On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic
Théo Moins, Julyan Arbel, Anne Dutfoy, Stéphane Girard

To cite this version:
Théo Moins, Julyan Arbel, Anne Dutfoy, Stéphane Girard. On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic. 2023. hal-03600407v2
On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic

Théo Moins * Julyan Arbel * Anne Dutfoy † Stéphane Girard *

March 7, 2023

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.

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 behaviour 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.

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 $i$th 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 $i$th 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 $\text{Var}[\theta]$, 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 $U(\mu - 2\sigma, \mu + 2\sigma)$, normal $\mathcal{N}(\mu, \tilde{\sigma}^2)$, or Laplace $\mathcal{L}(\mu, \sigma)$ distributions share the same mean (equal to $\mu$) and same mean over the median (equal to $\sigma$), 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. One illustration can also be found in the right column of Figure 2. 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 $R_\infty$. In addition to being more interpretable, $R_\infty$ 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, HMC [Neal, 2011], 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 $R_\infty$ 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 behaviour 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.

\(^1\)https://theomoins.github.io/localrhat/Simulations.html
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 $\text{Var}[\theta]$. The first one is the within-variance $\hat{W}$, which underestimates $\text{Var}[\theta]$ 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 $\text{Var}[\theta]$ 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 $\theta$ is sampled using $m$ chains, and let $Z \in \{1, \ldots, m\}$ be the corresponding index of the chain. Then,

$$\text{Var}[\theta] = \mathbb{E}_Z[\text{Var}[\theta | Z]] + \text{Var}_Z[\mathbb{E}[\theta | Z]].$$

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 $j \in \{1, \ldots, m\}$ 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 $\theta$ in the original Gelman–Rubin construction. The population within-chain and between-chain variances at point $x$ are then defined respectively as $W(x) = \mathbb{E}_Z[\text{Var}[I_x | Z]]$ and $B(x) = \text{Var}_Z[\mathbb{E}[I_x | Z]]$. Note that both quantities exist whatever the tail heaviness of $\theta$ 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 $\theta$ given $Z = j$ has cumulative distribution function (cdf) $F_j$. Then, one has for any $x \in \mathbb{R}$:

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

(2)
Thus, using $I_x$ instead of $\theta$ 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 $\mathbb{R}$:

$$R_\infty = \sup_{x \in \mathbb{R}} R(x).$$  \hspace{1cm} (3)

Note that, in view of Proposition 2.2(iv), $R_\infty$ is finite simply as soon as the $F_j$’s are continuous with overlapping supports. Considering $R_\infty$ 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 $\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\theta(i,j) \leq x\}$. This is equivalent to computing the original version of $\hat{R}$ on indicator variables $I_{x}^{(i,j)} = \mathbb{1}\{\theta^{(i,j)} \leq x\}$ 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 $R_\infty$.

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
\( \sqrt{n}(\hat{F}_j(x) - F(x)) \overset{d}{\to} N(0, \sigma^2(x)) \), \hspace{1cm} (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 \( \hat{F}(x) = \frac{1}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} I\{\theta^{(i,j)} \leq x\} = \frac{1}{m} \sum_{j=1}^{m} \hat{F}_j(x) \) and taking into account of the independence between chains yield
\( \sqrt{nm}(\hat{F}(x) - F(x)) \overset{d}{\to} N(0, \sigma^2(x)) \), \hspace{1cm} (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 estimator
\[ \hat{\text{ESS}} \]of variance from multiple sequences dates back to Gelman and Rubin [1992], Section 2.2, note that casting the problem of convergence monitoring in terms of analysing components of variance or ANOVA test statistics: \( \hat{\text{ESS}}(x) = \frac{F(x)(1 - F(x))}{\hat{\sigma}^2(x)} \).

