Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–26

On the Use of a Local $\hat{\mathrm{R}}$ to Improve MCMC Convergence Diagnostic*

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

Abstract. Diagnosing convergence of Markov chain Monte Carlo is crucial and remains an essentially unsolved problem. Among the most popular methods, the potential scale reduction factor, commonly named $\hat{R}$, is an indicator that monitors the convergence of output chains to a target distribution, based on a comparison of the between- and within-variances. Several improvements have been suggested since its introduction in the 90s. Here, we aim at better understanding the $\hat{R}$ behavior by proposing a localized version that focuses on quantiles of the target distribution. This new version relies on key theoretical properties of the associated population value. It naturally leads to proposing a new indicator $\hat{R}_\infty$, which is shown to allow both for localizing the Markov chain Monte Carlo convergence in different quantiles of the target distribution, and at the same time for handling some convergence issues not detected by other $\hat{R}$ versions.

Keywords: computational statistics, convergence diagnostics, Markov chain Monte Carlo, potential scale reduction factor.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms have strongly contributed to the popularity of Bayesian models to sample from posterior distributions, especially in high-dimensional or high computational settings. This success results in a variety of softwares increasingly used for a wide range of applications: Stan (Carpenter et al., 2017), PyMC3 (Salvatier et al., 2016), NIMBLE (de Valpine et al., 2017), or Pyro (Bingham et al., 2019), to cite a few. The fundamental idea behind these algorithms is the convergence of the sampling distribution to the target (typically the posterior) when the number of samples goes to infinity. A major challenge is therefore to know if the behavior for a finite number of draws is satisfactory or not. This allows for a handle on the number of iterations to be drawn, which is all the more crucial in complex models with costly sampling schemes. See Roy (2020) for a recent literature review on convergence diagnostics.

*SG acknowledges the support of the Chair Stress Test, Risk Management and Financial Steering, led by the École polytechnique and its Foundation and sponsored by BNP Paribas. JA acknowledges the support of the French National Research Agency (ANR-21-JSTM-0001).

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

‡Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK 38000 Grenoble, France, julyan.arbel@inria.fr

§EDF R&D dept. Péricles 91120 Palaiseau, France

 https://doi.org/10.1214/23-BA1399

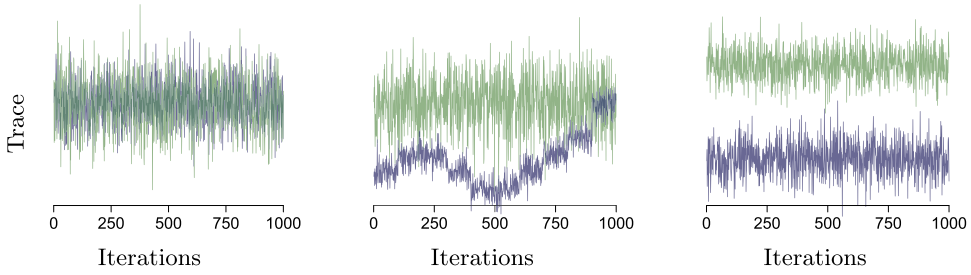

Figure 1: Traceplots illustrating convergence and two types of non-convergence of MCMC. Left: nothing indicates a convergence issue, as the two chains seem to have the same stationary distribution. Middle: the blue chain is still in an exploration phase and therefore is not stationary. Right: example where having multiple chains helps detecting a mixing issue despite a stationarity appearance of each.

1.1 Diagnosing MCMC convergence

Two frequently used properties to verify chains convergence are stationarity and mixing (see Vats and Flegal, 2021, for a discussion). *Stationarity* is related to the invariance property of the target distribution F for standard MCMC algorithms like Metropolis–Hastings or Gibbs sampling (Robert and Casella, 2004): if $\theta^{(i)}$ is the ith element of an MCMC chain, then $\theta^{(i)} \sim F$ implies $\theta^{(i+1)} \sim F$, so that as soon as an element of the chain is distributed according to F, all the following ones will be too. Thus, a chain whose distribution changes drastically during iterations is still in the exploration phase and is therefore not stationary (see middle panel in Figure 1). *Mixing* refers in practice to the exploration of the support of F: slow mixing chains correspond to chains that only explore a subset of the parameter space, which can lead to strong bias in the distribution (see Robert, 1995, for a more rigorous definition). A common way to limitate mixing issues is to run several chains in parallel with different starting points, which also allows comparing the chains together. Stationarity and mixing are two properties that can be treated separately: in principle, being stationary implies convergence to the target distribution and thus necessarily also mixing, but in practice there are examples of chains that seem to have reached stationarity but are not mixing (see right panel in Figure 1), hence the need for comparing multiple chains.

We place ourselves in the case of several chains: consider m chains of size n, with $\theta^{(i,j)}$ denoting the ith draw from chain j. We focus here on the Gelman–Rubin diagnostic (Gelman and Rubin, 1992), named potential reduction scale factor and commonly denoted by $\hat{R}$. It is by far one of the most popular methods to assess MCMC convergence, used in particular in Stan, PyMC3, or NIMBLE. The original heuristic for $\hat{R}$ construction is the comparison between two estimators that converge to the target variance $\mathrm{Var}[\theta]$, based on $\hat{W}$ and $\hat{B}$, respectively the estimated within- and between-variances. This diagnostic has the advantage of being scalar even in the case of a huge number of chains and comes with a rule of thumb that makes it very easy to use: generally $\hat{R} \geq 1$, and if it is greater than a given threshold (for example 1.01), then a convergence issue

is raised. This was originally constructed to diagnose mixing issues only, but Gelman et al. (2013, Section 11.4) suggest splitting the chains in two before computing $\hat{R}$ to check for stationarity at the same time. We will also always consider this split version of $\hat{R}$ throughout this paper, thus focusing only on the problem of mixing diagnostic.

1.2 Different $\hat{R}$ versions and their limitations

The original $\hat{R}$ of Gelman and Rubin (1992) has some limitations that are listed here with associated improvements suggested in the literature.

L1. It must be compared to an arbitrary chosen threshold. To use $\hat{R}$, a threshold must be set to determine a convergence issue. Originally set to 1.1, Vats and Knudson (2021) note that this choice is arbitrary and usually too optimistic. Thus, the authors propose a threshold according to a confidence level based on a relationship made with effective sample size (ESS). This observation was then shared by Vehtari et al. (2021) who suggest dropping the threshold to 1.01. Driven by practical arguments, this choice remains unprincipled nor theoretically justified, which is related to the next limitation.

L2. It suffers from a lack of interpretability. How to interpret a given value of $\hat{R}$? By construction, $\hat{R}$ is a ratio of two quantities that must estimate the posterior variance. Therefore, having a value close to one can be seen as having two correct estimations of the same quantity, which is an indication of convergence. However to our knowledge, no study investigates the theoretical or population value R that $\hat{R}$ aims at estimating, which would shed light on what is actually diagnosed. Typically, chains such that $\hat{R} \approx 1$ do not necessarily correspond to mixing chains: Vehtari et al. (2021) exhibit some counter-examples in order to motivate a more robust version called rank-$\hat{R}$. Still, the different versions of $\hat{R}$ only allow to draw conclusions when they are significantly greater than 1, and the common properties of chains producing $\hat{R} \approx 1$ are not well known as they are constructed at the estimator level.

L3. It is not robust to certain types of non-convergence. Traditional $\hat{R}$ can be fooled, in the sense that $\hat{R} \approx 1$ without convergence. This motivates the construction of rank-$\hat{R}$ (Vehtari et al., 2021), based on two cases where the original $\hat{R}$ is not robust:

(i) When the mean of the target distribution is infinite: in that case $\hat{W}$ and $\hat{B}$ are ill-defined and $\hat{R} \approx 1$ even though the chains follow different distributions. One solution is to apply rank transformation on the chains before computing $\hat{R}$ (this version is named bulk-$\hat{R}$ by Vehtari et al., 2021).

(ii) When the means of the chains are equal: in that case, the variance of means $\hat{B}$ is zero, and so $\hat{R} \approx 1$ even if the variances of chains are different. Here in addition to the rank-transformation, transforming the chains to get the deviation from their median allows to overcome this problem (this version is named tail-$\hat{R}$ by Vehtari et al., 2021).

Defining rank-$\hat{R} = \max\{\text{bulk-}\hat{R}, \text{tail-}\hat{R}\}$ overcomes the two issues at the same time. However, this robustness can be seen as very specific and can easily be fooled by simple

examples. One way is to consider chains with different distributions, but with (i) same mean (to fool bulk-$\hat{R}$), and (ii) same mean over the median (to fool tail-$\hat{R}$). For example, uniform $\mathcal{U}(\mu - 2\sigma, \mu + 2\sigma)$, normal $\mathcal{N}(\mu, \frac{\pi}{2}\sigma^2)$, or Laplace $\mathcal{L}(\mu, \sigma)$ distributions share the same mean (equal to μ) and same mean over the median (equal to σ), and thus mixing them yields rank-$\hat{R} \approx 1$. We provide an example and a more general framework to construct such cases in Appendix A of the supplementary material (Moins et al., 2023). One illustration can also be found in the right column of Figure 3. Although these counter-examples may never appear in practice, they do show some fairly counter-intuitive results that the additional layer of computation carried by rank-$\hat{R}$ makes even more difficult to analyse.

L4. It does not target a specific quantity of interest. Another point raised by Vehtari et al. (2021) is that the convergence diagnostic does not depend on inferential features of interest. It might be more precise to speak of convergence for a given posterior quantity, typically a mean, higher order moment, or quantile. Typically, practitioners apply $\hat{R}$ on quantities of interest such as the log-likelihood, the posterior density, or quantiles. On their side, Vehtari et al. (2021) suggest a local transformation on ESS to obtain a tail-ESS associated with 5% and 95% quantiles.