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

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

(ii) For any \( x \in \mathbb{R} \), \( \text{ESS}(x)(\hat{R}^2(x) - 1) \overset{d}{\to} \chi^2_{m-1} \) 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 hypothesis \( \mathbb{E}(I_{x}^{(1,m)}) = \cdots = \mathbb{E}(I_{x}^{(1,m)}) \) 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 \( \text{ESS}(x) \) is replaced by a consistent estimator \( \hat{\text{ESS}}(x) \). 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 \( \chi^2_{m-1} \) distribution, and introduce the associated threshold

\[
R_{\text{lim},\alpha}(x) := \sqrt{1 + \frac{z_{m-1,1-\alpha}}{\text{ESS}(x)}}. \tag{7}
\]

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

### 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 \( \alpha \), using the definition of \( R_{\text{lim},\alpha}(x) \) in (7). Some values are displayed in the right panel of Table 1 for a fixed \( \text{ESS}(x) = 400 \) 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 \((m = 3)\) 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 \( \alpha \), \( m \) and a target value \( \text{ESS}(x) \) 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

<table>
<thead>
<tr>
<th>( m )</th>
<th>( R_{\text{lim},\alpha}(x) )</th>
<th>( \text{ESS}(x) )</th>
<th>( \alpha )</th>
<th>( m )</th>
<th>( R_{\text{lim},\alpha}(x) )</th>
<th>( \text{ESS} )</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.57</td>
<td></td>
<td></td>
<td>2</td>
<td>1.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.26</td>
<td></td>
<td></td>
<td>4</td>
<td>1.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>0.04</td>
<td></td>
<td></td>
<td>8</td>
<td>1.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>&lt; 10(^{-3})</td>
<td></td>
<td></td>
<td>15</td>
<td>1.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td>1.080</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>1.144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Left: Type I error \( \alpha \) as a function of \( \text{ESS}(x) \) when \( R_{\text{lim},\alpha}(x) = 1.01 \) and \( m = 4 \). Right: \( R_{\text{lim},\alpha}(x) \) as a function of \( m \) when \( \text{ESS}(x) = 400 \) and \( \alpha = 0.05 \).
Table 2: Left: Empirical quantiles $R_{\infty, \text{lim}}$ of the $\hat{R}_\infty$ distribution under the null hypothesis (i.e. all chains follow the same distribution) for a target effective sample size of 400, based on 2000 replications. This table is illustrated by the figure on the right.

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 $R_\infty$ 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 $\hat{R}_\infty$.

2.5 Illustrative examples

In this section, we consider toy distributions for the chains, where the computation of $R_\infty$ can be done explicitly. In particular, we first focus on two cases raised by Vehtari et al. [2021] of deficient behaviour of the traditional $\hat{R}$. Then, we exhibit a failure situation for rank-$\hat{R}$.

All these theoretical behaviours 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 behaviour 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 $U(-\sigma, \sigma)$ while the last chain has the cdf $F_m$ of the uniform distribution $U(-\sigma_m, \sigma_m)$ 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,
Example 1: \( U(-\sigma, \sigma) - U(-\sigma_m, \sigma_m) \)

Example 2: Pareto(\(\alpha\), \(\eta\)) - Pareto(\(\alpha\), \(\eta_m\))

Example 3: \( \text{Exp}(1) - U\left(1 - \frac{\lambda}{2}, 1 + \frac{\lambda}{2}\right) \)

Figure 2: 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).
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 $R_{\infty}$ is an increasing function of $\sigma_m/\sigma$ 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 2. In the bottom row, the histograms of replications confirm that $R_{\infty}$ 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 behaviour of $\hat{R}$, we consider chains following Pareto $(\alpha, \eta)$ distributions, with cdf

$$F(x | \alpha, \eta) = 1 - \left(\frac{x}{\eta}\right)^{-\alpha}, \quad \forall x \in [\eta, +\infty),$$

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: $F_1(x) = \cdots = F_{m-1}(x) = F(x | \alpha, \eta)$ and $F_m(x) = F(x | \alpha, \eta_m)$ 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, $R_{\infty}$ can be written as

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

see Lemma D.3 in the supplementary material. Clearly, $R_{\infty}$ is an increasing function of $\eta_m/\eta$ 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 2. 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 $\text{Exp}(1)$ and one uniform $U(1-2\log 2, 1+2\log 2)$. This results in chains with same mean and mean over the median. Results are illustrated in the right panel of Figure 2: 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 $R_{\infty}$ 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 $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ with $d \geq 2$, and denote by $\theta_p^{(j)}$ the coordinate
Let \( p \in \{1, \ldots, d\} \), with \( m \) values corresponding to chain \( j = 1, \ldots, m \). Similarly to the univariate case, \( \hat{R} \) can be computed on the indicator variables \( I^{(j)} = \mathbb{I}(\theta^{(j)}_1 \leq x_1, \ldots, \theta^{(j)}_d \leq x_d) \) for any \( x = (x_1, \ldots, x_d) \in \mathbb{R}^d \). Under the assumptions of Proposition 2.1, all calculations remain valid in dimension \( d \) and therefore the expression of \( R(x) \) is formally the same as in (2):

\[
R(x) = \sqrt{W(x) + B(x)} = \sqrt{1 + \frac{\sum_{j=1}^{m} \sum_{k=j+1}^{m} (F_j(x) - F_k(x))^2}{m \sum_{j=1}^{m} F_j(x)(1 - F_j(x))}}.
\]