L5. It is associated with a univariate parameter. Although the vast majority of Bayesian models have multivariate parameters, $\hat{R}$ focuses on univariate convergence (i.e. convergence of margins). Some multivariate extensions exist, like Brooks and Gelman (1998) or Vats et al. (2019), but do not seem to be universally accepted: for example Stan or PyMC3 use instead a table containing univariate $\hat{R}$ with one value per parameter. However, assessing convergence on margins misses the point of dependence among parameter components, and does not guarantee the convergence of the joint distribution. Another version of $\hat{R}$ called R^* is suggested by Lambert and Vehtari (2021) and can deal with multivariate parameters: the idea is to use a classification algorithm which, in the case of converging chains, would not be able to identify to which chain a sample belongs. To avoid a result depending on the seed of the experiment, the authors suggest to draw several samples from the simplex obtained with the classification algorithm. In addition to the interpretability issues mentioned previously, this method has the constraint of not being able to study only a scalar value but a histogram, to check to what extent it contains or not the value 1.

We take a step forward in addressing all these limitations with a localized version of $\hat{R}$ briefly introduced in Moins et al. (2021) and developed here: we analyze $\hat{R}(x)$, a local version of $\hat{R}$ associated with a given quantile x, and the corresponding population value $R(x)$. This study leads us to propose a new indicator $\hat{R}_\infty$. In addition to being more interpretable, $\hat{R}_\infty$ shows better results than $\hat{R}$ in terms of MCMC convergence diagnostic, both on simulated experiments and on Bayesian models. As with all other versions of $\hat{R}$, this one can be applied to any MCMC algorithm: Metropolis–Hastings, Hamiltonian Monte-Carlo (HMC, Neal, 2011), No-U-Turn Sampler (NUTS, Hoffman and Gelman, 2014), etc.

The rest of the paper is organized as follows: we introduce in Section 2 the population version $R(x)$ and the corresponding sample version $\hat{R}(x)$, as well as their scalar counterparts R_∞ and $\hat{R}_\infty$. Since this proposed version depends on a quantile x and is

constructed at a population level, it is both targeting a specific quantity of interest and interpretable, addressing respectively limitations L4 and L2. We also establish several properties on the behavior of $R(x)$ function and on the convergence of the estimator $\hat{R}(x)$, helping in establishing a threshold and addressing limitation L1. Our proposed approach to deal with the multivariate case of limitation L5 is described in Section 3. Some empirical results are given in Section 4, showing that our proposed solution helps overcoming many of convergence issues identified in limitation L3. We conclude in Section 5. All proofs and details of the calculations are provided in the supplementary material, and experiments are available online[1] as well as the R package `localrhat` (Moins et al., 2022) containing our diagnostic implementation.

2 Local version of $\hat{R}$

Since the original version of Gelman and Rubin (1992), the heuristic for the construction of $\hat{R}$ was based on an analysis of variance. It consists in comparing two estimators of the posterior variance $\mathrm{Var}[\theta]$. The first one is the within-variance $\hat{W}$, which underestimates $\mathrm{Var}[\theta]$ as the bias of the estimator is (most of the time) strictly negative if the elements of the chains are not i.i.d, see Vats and Knudson (2021). The second one adds the between-variance $\hat{B}$ as a bias correction. This typically overestimates $\mathrm{Var}[\theta]$ if the initial values are chosen over-dispersed. As pointed out by Vats and Knudson (2021), following this heuristic does not exclude the use of other estimators of the bias than $\hat{B}$. Moreover, defining $\hat{R}$ at the sample level hinders a theoretical study of a population version to be conducted. Another justification can start with the law of total variance: assume that a univariate θ is sampled using m chains, and let $Z \in \{1, \dots, m\}$ be the corresponding index of the chain. Then,

$$\mathrm{Var}[\theta] = \mathbb{E}_Z[\mathrm{Var}_{\theta|Z}[\theta \mid Z]] + \mathrm{Var}_Z[\mathbb{E}_{\theta|Z}[\theta \mid Z]]. \tag{1}$$

The two terms in the right-hand side correspond respectively to the population versions of the within-variance W and the between-variance B. Replacing them by their estimated versions yields the original $\hat{R}$ formula of Gelman and Rubin (1992). In the following, we use (1) on a chains transformation which allows to localise convergence at a given quantile. For the theoretical study, we suppose stationarity of the chains to focus only on chain mixing issues. Thus, samples within a chain $j \in \{1, \dots, m\}$ may be correlated but are all distributed according to the same distribution F_j which may vary with j.

2.1 Population version

For all $x \in \mathbb{R}$, introduce the Bernoulli random variable $I_x = \mathbb{I}\{\theta \leq x\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function. Similarly to the Raftery–Lewis diagnostic (Raftery and Lewis, 1992), the idea of our local convergence estimate is decidedly simple: we use I_x in place of θ in the original Gelman–Rubin construction. The population within-chain and between-chain variances at point x are then defined respectively as $W(x) = \mathbb{E}\left[\mathrm{Var}[I_x \mid Z]\right]$ and

[1] https://theomoins.github.io/localrhat/Simulations.html.

$B(x) = \mathrm{Var}[\mathbb{E}[I_x \mid Z]]$. Note that both quantities exist whatever the tail heaviness of θ distribution thanks to introduction of the indicator function, thus relaxing moment conditions of the original $\hat{R}$. We define the associated population $R(x)$ as

$$R(x) = \sqrt{\frac{W(x) + B(x)}{W(x)}}.$$

It turns out that under the assumption of stationarity for each chain, $R(x)$ can be expressed in closed-form with respect to the chains' distribution.

Proposition 2.1. *Suppose that, for any $j \in \{1, \ldots, m\}$, $\mathbb{P}(Z = j) = 1/m$ and θ given $Z = j$ has cumulative distribution function (cdf) F_j. Then, one has for any $x \in \mathbb{R}$:*

$$R(x) = \sqrt{1 + \frac{\sum_{j=1}^{m} \sum_{k=j+1}^{m} \left(F_j(x) - F_k(x)\right)^2}{m \sum_{j=1}^{m} F_j(x)(1 - F_j(x))}}. \tag{2}$$

Thus, using I_x instead of θ defines a local convergence estimate at any point x which quantifies a distance between the F_j's. This allows for diagnosing convergence relatively to a quantile one wants to estimate (for a posterior credible interval for example). The following proposition states straightforward properties of $R(x)$ emanating from (2):

Proposition 2.2. *The population $R(x)$ satisfies the following properties:*

(i) *$R(x) \geq 1$ for all $x \in \mathbb{R}$.*

(ii) *$R(x) = 1$ for all $x \in \mathbb{R}$ if and only if $F_1 = \cdots = F_m$.*

(iii) *$R(x) \to 1$ as $|x| \to \infty$.*

(iv) *$R(x)$ inherits continuity property of $F_1, \ldots, F_m$ if the support of the F_j's are overlapping.*

Based on these results and in order to summarize this continuous index into a scalar one, we may also consider its supremum over $\mathbb{R}$:

$$R_\infty = \sup_{x \in \mathbb{R}} R(x). \tag{3}$$

Note that, in view of Proposition 2.2(iv), R_∞ is finite simply as soon as the F_j's are continuous with overlapping supports. Considering R_∞ amounts to considering the local version $R(x)$ corresponding to the quantile x with the poorest convergence when no information is given on the posterior interval used for inference.

2.2 Sample version

Population version $R(x)$ can be estimated by replacing the $F_j(x)$'s in (2) by their empirical counterparts $\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\theta^{(i,j)} \leq x\}$. This is equivalent to computing

the original version of $\hat{R}$ on indicator variables $I_x^{(i,j)} = \mathbb{I}\{\theta^{(i,j)} \leq x\}$ instead of $\theta^{(i,j)}$. This connects with the Raftery–Lewis diagnostic (Raftery and Lewis, 1992) and more recently with Vehtari et al. (2021) who suggest this transformation for effective sample size (ESS) to construct graphical diagnostics or "tail-versions" of this diagnostic. Moreover, a rank-normalization step is added in Vehtari et al. (2021)'s to prevent from infinite moments, although using $I_x^{(i,j)}$ ensures the index existence whatever the $\theta^{(i,j)}$ distribution is. Skipping this step for $\hat{R}$ yields an explicit expression of what is estimated in the stationary case with (2). This makes the diagnostic more interpretable and allows us to obtain key theoretical results for the associated theoretical R and R_∞.

Note that for a given number of chains m and chain length n, $\hat{R}(x)$ can only take $m(n+1)$ different values, as the computation is based on nm indicator variables. Thus, the best accuracy we can obtain for $\hat{R}_\infty$ for a given n and m consists in evaluating $\hat{R}(x)$ at all the $\theta^{(i,j)}$'s. This can be accelerated by subsampling, often with limited decrease in accuracy.

2.3 Convergence properties

Let us assume that all m chains are mutually independent and have converged to a common distribution so that $F_1 = \cdots = F_m =: F$. Assume, moreover, that a Markov chain central limit theorem holds (see for instance Robert and Casella, 2004, Theorem 6.65), so that we can write

$$\sqrt{n}(\hat{F}_j(x) - F(x)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(x)\right), \tag{4}$$

as $n \to \infty$, for all $j \in \{1, \ldots, m\}$ and where $\sigma^2(x)$ is some asymptotic variance. In particular in the i.i.d. setting, $\sigma^2(x) = F(x)(1 - F(x))$. Letting $\hat{F}(x) = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{I}\{\theta^{(i,j)} \leq x\} = \frac{1}{m}\sum_{j=1}^{m}\hat{F}_j(x)$ and taking into account of the independence between chains yield

$$\sqrt{nm}(\hat{F}(x) - F(x)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(x)\right), \tag{5}$$

as $n \to \infty$, and $\sigma(x)/\sqrt{nm}$ can be interpreted as the Monte Carlo standard error (MCSE) associated with $\hat{F}(x)$. Following the definition of the ESS used in Gong and Flegal (2016) or Vats et al. (2019), we can define a local-ESS as the ratio of the target variance to the squared MCSE:

$$\mathrm{ESS}(x) = nm\frac{F(x)(1 - F(x))}{\sigma^2(x)}. \tag{6}$$

This quantity is in line with the definition of ESS for quantile of Vehtari et al. (2021), and has already been studied by Raftery and Lewis (1992) who focus on this indicator transformation and approximate the resulting process as a two-state Markov chain. This yields an explicit expression of the stationary distribution F, which can be used to obtain an expression of $\mathrm{ESS}(x)$ as a function of the transition probabilities. Several limitations of this two-state Markov chain approximation are raised by Brooks and Roberts (1999); Doss et al. (2014), for example. A more general way to estimate $\mathrm{ESS}(x)$ is to apply the

same idea as in the definition of the local $\hat{R}(x)$: use any estimator of ESS (Robert and Casella, 2004; Gelman et al., 2013) on indicator variables $I_x^{(i,j)}$ instead of $\theta^{(i,j)}$.