The properties listed in Proposition 2.2 in the univariate case remain true as well. The associated \( R_{\infty} \) is defined as \( R_{\infty}(F_1, \ldots, F_m) = \sup_{x \in \mathbb{R}^d} R(x) \), while \( \hat{R}(x) \) 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 \( F_1, \ldots, F_m \) coincide, then \( R_{\infty} \) is the same as the one associated with uniform margins (see Lemma B.1 in the supplementary material). In other words, we have \( R_{\infty}(F_1, \ldots, F_m) = R_{\infty}(C_1, \ldots, C_m) \) where \( C_j \) is the copula defined in \([0, 1]^d\) associated with \( F_j, j \in \{1, \ldots, m\} \). 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 \( p \in \{1, \ldots, d\} \). If \( \hat{R}_{\infty, p} < R_{\infty, \text{lim}}^{(M)} \) for all \( p \in \{1, \ldots, d\} \), with \( R_{\infty, \text{lim}}^{(M)} \) 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 \( \hat{R}_{\infty} < R_{\infty, \text{lim}}^{(C)} \), with \( R_{\infty, \text{lim}}^{(C)} \) 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 \( \alpha \) 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 \( R_{\infty, \text{lim}}^{(M)} \) one can use the univariate threshold \( R_{\infty, \text{lim}}^{(M)} \) 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 \( R_{\infty, \text{lim}}^{(C)} \). Values of \( R_{\infty, \text{lim}}^{(M)} \) and \( R_{\infty, \text{lim}}^{(C)} \) are then given as functions of \( (\alpha, d, m) \) in Table C.3. As a general rule, one can reasonably use for \( \alpha = 0.05 \) the values \((R_{\infty, \text{lim}}^{(M)}, R_{\infty, \text{lim}}^{(C)}) = (1.03, 1.03)\) in the case of \( m = 4 \) chains, and \((R_{\infty, \text{lim}}^{(M)}, R_{\infty, \text{lim}}^{(C)}) = (1.04, 1.05)\) 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 \( \mathbf{u} = (u_1, \ldots, u_d) \in [0,1]^d \), one has

\[
R(\mathbf{u}) = \sqrt{1 + \frac{(C_1(\mathbf{u}) - C_2(\mathbf{u}))^2}{2(C_1(\mathbf{u})(1 - C_1(\mathbf{u})) + C_2(\mathbf{u})(1 - C_2(\mathbf{u})))}}. \tag{9}
\]

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:

\[
\forall \mathbf{u} \in [0,1]^d, \begin{cases}
C_-(\mathbf{u}) \leq C_1(\mathbf{u}) \leq C_+(\mathbf{u}), \\
C_-(\mathbf{u}) \leq C_2(\mathbf{u}) \leq C_+(\mathbf{u}).
\end{cases} \tag{10}
\]

Then, \( R_{\infty}(C_1, C_2) \leq R_{\infty}(C_-, C_+) \).

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

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

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 \( R_{\infty} \) associated with the convergence of the dependence structure is upper-bounded, with a bound that grows with the dimension. This difference of behaviour 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 \( R_{\infty} \)’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 $R_\infty$

When $m = 2$, one way to overcome this issue in the bivariate case is to compute two versions of the sensitivity of $R_\infty$ in the NLOD case than in the PLOD one. However, this observation is quickly inverted lower than the value $d$ due to the computation of $\mathbb{P}\{\theta_1^{(i)} \leq x_1, \ldots, \theta_d^{(i)} \leq x_d\}$. One way to overcome this issue in the bivariate case is to compute two versions of $R_\infty$.

Let us stress that positive and negative dependence are handled differently by $R_\infty$. 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 converges to $\sqrt{1 + \frac{1}{2(\frac{1}{2} - \frac{1}{d})}} \approx 1.136$ 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 $R_\infty$ strongly depends on the sign of dependence and asymptotically favours PLOD dependence when $d$ increases.

This difference can be explained by the construction of $R(x)$ itself (and thus $R_\infty$), which favours a dependence direction in $\mathbb{R}^d$ due to the computation of $\mathbb{P}\{\theta_1^{(i)} \leq x_1, \ldots, \theta_d^{(i)} \leq x_d\}$.
denoted respectively by \( R_∞^+ \) and \( R_∞^- \), based respectively on \( I\{\theta_1^{(i)} ≤ x_1, \theta_2^{(i)} ≤ x_2\} \) and \( I\{\theta_1^{(i)} ≤ x_1, \theta_2^{(i)} ≥ x_2\} \). Note that \( R_∞^- \) coincides with the construction proposed in Section 3.1.

**Corollary 3.4** Let \( m = 2 \). Then, \( R_∞^+(Π_2, M_2) = R_∞^-(W_2, Π_2) \) and \( R_∞^+(W_2, Π_2) = R_∞^-(Π_2, M_2) \).

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 \( \hat{R}_∞^{(max)} = \max(R_∞^+, R_∞^-) \). 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 \( \hat{R}_∞^{(max)} \) is still feasible for small values of \( d \): typically for \( d ≤ 6 \) we were able to replicate values in our experiments. Therefore, we provide in Table C.3 (Appendix C) the estimated threshold \( \hat{R}_∞^{(C)lim} \) associated with the maximum of \( \hat{R}_∞ \) in all possible directions when \( d ≤ 6 \).

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 \( l_\theta = \log p(y \mid \theta) \) when it is available, and compute \( \hat{R}_∞ \) with \( I\{\theta ≤ x\} \). 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 8.

### 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

\[
\theta^{(i,j)} \sim \mathcal{N}(0, \Sigma_j),
\]

\( i \in \{1, \ldots, n\} \) and \( j \in \{1, \ldots, m\} \), where \( \Sigma_j \) 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 \( \Sigma_j \). 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 3, with a comparison of \( \hat{R}_∞ \) 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 \( \hat{R}_∞ \). Due
Example 4: Bivariate normal Brooks–Gelman $\hat{R}$, $\hat{R}_\infty$, and $R_\infty$ wrt $\rho_m$

Figure 3: Behaviour 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 $\rho_m$, plotted as a function of $\rho_m$, and the corresponding population $R_\infty$ in blue.

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 either as the diagnostic is not based on a comparison between chains. This difference of behaviour is confirmed on the right panel of Figure 3, which illustrates 10 replications of both diagnostics as a function of $\rho_m$. 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 $\rho_m$, which is not satisfactory, contrary to $\hat{R}_\infty$ which has a parabolic shape.

As discussed in Section 3.3, the behaviour 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 3 on theoretical $R_\infty$ (in blue) and estimations $\hat{R}_\infty$ (in purple).

Example 5: Evolution of the behaviour 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 $\Sigma_m$. To obtain $\Sigma_m$, we generate a matrix $S$ according to Wishart distribution with $d$ degrees of freedom, and we transform $S$ in order to have one on the diagonal to keep the same margins for all chains (while remaining semi definite positive):

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

To illustrate the influence of the dependence direction (Section 3.3), a new matrix $\Sigma_m$ is generated for each simulation, in order to have varying directions across replications. Then,
Example 5: Multivariate normal with \( d \in \{2, \ldots, 6\} \)

Figure 4: Comparison between \( \hat{R}_\infty \) computed on one direction (in green), and \( \hat{R}_{\infty}^{(\text{max})} \), 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.

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

Results are shown in Figure 4, where 200 replications are shown for \( \hat{R}_\infty \) and \( \hat{R}_{\infty}^{(\text{max})} \) for \( d \in \{2, \ldots, 6\} \). As \( \hat{R}_{\infty}^{(\text{max})} \) 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 \( \hat{R}_{\infty}^{(\text{max})} \), 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, \( \hat{R}_{\infty}^{(\text{max})} \) 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 \( R_\infty \) 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 Vehtari et al. [2021] and Lambert and Vehtari [2021] to illustrate the behaviour 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 \( R_{\infty,\lim} = 1.02 \) in the univariate case (according to Section 2.3), and \( (R_{\infty,\lim}^{(C)}, R_{\infty,\lim}^{(M)}) = (1.03, 1.03) \) 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
**Example 6:** Autoregressive model

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 $i \in \{1, \ldots, n-1\}$ and $j \in \{1, \ldots, m\}$,

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

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 5 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 $\hat{R}(0) = 1$ is due to the fact that all the chains share the same median equal to zero.