Combining the asymptotic result (5) with expression (6) yields the following large n limiting distribution result on $\hat{R}(x)$ (χ^2_{m-1} denotes the chi-square distribution with $m-1$ degrees of freedom).

Proposition 2.3. *Assume that all m chains are mutually independent and have converged to a common distribution $F \coloneqq F_1 = \cdots = F_m$. Then:*

(i) *The distribution of $\hat{R}_\infty$ does not depend on the underlying distribution F.*

(ii) *For any $x \in \mathbb{R}$, $\mathrm{ESS}(x)(\hat{R}^2(x) - 1) \xrightarrow{d} \chi^2_{m-1}$ as $n \to \infty$.*

Note that casting the problem of convergence monitoring in terms of analysing components of variance from multiple sequences dates back to Gelman and Rubin (1992), Section 2.2, and earlier works by Fosdick (1959); Gelfand and Smith (1990). Let us highlight that the assumption $F_1(x) = \cdots = F_m(x)$ is equivalent to the ANOVA (analysis of variance) hypothesis $\mathbb{E}(I_x^{(\cdot,1)}) = \cdots = \mathbb{E}(I_x^{(\cdot,m)})$ and that the statistics studied in Proposition 2.3(ii) can similarly be rewritten in terms of the ANOVA test statistics: $\hat{R}^2(x) - 1 = \hat{B}(x)/\hat{W}(x)$, where $\hat{B}(x)$ and $\hat{W}(x)$ are the respective empirical counterparts of $B(x)$ and $W(x)$. These interpretations can then be used to derive a statistical test on the convergence of the chains. To this end, note also that the limit in distribution of Proposition 2.3(ii) still holds when $\mathrm{ESS}(x)$ is replaced by a consistent estimator $\widehat{\mathrm{ESS}}(x)$. This result allows computing the type I error associated with the null hypothesis that $\hat{R}(x) = 1$, in other terms that all the chains have converged to a common distribution at x. Let $z_{m-1,1-\alpha}$ be the quantile of level $1-\alpha$ of the χ^2_{m-1} distribution, and introduce the associated threshold

$$R_{\mathrm{lim},\alpha}(x) \coloneqq \sqrt{1 + \frac{z_{m-1,1-\alpha}}{\mathrm{ESS}(x)}}. \tag{7}$$

The type I error is then given by $\mathbb{P}(\hat{R}(x) \geq R_{\mathrm{lim},\alpha}(x)) \simeq \alpha$. As an illustration, some values of α are reported for the threshold $R_{\mathrm{lim},\alpha}(x) = 1.01$, $m = 4$ chains and different values of $\mathrm{ESS}(x)$ in the left panel of Table 1. For example, it appears that the probability of having $\hat{R}(x) > 1.01$ and $\mathrm{ESS}(x) = 400$ when convergence is reached is 0.04, and decreases quickly for larger values of $\mathrm{ESS}(x)$.

2.4 Threshold elicitation

Threshold for the local $\hat{R}(x)$ Proposition 2.3(ii) allows us to associate a threshold for $\hat{R}(x)$ to a type I error α, using the definition of $R_{\mathrm{lim},\alpha}(x)$ in (7). Some values are displayed in the right panel of Table 1 for a fixed $\mathrm{ESS}(x) = 400$ and $\alpha = 0.05$. It appears that the value of 1.01, the recent recommendation of Vehtari et al. (2021), seems to be coherent for $\hat{R}(x)$ and a moderate number of chains, typically the default configuration in Stan ($m = 4$), JAGS (Just Another Gibbs Sampler, Plummer, 2003)

m	$R_{\text{lim},\alpha}(x)$	ESS(x)	α
4	1.01	50	0.80
		100	0.57
		200	0.26
		400	0.04
		800	$< 10^{-3}$
		1500	$< 10^{-6}$

m	$R_{\text{lim},\alpha}(x)$	ESS(x)	α
2	1.005	400	0.05
4	1.010		
8	1.017		
15	1.029		
50	1.080		
100	1.144		

Table 1: Left: Type I error α as a function of ESS(x) when $R_{\text{lim},\alpha}(x) = 1.01$ and $m = 4$. Right: $R_{\text{lim},\alpha}(x)$ as a function of m when ESS$(x) = 400$ and $\alpha = 0.05$.

$(m = 3)$ or PyMC3 $(m = \max\{n_c, 2\}$ with n_c the number of cores). However, the value of m must be doubled if a split version is used, and when m increases the threshold becomes more severe and it may be appropriate to consider a higher (i.e. less stringent) one: for example, a threshold of 1.1 can be enough provided the number of chains m is larger than 100. The case of a large number of chains has been recently studied by Margossian et al. (2022) who suggest a new version of $\hat{R}$ for this configuration. Note that a similar observation about the stringency of the threshold can be made with rank-$\hat{R}$, see Appendix C for more details.

Therefore, we recommend to keep the threshold of 1.01 as a general rule of thumb for $\hat{R}(x)$, except if the number of chains is too large or if one wants to have a more precise threshold. In such a case it only requires to provide α, m and a target value ESS(x) to compute $R_{\text{lim},\alpha}(x)$ using (7).

Threshold for the supremum $\hat{R}_\infty$ Proposition 2.3 does not induce any threshold $R_{\infty,\text{lim}}$ for $\hat{R}_\infty$, since Proposition 2.3(ii) only establishes the pointwise convergence of the empirical process $\hat{R}(\cdot)$. However, Proposition 2.3(i) shows that under the null hypothesis where all chains follow a common distribution F, the latter F is irrelevant to the $\hat{R}_\infty$ statistic. Such an independence to the underlying distribution F makes it possible the use of a quantile of $\hat{R}_\infty$ as a threshold associated with a given probability α and number of chains m. Table 2 provides estimations of $R_{\infty,\text{lim}}$ using replications for several values of α and m and a fixed number of effective samples of 400, as recommended by Vehtari et al. (2021) (more details are provided in Appendix C). Here, we can see that a fixed rule of thumb for a range of m would be too imprecise, as the quantile values increase rapidly with m. Nevertheless, Table 2 illustrates a linear relationship between m and the appropriate threshold for a given α.

In the simulations in Section 2.5 and in the experiments in Section 4, we mostly consider $m = 4$ and therefore choose a threshold of 1.02, which is a little more accurate than 1.01 by looking at Table 2. Note that if $m = 8$ or if a split version of $\hat{R}_\infty$ is used with $m = 4$, then a threshold of 1.03 should be preferred. In the `localrhat` R package (Moins et al., 2022), the computation of $\hat{R}_\infty$ comes with the associated threshold at 5% based on the calculations in Table 2, as well as a p-value associated with the obtained $\hat{R}_\infty$.

	$R_{\infty,\mathrm{lim}}$			
m \ α	0.005	0.01	0.05	0.1
2	1.018	1.016	1.012	1.010
3	1.023	1.022	1.016	1.014
4	1.027	1.025	1.020	1.018
8	1.038	1.037	1.031	1.028
10	1.043	1.041	1.036	1.033
20	1.080	1.076	1.062	1.056

Table 2: Empirical quantiles $R_{\infty,\mathrm{lim}}$ of the $\hat{R}_\infty$ distribution under the null hypothesis that all chains follow the same distribution for a target ESS of 400, based on 2000 replications.

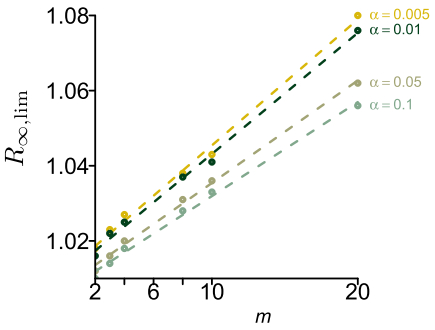

Figure 2: Illustration of the values in Table 2 and of the linearity with m for a fixed α.

2.5 Illustrative examples

In this section, we consider toy distributions for the chains, where the computation of R_∞ can be done explicitly. In particular, we first focus on two cases raised by Vehtari et al. (2021) of deficient behavior of the traditional $\hat{R}$. Then, we exhibit a failure situation for rank-$\hat{R}$. All these theoretical behaviors are illustrated on a simulation study. Further applications to Bayesian inference are provided in Section 4, and other examples where $\hat{R}$ and rank-$\hat{R}$ fail in Appendix D.

Example 1: Chains with same mean and different variances. To tackle the first situation of poor behavior of the traditional $\hat{R}$, we consider m chains following centered uniform distributions with different variances. More specifically, assume that the $m-1$ first chains have the cdf $F_1 = \cdots = F_{m-1}$ of the uniform distribution $\mathcal{U}(-\sigma, \sigma)$ while the last chain has the cdf F_m of the uniform distribution $\mathcal{U}(-\sigma_m, \sigma_m)$ with $0 < \sigma \leq \sigma_m$. In such a case, the between-variance is zero and it is thus expected that $\hat{R} \approx 1$. In contrast, Lemma D.2 in Appendix D provides an explicit expression for $R(x)$ as well as

$$R_\infty = \sqrt{1 + \frac{m-1}{m}\left(1 - \frac{2}{1+\sigma_m/\sigma}\right)}.$$

It appears that R_∞ is an increasing function of σ_m/σ starting from $R_\infty = 1$ when $\sigma_m/\sigma = 1$, and upper-bounded by $\sqrt{2 - 1/m}$ when $\sigma_m/\sigma \to \infty$. Results are illustrated in the left column of Figure 3. In the bottom row, the histograms of replications confirm that $\hat{R}_\infty$ is able to spot the same convergence issue as the one Vehtari et al. (2021) suggests.