**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 that avoids sampling from a heavy-tailed distribution:
Example 7.a: HMC on nominal Cauchy

Example 7.b: HMC on alternative Cauchy

Figure 6: Behaviour 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.a: Nominal parameterisation

\[ x_j \sim \text{Cauchy}(0, 1), \quad j \in \{1, \ldots, 50\}. \]

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 6. 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 6), 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 first parameterisation (CP), a prior dependence is between \((\mu, \tau)\) and the population parameters \(\theta_j\), whereas in the other case (NCP), \(\theta_j\) is a priori independent of \((\mu, \tau)\), and \(\theta_j\) is just a function of \(\bar{\theta}_j\) and \((\mu, \tau)\) [see for example Papaspiliopoulos et al., 2003, for more details]. Vehtari et al. [2021] argue in favour of the NCP for the eight-school example, by analysing the convergence of the chains associated with the parameter \(\tau\). We also focus on computing \(\hat{R}_\infty\) for \(\tau\): results for \(\hat{R}_\infty\) and comparison with other versions of \(\hat{R}\) are shown in Figure 7. 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
Example 9: Bayesian logistic regression

Figure 8: Behaviour 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 $\hat{R}^{(\max)}$. 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 $\hat{R}^{(\max)}$.

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:

$$\beta \sim N(0, 0.35^2I_4), \quad y_j \sim \text{Bernoulli} \left( \frac{1}{1 + e^{-x_j \beta}} \right).$$

Here the posterior is intractable and Vats et al. [2019] showed that the posterior coefficients $\beta$ 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 8: 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 8, we notice in this example that the sensitivity of $\hat{R}^{(\max)}$ is approximately the same as the univariate version on the left, as the proportion of values over the threshold is similar (the choice of $R^{(C)}_{\infty,\lim} = 1.03$ is made according to Table C.3). Although the computation of $\hat{R}^{(\max)}$ is possible here as the number of dimensions is small, computing a univariate $\hat{R}_\infty$ on the log posterior instead seems satisfactory here.
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.

Acknowledgement

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. S. Girard 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. J. Arbel acknowledges the support of the French National Research Agency (ANR-21-JSTM-0001).

Appendix A Construction of rank-$\hat{R}$ false negatives

In this section we detail different ways to construct false negative distributions on the chains that fool rank-$\hat{R}$, in the sense that they return rank-$\hat{R} \approx 1$ with non-identical distributions.

A.1 Two-parameter distributions

As mentioned in Section 1.2, one way to fool rank-$\hat{R}$ is to fix two constraints on the distributions, which are identical mean and mean over the median on all chains. Many two-parameter distributions can be tuned to respect these constraints. Consider for example a uniform chain $U(-2\sigma, 2\sigma)$ and another one from a Laplace distribution $\mathcal{L}(0, \sigma)$, where $\sigma > 0$. The resulting $R_\infty$ does not depend on the scale parameter $\sigma$ and is given by

$$R_\infty = \sqrt{1 + \frac{1}{2e^2 - 1}} \approx 1.018,$$

see Lemma D.4. Note that computations are done in the case of $m = 2$ chains, and so the traditional threshold of 1.01 holds for $\hat{R}_\infty$ (see Table 2). Therefore we expect $\hat{R}_\infty$ to diagnose
**Example A.9:** Laplace $\mathcal{L}(0, \sigma)$ and $\mathcal{U}(-2\sigma, 2\sigma)$

![Figure A.9: Behaviour of $\hat{R}_\infty$ on the case of $m = 2$ chains, $n = 500$ each where $F_1$ is a Laplace distribution $\mathcal{L}(0, 1/4)$ and $F_2$ is $\mathcal{U}(-1/2; 1/2)$. On the left: $\hat{R}(x)$ as a function of $x$ for one simulation. On the right: Histograms of 500 replications of $\hat{R}$, rank-$\hat{R}$ of and $\hat{R}_\infty$. The dashed line corresponds to the threshold of 1.01 which holds here for all versions of $\hat{R}$.](image)

convergence, and not $\hat{R}$ nor rank-$\hat{R}$. Results are provided in Figure A.9 and confirm this hypothesis.