Example 2: Chains with heavy-tails and different locations. As a second example of poor behavior of $\hat{R}$, we consider chains following Pareto(α, η) distributions, with cdf

$$F(x \mid \alpha, \eta) = 1 - (x/\eta)^{-\alpha}, \quad \forall x \in [\eta, +\infty),$$

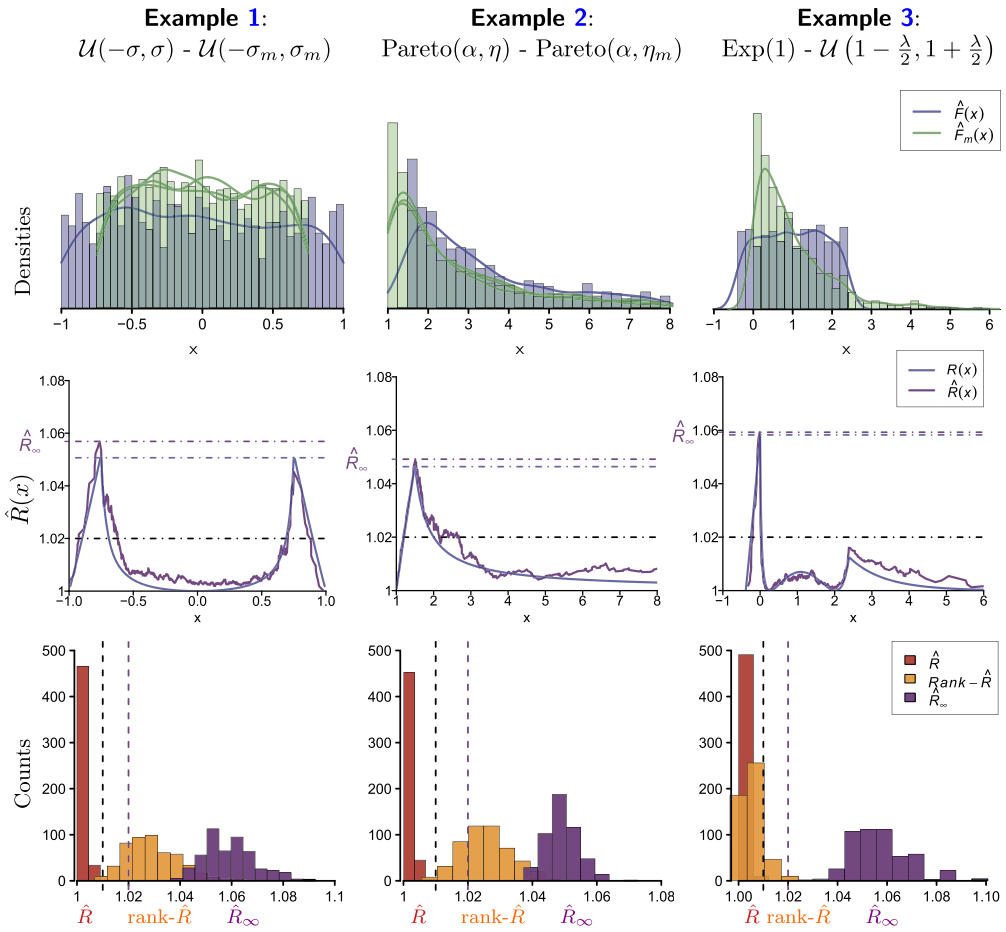

Figure 3: Illustrations with $m = 4$ chains, $n = 200$ independent iterations each. Top row: Simulation of $F_1 = \cdots = F_{m-1}$ in green distinct from F_m in blue. For the uniform example (left), $\sigma = 3/4$ and $\sigma_m = 1$, for the Pareto (middle) $\eta = 1$ and $\eta_m = 1.5$, and for the uniform (right) $\lambda = 4\log(2)$. Second row: The corresponding population version $R(x)$ and empirical version $\hat{R}(x)$ as functions of x for one replication. Bottom row: Histograms of 500 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. Dashed lines correspond to the threshold of 1.01 for $\hat{R}$ and rank-$\hat{R}$ and 1.02 for $\hat{R}_\infty$ (see Section 2.3).

shape parameter $\alpha > 0$ and lower bound $\eta > 0$. Let us recall that such a distribution is heavy-tailed (Embrechts et al., 2013, Table 3.4.2) and has an infinite first moment when $\alpha \leq 1$. We focus on the case where one chain is shifted from the other ones: $F_1(x) = \cdots = F_{m-1}(x) = F(x \mid \alpha, \eta)$ and $F_m(x) = F(x \mid \alpha, \eta_m)$ with $0 < \eta \leq \eta_m$ and $\alpha \leq 1$. Here, the within- and between-variances do not exist and it is expected in

practice that $\hat{R} \approx 1$. In contrast, R_∞ can be written as

$$R_\infty = \sqrt{1 + \frac{1}{m}\left(\left(\frac{\eta_m}{\eta}\right)^\alpha - 1\right)},$$

see Lemma D.3 in the supplementary material. Clearly, R_∞ is an increasing function of η_m/η starting from $R_\infty = 1$ when $\eta_m = \eta$ and such that $R_\infty \to \infty$ as $\eta_m/\eta \to \infty$. Results are shown in the middle column of Figure 3. This experiment corresponds to the second example of convergence issue raised by Vehtari et al. (2021). The same observations as for Example 1 can be made here: $\hat{R}_\infty$ is prone to indicating a convergence issue than rank-$\hat{R}$.

Example 3: Chains with same mean and mean over the median. Finally, we come back to the example described in Section 1.2 where both $\hat{R}$ and rank-$\hat{R}$ fail to detect non-convergence. Following the method described in Appendix A, we consider $m-1$ exponential chains Exp(1) and one uniform $\mathcal{U}(1-2\log 2, 1+2\log 2)$. This results in chains with same mean and mean over the median. Results are illustrated in the right panel of Figure 3: the histograms of replications confirm that $\hat{R}_\infty$ is able to detect the convergence issue that neither $\hat{R}$ nor rank-$\hat{R}$ are able to detect. Here, the explicit calculation of R_∞ is not feasible, but Lemma D.4 in the supplementary material provides another example where the computation can be done, with uniform and Laplace distributions.

3 Multivariate extension

3.1 Population version and algorithm for multivariate diagnosis

Our $R(x)$ can naively be adapted to the multivariate case: assume now that the parameter is multivariate and write $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ with $d \geq 2$, and denote by $\theta_p^{(j)}$ the coordinate $p \in \{1, \ldots, d\}$ from chain $j \in \{1, \ldots, m\}$. Similarly to the univariate case, $\hat{R}$ can be computed on the indicator variables $I_{\boldsymbol{x}}^{(j)} = \mathbb{I}\{\theta_1^{(j)} \leq x_1, \ldots, \theta_d^{(j)} \leq x_d\}$ for any $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Under the assumptions of Proposition 2.1, all calculations remain valid in dimension d and therefore the expression of $R(\boldsymbol{x})$ is formally the same as in (2):

$$R(\boldsymbol{x}) = \sqrt{\frac{W(\boldsymbol{x}) + B(\boldsymbol{x})}{W(\boldsymbol{x})}} = \sqrt{1 + \frac{\sum_{j=1}^m \sum_{k=j+1}^m \left(F_j(\boldsymbol{x}) - F_k(\boldsymbol{x})\right)^2}{m \sum_{j=1}^m F_j(\boldsymbol{x})(1 - F_j(\boldsymbol{x}))}}. \tag{8}$$

The properties listed in Proposition 2.2 in the univariate case remain true as well. The associated R_∞ is defined as $R_\infty(F_1, \ldots, F_m) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} R(\boldsymbol{x})$, while $\hat{R}(\boldsymbol{x})$ is computed by replacing the cumulative distribution functions in (8) by their empirical counterparts. Note also that all values computed in Table 1 and Table 2 remain identical in this multivariate extension. However, those results are not giving information about the sensitivity to convergence issues, which in the multivariate case can come from margins but also from the dependence structure.

It is easily seen that, if the marginal distributions of $F_1, \dots, F_m$ coincide, then R_∞ is the same as the one associated with uniform margins (see Lemma B.1 in the supplementary material). In other words, we have $R_\infty(F_1, \dots, F_m) = R_\infty(C_1, \dots, C_m)$ where C_j is the copula defined in $[0,1]^d$ associated with F_j, $j \in \{1, \dots, m\}$. This suggests that a multivariate diagnosis can be conducted in two steps as follows:

1. Compute the univariate $\hat{R}_{\infty,p}$ separately on each of the coordinates $p \in \{1, \dots, d\}$. If $\hat{R}_{\infty,p} < R^{(M)}_{\infty,\lim}$ for all $p \in \{1, \dots, d\}$, with $R^{(M)}_{\infty,\lim}$ a choice of margins threshold, then all of them are deemed to have converged and to be identically distributed.
2. Compute the multivariate $\hat{R}_\infty$ to check the dependence structure convergence. If $\hat{R}_\infty < R^{(C)}_{\infty,\lim}$, with $R^{(C)}_{\infty,\lim}$ a copula threshold, then the dependence structure is also deemed to have converged, and so has the multivariate distribution.

The test for convergence is now separated in two parts: 1. convergence of the margins, and 2. convergence of the copula knowing that the margins have converged. It can easily be shown that, up to a first order approximation, one way to obtain a type I error α for the global two-step test is to consider a level $\alpha/2$ for each of the two components. The first step corresponds to d univariate tests, so for $R^{(M)}_{\infty,\lim}$ one can use the univariate threshold $R_{\infty,\lim}$ defined in Section 2.4 with a level $\alpha/2d$, corresponding to a Bonferroni correction for the error level $\alpha/2$. In the following subsections, we focus on the second step of the algorithm: the theoretical properties of the multivariate $\hat{R}_\infty$ in the case of convergence on the margins, which will provide insights for choosing $R^{(C)}_{\infty,\lim}$. Values of $R^{(M)}_{\infty,\lim}$ and $R^{(C)}_{\infty,\lim}$ are then given as functions of (α, d, m) in Table 1. As a general rule, one can reasonably use for $\alpha = 0.05$ the values $(R^{(M)}_{\infty,\lim}, R^{(C)}_{\infty,\lim}) = (1.03, 1.03)$ in the case of $m = 4$ chains, and $(R^{(M)}_{\infty,\lim}, R^{(C)}_{\infty,\lim}) = (1.04, 1.05)$ if $m = 8$ or if a split version is used with $m = 4$, with limited variations around these values for varying dimension d.

3.2 Upper bounds

Let us first consider the case of $m = 2$ chains with uniform margins and associated copulas C_1 and C_2. For all $\boldsymbol{u} = (u_1, \dots, u_d) \in [0,1]^d$, one has

$$R(\boldsymbol{u}) = \sqrt{1 + \frac{\left(C_1(\boldsymbol{u}) - C_2(\boldsymbol{u})\right)^2}{2\left(C_1(\boldsymbol{u})(1 - C_1(\boldsymbol{u})) + C_2(\boldsymbol{u})(1 - C_2(\boldsymbol{u}))\right)}}. \tag{9}$$

In addition to having the usual lower bound of 1, the next lemma allows establishing an upper bound on $R_\infty(C_1, C_2)$.

Lemma 3.1. *Let C_1, C_2, C_- and C_+ be copulas such that:*

$$\text{for all } \boldsymbol{u} \in [0,1]^d, \begin{cases} C_-(\boldsymbol{u}) \le C_1(\boldsymbol{u}) \le C_+(\boldsymbol{u}), \\ C_-(\boldsymbol{u}) \le C_2(\boldsymbol{u}) \le C_+(\boldsymbol{u}). \end{cases} \tag{10}$$

Then, $R_\infty(C_1, C_2) \le R_\infty(C_-, C_+)$.

Let W_d and M_d the lower and upper Fréchet–Hoeffding bounds in dimension d (see Nelsen, 2006, Theorem 2.10.12):

$$W_d(\boldsymbol{u}) \coloneqq \max\left\{1 - d + \sum_{i=1}^{d} u_i, 0\right\} \quad \text{and} \quad M_d(\boldsymbol{u}) \coloneqq \min\left\{u_1, \ldots, u_d\right\}.$$

Any copula is bounded from below and from above by W_d and M_d respectively, in the sense of (10). Thus, applying Lemma 3.1 with $(C_-, C_+) = (W_d, M_d)$ yields:

Proposition 3.1. *For any d-variate copulas C_1 and C_2,*

$$R_\infty(C_1, C_2) \leq \sqrt{\frac{d+1}{2}}.$$

Unlike the univariate version (see for instance Example 2 in Section 2.5), the value of R_∞ associated with the convergence of the dependence structure is upper-bounded, with a bound that grows with the dimension. This difference of behavior could be used for example to tune the threshold for the multivariate case. However this bound, although it is the "best possible" (Nelsen, 2006, Theorem 2.10.13), is tight only in the case $d = 2$ since W_d is no more a copula when $d > 2$. It may also be too loose since it compares the extreme case of one chain with comonotonic dependence and another one with anti-comonotonic dependence. Some refinements are proposed in Section 3.3.

In the case of $m > 2$ chains, the previous bounding technique does not apply anymore, and we propose the following result based on bounding pairwise R_∞'s:

Corollary 3.1. *For any $m \geq 2$ and d-variate copulas $(C_1, \ldots, C_m)$,*

$$R_\infty(C_1, \ldots, C_m) \leq \sqrt{1 + \frac{m-1}{2}(d-1)}.$$

Although this limit is not tight in the general case, it coincides with the upper bound of Proposition 3.1 when $m = 2$. Let us also note that, for any fixed $m \geq 2$, the upper bound of $R_\infty(C_1, \ldots, C_m)$ diverges at a fixed $\sqrt{d}$ rate as the dimension increases.

3.3 Influence of the dependence direction on the sensitivity of $\hat{R}_\infty$

When $m = 2$, one way to refine the upper bound established in Proposition 3.1 is to assume that both copulas are modelling either positive of negative dependence. More specifically, let us recall the notions of positive lower orthant dependence (PLOD) and negative lower orthant dependence (NLOD) (see Nelsen, 2006, Section 5.7). The random vector $(\theta_1, \ldots, \theta_d)$ is

- PLOD if $\forall \boldsymbol{x} \in \mathbb{R}^d$, $\quad \mathbb{P}(\theta_1 \leq x_1, \ldots, \theta_d \leq x_d) \geq \prod_{i=1}^{d} \mathbb{P}(\theta_i \leq x_i)$,
- NLOD if $\forall \boldsymbol{x} \in \mathbb{R}^d$, $\quad \mathbb{P}(\theta_1 \leq x_1, \ldots, \theta_d \leq x_d) \leq \prod_{i=1}^{d} \mathbb{P}(\theta_i \leq x_i)$.

Both properties can be characterized in terms of the associated copula. The PLOD (resp. NLOD) property holds if and only if $C(\boldsymbol{u}) \geq \Pi_d(\boldsymbol{u})$ (resp. $C(\boldsymbol{u}) \leq \Pi_d(\boldsymbol{u})$) for all $\boldsymbol{u} \in [0,1]^d$ where Π_d is the independent copula defined by $\Pi_d(\boldsymbol{u}) \coloneqq \prod_{i=1}^d u_i$. Note that this does not define a total order on copulas since some copulas are neither PLOD nor NLOD. Nevertheless, it allows us to derive refined bounds for R_∞ in the NLOD and PLOD cases.

For PLOD, the upper bound is in not closed-form for any dimension d, but simple bounds can be derived in the two extreme cases $d = 2$ and $d \to \infty$.

Corollary 3.2. *Let $m = 2$. For any two PLOD d-variate copulas C_1 and C_2, $R_\infty(C_1, C_2) \leq R_\infty(\Pi_d, M_d)$ with*

$$\begin{cases} R_\infty(\Pi_2, M_2) = \sqrt{\frac{1}{2} + \frac{1}{\sqrt{3}}} \approx 1.038 & \text{if } d = 2, \\ \sqrt{\frac{d}{2\log d}}(1 + o(1)) \leq R_\infty(\Pi_d, M_d) \leq \sqrt{\frac{d+1}{2}} & \text{as } d \to \infty. \end{cases}$$

Conversely, the upper bound can be computed explicitly in the NLOD case.

Corollary 3.3. *Let $m = 2$. For any two NLOD d-variate copulas C_1 and C_2, $R_\infty(C_1, C_2) \leq R_\infty(\Pi_d, W_d)$ with*

$$R_\infty(\Pi_d, W_d) = \sqrt{1 + \frac{1}{2}\frac{1}{\left(1 - \frac{1}{d}\right)^{-d} - 1}}.$$

Let us stress that positive and negative dependence are handled differently by R_∞. When $d = 2$, the PLOD and NLOD bounds (respectively equal to 1.04 and 1.08) are significantly lower than the value $\sqrt{3/2} \approx 1.22$ corresponding to the global bound, with a value higher in the NLOD case than in the PLOD one. However, this observation is quickly inverted when d increases: for NLOD, $R_\infty(\Pi_d, W_d)$ is bounded and converges to $\sqrt{1 + \frac{1}{2(e-1)}} \approx 1.136$ as $d \to \infty$, which strongly constrains the range of values that can be obtained whatever the dimension. In contrast, the upper bound $R_\infty(\Pi_d, M_d)$ in the PLOD case diverges with the dimension, at the same rate (up to a logarithmic factor) as in the general case, see Proposition 3.1. Thus, the sensitivity of R_∞ strongly depends on the sign of dependence and asymptotically favors PLOD dependence when d increases.

This difference can be explained by the construction of $R(x)$ itself (and thus R_∞), which favors a dependence direction in $\mathbb{R}^d$ due to the computation of $\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \ldots, \theta_d^{(\cdot)} \leq x_d\}$. One way to overcome this issue in the bivariate case is to compute two versions of R_∞, denoted respectively by R_∞^+ and R_∞^-, based respectively on $\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \leq x_2\}$ and $\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \geq x_2\}$. Note that R_∞^+ coincides with the construction proposed in Section 3.1.

Corollary 3.4. *Let $m = 2$. Then, $R_\infty^+(\Pi_2, M_2) = R_\infty^-(W_2, \Pi_2)$ and $R_\infty^+(W_2, \Pi_2) = R_\infty^-(\Pi_2, M_2)$.*

It appears that PLOD and NLOD upper bounds are exchanged by computing R_∞^- instead of R_∞^+, which makes R_∞^- more sensitive to negative dependence than positive dependence (in the bivariate case). One way to consider symmetrically both dependencies would be to consider $\hat{R}_\infty^{(\max)} = \max(R_\infty^+, R_\infty^-)$. However, in dimension d, considering all directions would imply the computation of 2^{d-1} different R_∞, which would be too expensive for large d. Similar curse of dimensionality occurs in the multivariate extension of the Kolmogorov–Smirnov test, see for example Lopes et al. (2007) for improvements of the naive multidimensional version of the test. Computing $\hat{R}_\infty^{(\max)}$ is still feasible for small values of d: typically for $d \leq 6$ we were able to replicate values in our experiments. Therefore, we provide in Table 1 (Appendix C) the estimated threshold $R_{\infty,\lim}^{(C)}$ associated with the maximum of $\hat{R}_\infty$ in all possible directions when $d \leq 6$.

One alternative in the high-dimensional case could be to apply $\hat{R}$ on an indicator function associated with a univariate function of the parameters, to return to the case described in Section 2. Typically in a Bayesian model, one could use the log-likelihood $l_{\boldsymbol{\theta}} = \log p(y \mid \boldsymbol{\theta})$ when it is available, and compute $\hat{R}_\infty$ with $\mathbb{I}\{l_{\boldsymbol{\theta}} \leq x\}$. Similarly, the log posterior as implemented in Stan can also be used, as suggested in the Stan reference manual (Carpenter et al., 2017). Ensuring convergence for all x on the log posterior may be satisfying for multivariate diagnosis, as it is illustrated in Example 9.