### A.2 General framework with Generalized Pareto Distribution

To find a general way to construct counter-examples, let us consider the Generalised Pareto Distribution (GPD), parametrised by $(\mu, \sigma, \xi)$:

$$
F_{\text{GPD}}(x) = \begin{cases} 
1 - \left\{ 1 + \frac{\xi}{\sigma} \left( \frac{x - \mu}{\sigma} \right) \right\}_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\
1 - \exp \left( - \frac{x - \mu}{\sigma} \right) & \text{if } \xi = 0,
\end{cases}
$$

with $\{ \cdot \}_+ = \max\{0, \cdot\}$. The support of this distribution depends on the parameters: $[\mu; +\infty)$ if $\xi \geq 0$, $[\mu; \mu - \sigma/\xi)$ otherwise. The expectation exists only if $\xi < 1$ and is equal to $\mu + \frac{\sigma}{\xi}$, and the median is given by $x_{\text{med}} = \mu + \frac{\sigma}{\xi} \xi$.

An interesting property here is that conditioned on exceeding a value, the distribution is still a GPD distribution: If $X \sim \text{GPD}(\mu, \sigma, \xi)$ then $X \mid X > u \sim \text{GPD}(u, \sigma, \xi)$ with $\sigma = \sigma + \xi (u - \mu)$. Therefore, $X \mid X > x_{\text{med}} \sim \text{GPD}(u, \sigma, \xi)$, and

$$
E(X \mid X > x_{\text{med}}) = x_{\text{med}} + \frac{\sigma + \xi (x_{\text{med}} - \mu)}{1 - \xi},
$$

$$
= \mu + \frac{\sigma}{1 - \xi} \left( 1 + \frac{2\xi - 1}{\xi} \right),
$$

$$
= E(X) + \sigma \frac{(2\xi - 1)}{\xi(1 - \xi)}.
$$

Then, by considering $(m - 1)$ chains that follow a GPD$(\mu_1, \sigma_1, \xi_1)$, and one that follows a GPD$(\mu_2, \sigma_2, \xi_2)$, we can solve the system of two equations that links the two means and the two means over the medians, to obtain a range of possible parameters:

$$
\begin{align*}
\mu_1 + \frac{\sigma_1}{1 - \xi_1} &= \mu_2 + \frac{\sigma_2}{1 - \xi_2}, \\
\sigma_1 \frac{(2\xi_1 - 1)}{\xi_1(1 - \xi_1)} &= \sigma_2 \frac{(2\xi_2 - 1)}{\xi_2(1 - \xi_2)}.
\end{align*}
$$

(11)
In order to obtain different values of parameters, we should choose $\xi_1 \neq \xi_2$ and $\sigma_1 \neq \sigma_2$.

One way to characterize the set of solutions is as follows:

1. Fix $\xi_1$ and $\xi_2$ such that $\xi_1 \neq \xi_2$, and define $\lambda = \frac{f(\xi_1)}{f(\xi_2)}$ with $f(\xi) = \frac{(2^\xi - 1)}{\xi(1-\xi)}$.

2. Fix $\sigma_1$, and using the second equation of (11) define $\sigma_2 = \lambda \sigma_1$.

3. Finally, choose $\mu_1$ and $\mu_2$ such that $\mu_1 - \mu_2 = \sigma_1(\frac{\lambda}{1-\xi_2} - \frac{1}{1-\xi_1})$.

An example of a solution is as follows:

1. Choose $(\mu_1, \sigma_1, \xi_1) = (0, 1, 0)$, the standard exponential distribution Exp(1), and $\xi_2 = -1$, a uniform distribution.

2. Following the method described before, we obtain $\lambda = \frac{f(\xi_1)}{f(\xi_2)} = 4 \log(2)$, so $\sigma_2 = \lambda \sigma_1 = 4 \log(2)$.

3. Following the last point, we set $\mu_2 = \mu_1 - \sigma_1(\frac{\lambda}{1-\xi_2} - \frac{1}{1-\xi_1}) = 1 - 2 \log(2)$.