3.4 Multivariate illustrative examples

Similarly to Section 2.5, we illustrate our theoretical study in the multivariate case with simulations based on toy distributions for the chains. Especially, we consider multivariate normal distributions, and focus on the case where all the margins are the same (typically distributed according to a standard normal distribution). This leads to

$$\boldsymbol{\theta}^{(i,j)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_j),$$

$i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, where $\boldsymbol{\Sigma}_j$ is the covariance matrix of the chain j, with diagonal elements equal to one to keep standard Gaussian margins.

Example 4: Bivariate normal distributions with different correlation terms. In the bivariate case, the dependence structure is driven by only one value, which is the off-diagonal element $\rho_j \in (-1, 1)$ of $\boldsymbol{\Sigma}_j$. Similarly to other examples, we suppose that we have $m - 1$ converging chains with identity covariance matrix ($\rho_1 = \cdots = \rho_{m-1} = 0$) while $\rho_m \in (-1, 1)$ for the last one.

Results are shown in Figure 4, with a comparison of $\hat{R}_\infty$ with the multivariate $\hat{R}$ of Brooks and Gelman (1998). The histogram on the left represents the values of the two diagnostics for 100 replications with $m = 2$, $n = 200$ and $\rho_m = 0.9$. Despite a large difference on the covariance term between the chains, we can see that Brooks–Gelman $\hat{R}$ fails to correctly diagnose this difference, as most of the values are between 1 and 1.01, contrary to $\hat{R}_\infty$. Due to the i.i.d nature of the example, the recent proposal of Vats and Knudson (2021) for a multivariate $\hat{R}$ does not detect any convergence issue as the diagnostic is not based on a comparison between chains. This difference of

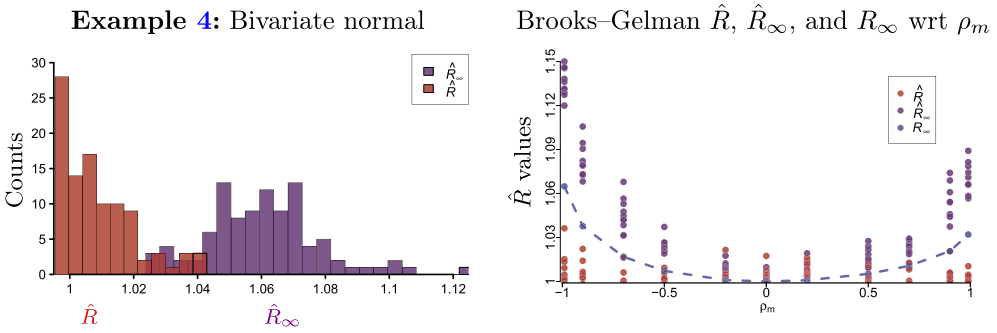

Figure 4: Behavior of Brooks–Gelman $\hat{R}$ (in orange) and multivariate $\hat{R}_\infty$ (in violet) in the case of chains with bivariate normal distributions, with different off-diagonal elements in the covariance matrix. On the left: Histograms with 100 replications with one standard normal chain and one with $\rho_m = 0.9$. On the right: The same experiment with 10 replications for different values of ρ_m, plotted as a function of ρ_m, and the corresponding population R_∞ in blue.

behavior is confirmed on the right panel of Figure 4, which illustrates 10 replications of both diagnostics as a function of ρ_m. For instance, if $\rho_m = 0$ then the four chains are identically distributed and no convergence issue should be raised. Conversely, the value of $\hat{R}$ should increase when $|\rho_m| \to 1$, as the difference between the last chain and the other ones increases. For the Brooks–Gelman version, we can see that the value of $\hat{R}$ is almost constant and thus insensitive to ρ_m, which is not satisfactory, contrary to $\hat{R}_\infty$ which has a parabolic shape.

As discussed in Section 3.3, the behavior of $\hat{R}_\infty$ is not symmetric when $\rho_m \to -1$ and $\rho_m \to 1$: the upper bound corresponding to positive dependence diverges with the dimension (Corollary 3.2 for PLOD copulas) whereas the one for negative dependence is bounded by approximately 1.14 (Corollary 3.3 for NLOD copulas). This leads to the intuition that the convergence diagnostic is more sensitive in the PLOD case than in the NLOD, but this observation is asymptotic and when $d = 2$, the two bounds are respectively equal to 1.08 and 1.04, so the statement is reversed. This asymmetry is illustrated in Figure 4 on theoretical R_∞ (in blue) and estimations $\hat{R}_\infty$ (in purple).

Example 5: Evolution of the behavior when the dimension increases. In the general case of dimensionality $d > 2$, we still compare $m - 1$ chains that follow a multivariate standard normal distribution with one that has a given covariance matrix $\boldsymbol{\Sigma}_m$. To obtain $\boldsymbol{\Sigma}_m$, we generate a matrix $\boldsymbol{S}$ according to Wishart distribution with d degrees of freedom, and we transform $\boldsymbol{S}$ in order to have one on the diagonal to keep the same margins for all chains (while remaining semi definite positive):

$$\boldsymbol{\Sigma}_m = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}, \quad \text{with} \quad \boldsymbol{D} = \operatorname{diag}(s_{1,1}, \ldots, s_{d,d}).$$

To illustrate the influence of the dependence direction (Section 3.3), a new matrix $\boldsymbol{\Sigma}_m$ is generated for each simulation, in order to have varying directions across replications.

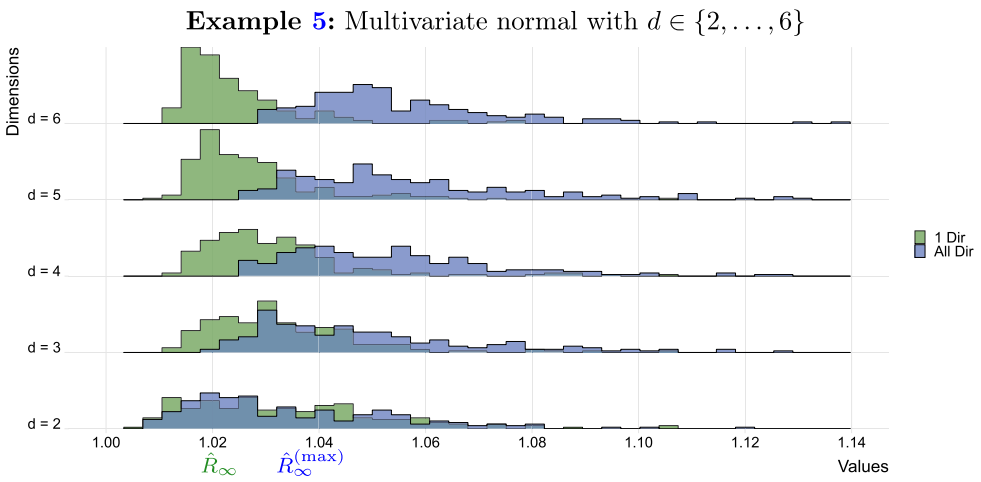

Figure 5: Comparison between $\hat{R}_\infty$ computed on one direction (in green), and $\hat{R}_\infty^{(\max)}$, the maximum of $\hat{R}_\infty$ computed on all possible indicator functions (in blue). For each $d \in \{2, \ldots, 6\}$, 200 replications are done where a new covariance matrix is generated for the normal distribution, which leads to different directions of dependence among the replications.

Then, we compare $\hat{R}_\infty$ with $\hat{R}_\infty^{(\max)}$, the maximum of $\hat{R}_\infty$ over all 2^{d-1} possible directions for the indicator functions.

Results are shown in Figure 5, where 200 replications are shown for $\hat{R}_\infty$ and $\hat{R}_\infty^{(\max)}$ for $d \in \{2, \ldots, 6\}$. As $\hat{R}_\infty^{(\max)}$ requires the computation of 2^{d-1} different $\hat{R}_\infty$, obtaining these histograms quickly becomes infeasible for larger dimensions. When $d = 2$, we can see that there is no significant difference between $\hat{R}_\infty$ and $\hat{R}_\infty^{(\max)}$, but as the dimension increases the values of $\hat{R}_\infty$ become more concentrated and closer to one. Indeed, as the number of possible directions increases exponentially, it is more and more rare to obtain the one to which $\hat{R}_\infty$ is sensitive. On the contrary, $\hat{R}_\infty^{(\max)}$ seems to stay robust with respect to this curse of dimensionality in terms of sensitivity, as the histograms look invariant when d increases.

4 Empirical results

In Section 2.5 and Section 3.4, we considered toy examples where the distribution of the chains is known in order to control the value of the population R_∞ and illustrate the robustness when other versions of $\hat{R}$ fail. Here we extend to other models in a more practical case for Bayesian inference. We adopt a baseline similar to the one used by Lambert and Vehtari (2021) to illustrate the behavior of $\hat{R}_\infty$ on Bayesian models, and add a multivariate example studied in Vats et al. (2019). For all examples in this section, we choose 4 chains and therefore a threshold $R_{\infty,\mathrm{lim}} = 1.02$ in the univariate

Example 6: Autoregressive model

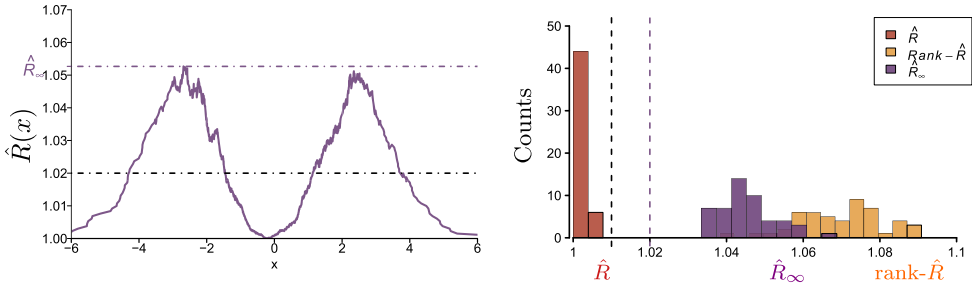

Figure 6: Behavior of $\hat{R}_\infty$ on the autoregressive example described in Section 4, with $m = 4$ chains of size $n = 500$ and $(\sigma, \sigma_m, \rho) = (1, 2, 1/2)$. On the left: $\hat{R}(x)$ as a function of x for one replication. On the right: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to thresholds of 1.01 and 1.02.

case (according to Section 2.3), and $(R^{(C)}_{\infty,\text{lim}}, R^{(M)}_{\infty,\text{lim}}) = (1.03, 1.03)$ in the multivariate one (according to Section 3.1). For each univariate study, we plot an example of $\hat{R}(x)$ as a function of x, and we recommend this illustration to users who want to analyse more carefully a given value of $\hat{R}_\infty$. Together with this figure, we also show histograms of replications to check the behavior of the different $\hat{R}$ more rigorously. All experiments are done on R using `rstan` library (Stan Development Team, 2021) and the package `localrhat` that we propose with this paper (Moins et al., 2022). Additional experiments have also been conducted on Python using OpenTURNS (Baudin et al., 2017). All the code concerning these experiments and the additional ones are available in the online appendix (link in the Introduction).

Example 6: Autoregressive model with different variances. The first example is a basic autoregressive model to study the behavior of $\hat{R}_\infty$ in the case of Markov chains with different variances: we consider m chains of size n such that for $i \in \{1, \dots, n-1\}$ and $j \in \{1, \dots, m\}$,

$$\theta^{(i+1,j)} = \rho\theta^{(i,j)} + \epsilon_{i,j}, \quad \text{with} \quad \epsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2),$$

where $\rho \in (0, 1)$ and $\sigma_j > 0$. In particular, assume that the first $m-1$ chains are generated using the same process: $\sigma_1 = \cdots = \sigma_{m-1} = \sigma$, while for the last chain $\sigma_m \neq \sigma$.

Results are illustrated in Figure 6 with $m = 4$, $\sigma = 1$, $\sigma_m = 2$ and $\rho = 1/2$ on 50 replications, and an example of $\hat{R}(x)$ as a function of x on the left panel. Similarly to the rank-$\hat{R}$ replications, the $\hat{R}_\infty$ values remain far from the threshold of 1.02 which confirms the sensitivity to this convergence defect. This corroborates in a more practical case the results of Example 1 in Section 2.5, on the sensitivity of $\hat{R}_\infty$ on chains with same mean and different variances. Note that the value $R(0) = 1$ is due to the fact that all the chains share the same median equal to zero.

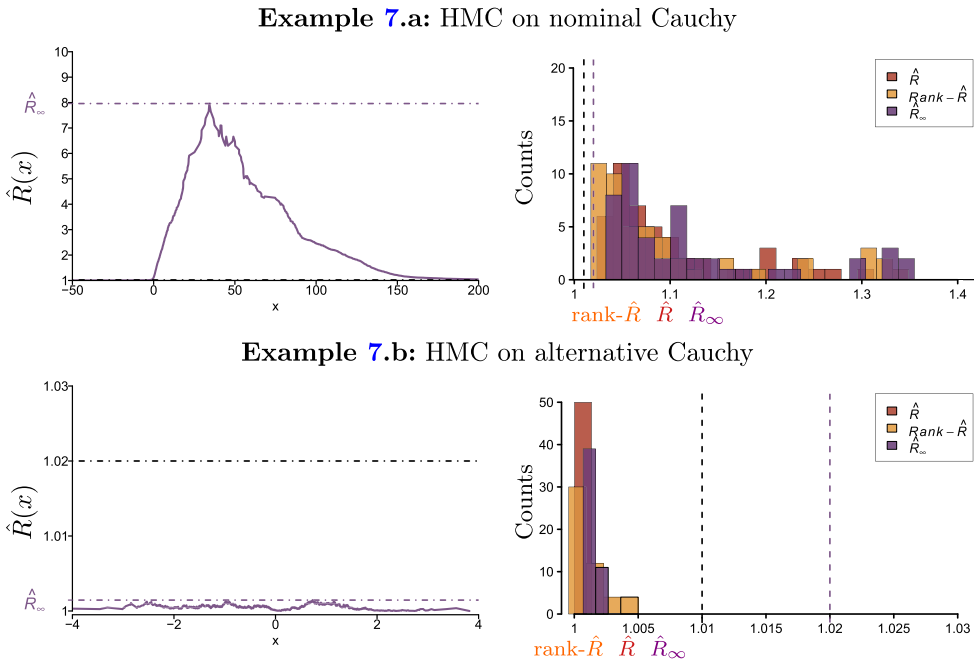

Figure 7: Behavior of $\hat{R}_\infty$ on the Cauchy example described in Section 4 for the two parameterisations. On the left: $\hat{R}(x)$ as a function of x for one replication. On the right: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to thresholds of 1.01 and 1.02.

Example 7: HMC on Cauchy distribution. As an extension of Example 2 in Section 2.5, we analyze the behavior of $\hat{R}_\infty$ in the case of heavy-tailed distributions. We run Hamiltonian Monte Carlo (HMC) (Neal, 2011) using Stan on Cauchy distributions for 50 variables. We consider the one with the most important mixing issue diagnosed with $\hat{R}_\infty$. Due to the tail heaviness of Cauchy distributions, the HMC iterations on a given chain can get trapped in a tail, which causes mixing issues. One solution to avoid this is to use an alternative parameterisation (Moins et al., 2023) that avoids sampling from a heavy-tailed distribution:

Example 7.a. Nominal parameterisation

$$x_j \sim \text{Cauchy}(0,1), \quad j \in \{1,\dots,50\}.$$

Example 7.b. Alternative parameterisation

$$x_j = a_j/\sqrt{b_j}, \quad a_j \sim \mathcal{N}(0,1), \quad b_j \sim \chi^2_1.$$

One would expect convergence issues with the nominal parameterisation and not with the alternative one. For both, the process of selecting the worst parameters among the

50 ones is iterated for the generation of replications, and results are shown in Figure 7. Histograms on the top right confirm the risk of diverging chains with the nominal parameterisation, as all the values are above 1.02 for all the versions of $\hat{R}$. This means that it is very likely to have at least one chain out of the 50 with a convergence issue in this experiment. This divergence can be really extreme, as it is shown on the top left panel where the value of $\hat{R}_\infty$ is over seven, due to a mixing issue in the right tail of the distribution. The opposite occurs with the other parameterisation, as all the convergence diagnostics indicate no mixing issues (see bottom row of Figure 7), which means no counter-indications that the chains for the 50 variables have converged. Looking at $\hat{R}(x)$ function on one replication in the bottom left panel, the curve seems to be very noisy and close to 1 compared to 1.02 (even sometimes less than 1) so the difference with 1 seems only due to Monte Carlo noise.

Example 8: Hierarchical Bayesian model on two parameterisations. As a classical Bayesian example, we consider using HMC on a hierarchical Bayesian model and in particular the eight-school (Gelman et al., 2013, Section 5.5), where two parameterisations are possible to model the problem:

Example 8.a. Centered parameterisation (CP)

$$\theta_j \sim \mathcal{N}(\mu, \tau), \quad y_j \sim \mathcal{N}(\theta_j, \sigma_j^2).$$

Example 8.b. Non-centered parameterisation (NCP)

$$\bar{\theta}_j \sim \mathcal{N}(0,1), \quad \theta_j = \mu + \tau\bar{\theta}_j, \quad y_j \sim \mathcal{N}(\theta_j, \sigma_j^2).$$

In the CP parameterisation, a prior dependence is between (μ, τ) and the population parameters θ_j, whereas in the other case (NCP), $\bar{\theta}_j$ is a priori independent of (μ, τ), and θ_j is just a function of $\bar{\theta}_j$ and (μ, τ) (see for example Papaspiliopoulos et al., 2003). Vehtari et al. (2021) argue in favor of the NCP for the eight-school example, by analysing the convergence of the chains associated with the parameter τ.

We also focus on computing $\hat{R}_\infty$ for τ: results and comparison with other versions of $\hat{R}$ are shown in Figure 8. In the first row, we can see that the $\hat{R}_\infty$ diagnostic confirms the one of rank-$\hat{R}$, as the two corresponding histograms are similar in the top right panel and conclude for a lack of convergence in most of the cases. However, for both diagnostics, a significant number of cases are also below 1.02 (respectively 1.01 for rank-$\hat{R}$), which is represented on the top left panel. In spite of this, the bottom row of Figure 8 shows a clear difference and NCP seems to help for chain convergence.

Example 9: Bayesian logistic regression. This example is related to the extension of $\hat{R}_\infty$ in the multivariate case as proposed in Section 3. As a multivariate Bayesian example, we run Stan on a basic hierarchical logistic model using the dataset `logit` available in the R package `mcmc`:

$$\boldsymbol{\beta} \sim \mathcal{N}(0, 0.35^2 \boldsymbol{I}_4), \quad y_j \sim \text{Bernoulli}\left(\frac{1}{1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}}}\right).$$

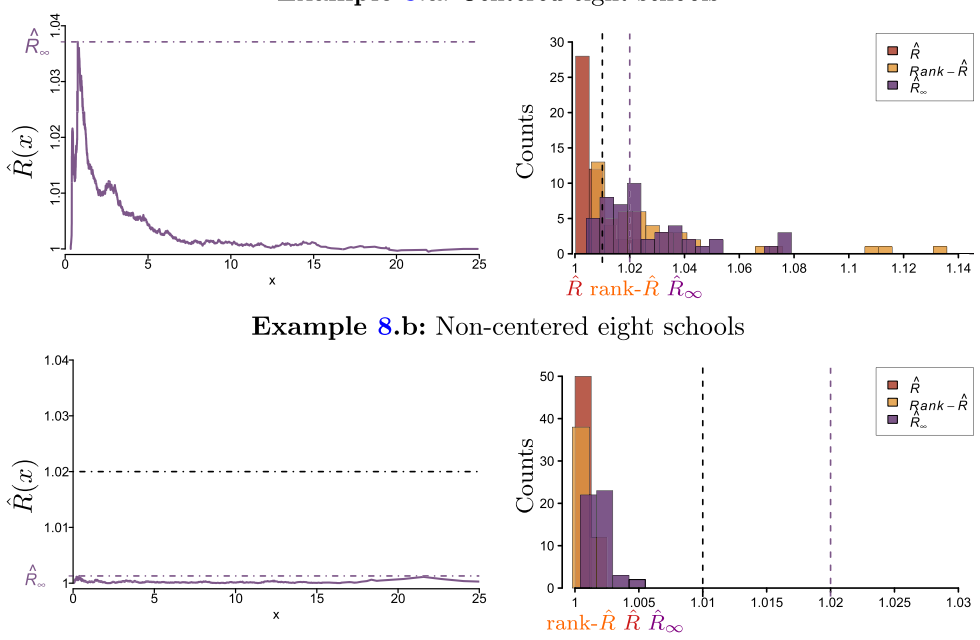

Figure 8: Behavior of $\hat{R}_\infty$ on the hierarchical example for τ described in Section 4 for the centered and non-centered version. On the left: $\hat{R}(x)$ as a function of x for one replication. On the right: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to thresholds of 1.01 and 1.02.

Here the posterior is intractable and Vats et al. (2019) showed that the posterior coefficients $\boldsymbol{\beta}$ could be significantly correlated, encouraging a multivariate diagnostic to check the convergence of the dependence structure. We run $m = 4$ chains each of size $n = 200$ after a burn-in of 100. In this configuration, despite a low number of iterations, all the different univariate $\hat{R}_\infty$ are mostly below 1.02 when replicated, and the rank-$\hat{R}$ are below 1.01.

When applied to the log posterior, the diagnostic is less clear and results are shown in the left panel of Figure 9: a significant part of the histogram for $\hat{R}_\infty$ is below the threshold, meaning that the number of iterations is almost sufficient but is not yet. Looking at the right plot of Figure 9, we notice in this example that the sensitivity of $\hat{R}_\infty^{(\max)}$ is approximately the same as the univariate version on the left, as the proportion of values over the threshold is similar (the choice of $R_{\infty,\mathrm{lim}}^{(C)} = 1.03$ is made according to Table 1). Although the computation of $\hat{R}_\infty^{(\max)}$ is possible here as the number of dimensions is small, computing a univariate $\hat{R}_\infty$ on the log posterior instead seems satisfactory here.

Example 9: Bayesian logistic regression

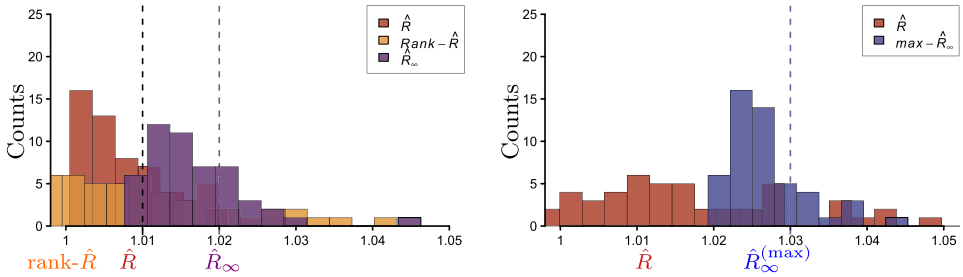

Figure 9: Behavior of multivariate and univariate $\hat{R}_\infty$ on the Bayesian logistic regression example, with $m = 4$ chains of size $n = 200$. On the left: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and univariate $\hat{R}_\infty$ all applied on the log-posterior. On the right: Histograms of 50 replications of Brooks–Gelman $\hat{R}$ and $\hat{R}_\infty^{(\max)}$. The dashed line corresponds to different thresholds: on the left, 1.01 in black for $\hat{R}$ and rank-$\hat{R}$, 1.02 in violet for $\hat{R}_\infty$, and on the right 1.03 in blue for $\hat{R}_\infty^{(\max)}$.

5 Discussion

In this paper we propose a new version of the Gelman–Rubin diagnostic called $\hat{R}_\infty$, which improves MCMC convergence diagnostics on several aspects. Firstly, it uses a localized version $\hat{R}(x)$ which assesses convergence at a given quantile x of the target distribution. Moreover, it is also based on a theoretical study of what $\hat{R}(x)$ is actually estimating: assuming stationarity to focus only on the mixing property, the population version can be seen as a distance measure between the distributions of the chains. This allows us to obtain convergence properties of $\hat{R}(x)$ and to tune the usual threshold of 1.01 (Section 2.3) based on a given confidence level and on the number of chains. We show theoretically (Section 2.5) and using experiments (Section 4) that our version is efficient to diagnose convergence. Finally, we suggest a two-step algorithm for a multivariate diagnosis (Section 3.1), and reinforce the second step to consider all the directions of the space, as we show that the natural extension cannot be used directly (Section 3.3). Therefore, in the high-dimensional case where this computation is likely to be too expensive, we suggest to replace it by a univariate calculation on the log-likelihood or the log-posterior. Diagnosing convergence in the multivariate case remains an open problem, and this is our hope that the local approach advocated here will trigger more research in this direction in the future.

Supplementary Material

Supplementary material for "On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic" (DOI: 10.1214/23-BA1399SUPP; .pdf). Proofs, details on calculations and additional experiments can be found on the supplementary material of the paper.

References

Baudin, M., Dutfoy, A., Iooss, B., and Popelin, A.-L. (2017). "OpenTURNS: An industrial software for uncertainty quantification in simulation." In R. Ghanem, D. H. and Owhadi, H. (eds.), *Handbook of Uncertainty Quantification*. Springer. 19

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). "Pyro: Deep universal probabilistic programming." *The Journal of Machine Learning Research*, 20(1): 973–978. 1

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. 4, 16

Brooks, S. P. and Roberts, G. O. (1999). "On Quantile Estimation and Markov Chain Monte Carlo Convergence." *Biometrika*, 86(3): 710–717. MR1723789. doi: https://doi.org/10.1093/biomet/86.3.710. 7

Carpenter, B., Gelman, A., Hoffman, M. D., 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*, 76(1): 1–32. 1, 16

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. 1

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. 7

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events*, volume 33. Springer Science & Business Media. MR1458613. doi: https://doi.org/10.1007/978-3-642-33483-2. 11

Fosdick, L. D. (1959). "Calculation of order parameters in a binary alloy by the Monte Carlo method." *Physical Review*, 116(3): 565. MR0110262. 8

Gelfand, A. E. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85(410): 398–409. MR1141740. 8

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press. MR3235677. 3, 8, 21

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7(4): 457–472. 2, 3, 5, 8

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(3): 684–700. MR3533633. doi: https://doi.org/10.1080/10618600.2015.1044092. 7

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *The Journal of Machine Learning Research*, 15(1): 1593–1623. MR3214779. 4

Lambert, B. and Vehtari, A. (2021). "R^*: A Robust MCMC Convergence Diagnostic with Uncertainty Using Decision Tree Classifiers." *Bayesian Analysis*, 1–27. MR4483223. doi: https://doi.org/10.1214/20-ba1252. 4, 18

Lopes, R. H., Reid, I., and Hobson, P. R. (2007). "The two-dimensional Kolmogorov-Smirnov test." In *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research*. Amsterdam, the Netherlands. 16

Margossian, C. C., Hoffman, M. D., Sountsov, P., Riou-Durand, L., Vehtari, A., and Gelman, A. (2022). "Nested $\hat{R}$: Assessing the convergence of Markov chain Monte Carlo when running many short chains." *arXiv:2110.13017*. 9

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. (2021). "Contributed discussion: "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC"." *Bayesian Analysis*, 16(2): 711–712. 4

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. (2022). *localrhat: a local $\widehat{R}$ to improve MCMC convergence diagnostic (R package)*. URL https://github.com/TheoMoins/localrhat. 5, 9, 19

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. (2023). "Supplementary material for "On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1399SUPP. 4

Moins, T., Arbel, J., Girard, S., and Dutfoy, A. (2023). "Reparameterization of extreme value framework for improved Bayesian workflow." *Computational Statistics and Data Analysis*, to appear. MR4610007. doi: https://doi.org/10.1016/j.csda.2023.107807. 20

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov Chain Monte Carlo*, 2(11): 2. MR2858447. 4, 20

Nelsen, R. B. (2006). *An introduction to copulas*. New York: Springer. MR2197664. doi: https://doi.org/10.1007/s11229-005-3715-x. 14

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2003). "Non-centered parameterisations for hierarchical models and data augmentation." In *Bayesian Statistics*, volume 7, 307–326. Oxford University Press, USA. MR2003180. 21

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling." Proceedings of the 3rd international workshop on distributed statistical computing, 124(10): 1–10. 8

Raftery, A. E. and Lewis, S. (1992). "How Many Iterations in the Gibbs Sampler?" *Bayesian Statistics*, 4: 763–773. 5, 7

Robert, C. P. (1995). "Convergence control methods for Markov chain Monte Carlo algorithms." *Statistical Science*, 10(3): 231–253. MR1390517. 2

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag. MR2080278. doi: https://doi.org/10.1007/978-1-4757-4145-2. 2, 7, 8

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. 1

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). "Probabilistic programming in Python using PyMC3." *PeerJ Computer Science*, 2: e55. 1

Stan Development Team (2021). "RStan: the R interface to Stan." R package version 2.21.3. 19

Vats, D. and Flegal, J. M. (2021). "Invited discussion: "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC"." *Bayesian Analysis*, 16(2): 695–701. 2

Vats, D., Flegal, J. M., and Jones, G. L. (2019). "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika*, 106(2): 321–337. MR3949306. doi: https://doi.org/10.1093/biomet/asz002. 4, 7, 18, 22

Vats, D. and Knudson, C. (2021). "Revisiting the Gelman–Rubin Diagnostic." *Statistical Science*, 36(4): 518 – 529. MR4323050. doi: https://doi.org/10.1214/20-sts812. 3, 5, 16

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 (with discussion)." *Bayesian Analysis*, 16(2): 667–718. MR4298989. doi: https://doi.org/10.1214/20-ba1221. 3, 4, 7, 8, 9, 10, 12, 21

Acknowledgments

We would like to thank a Reviewer and an Editor for providing us with valuable comments that helped us improving the manuscript. Specifically, comments from an Editor allowed us to deal with multi-stage testing in an appealing and satisfactory way.