To conclude, if some chains follow an Exp(1) distribution and the other ones a $U(1 - 2 \log(2); 1 + 2 \log(2))$ distribution, the difference between the chains should not be detected by the rank-$\hat{R}$, which is illustrated in the last column of Figure 2. Similarly to the other examples in Section 2.5, $\hat{R}_\infty$ manages to diagnose the convergence issue contrary to other versions.
Appendix B  Proofs

B.1 Proofs in the univariate case

Proof of Proposition 2.1. Let $x \in \mathbb{R}$. The within- and between-variances are given by:

$$ W(x) = \mathbb{E}[\text{Var}[I_x | Z]] $$

$$ = \frac{1}{m} \sum_{j=1}^{m} (\mathbb{E} \left[ I_x^2 | Z = j \right] - \mathbb{E}^2 \left[ I_x | Z = j \right]) \quad (12) $$

$$ = \frac{1}{m} \sum_{j=1}^{m} (\mathbb{P}(\theta \leq x | Z = j) - \mathbb{P}^2(\theta \leq x | Z = j)) $$

$$ = \frac{1}{m} \sum_{j=1}^{m} (F_j(x) - F_j^2(x)), \quad (13) $$

$$ B(x) = \text{Var}[\mathbb{E}[I_x | Z]] = \mathbb{E} \left[ (F_Z(x))^2 \right] - \mathbb{E}[F_Z(x)]^2 $$

$$ = \frac{1}{m} \sum_{j=1}^{m} F_j^2(x) - \left( \frac{1}{m} \sum_{j=1}^{m} F_j(x) \right)^2 \quad (14) $$

$$ = \frac{m - 1}{m^2} \sum_{j=1}^{m} F_j^2(x) - \frac{2}{m^2} \sum_{j<k} F_j(x) F_k(x) $$

$$ = \frac{1}{m^2} \sum_{j<k} (F_j^2(x) + F_k^2(x) - 2F_j(x) F_k(x)) $$

$$ = \frac{1}{m^2} \sum_{j<k} (F_j(x) - F_k(x))^2, \quad (15) $$

where (12) and (14) are a consequence of $\mathbb{P}(Z = j) = 1/m$ for all $j \in \{1, \cdots, m\}$. The conclusion follows.

Proof of Proposition 2.2. Proofs of (i) and (ii) are straightforward, let us focus on (iii) and (iv).

(iii): Without loss of generality, we assume that all $F_j$’s are defined on $\mathbb{R}$ and we denote $\bar{F}_j = 1 - F_j$ the associated survival function. From Proposition 2.1, we can write, for any $x \in \mathbb{R}$:

$$ R^2(x) = \frac{\left( \frac{1}{m} \sum_{j=1}^{m} F_j(x) \right) \left( \frac{1}{m} \sum_{j=1}^{m} \bar{F}_j(x) \right)}{\frac{1}{m} \sum_{j=1}^{m} F_j(x) \bar{F}_j(x)}. $$

Denote by $f(x) \simeq g(x)$ two functions $f$ and $g$ that are asymptotically equivalent when $x \to a$, $a \in \mathbb{R} \cup \{-\infty, +\infty\}$. Consider first the case $x \to -\infty$. Clearly, for all $j \in \{1, \ldots, m\}$,
$F_j(x) \to 0$ and $\tilde{F}_j(x) \to 1$ as $x \to -\infty$, so that

$$
\left( \frac{1}{m} \sum_{j=1}^{m} F_j(x) \right) \left( \frac{1}{m} \sum_{j=1}^{m} \tilde{F}_j(x) \right) \simeq \frac{1}{m} \sum_{j=1}^{m} F_j(x),
$$

and

$$
\frac{1}{m} \sum_{j=1}^{m} \tilde{F}_j(x) F_j(x) \simeq \frac{1}{m} \sum_{j=1}^{m} F_j(x),
$$

leading to $R(x) \to 1$ as $x \to -\infty$. Second, remarking that $R(x)$ is symmetric with respect to $(F_j(x), \tilde{F}_j(x))$, we similarly have $R(x) \to 1$ as $x \to \infty$.

(iv): It is clear that $B(x)$ and $W(x)$ are continuous since the $F_j$’s are continuous, so the only thing to prove is that the denominator $W(x)$ never vanishes, except if extending by continuity is possible. Remark that for all $x$, we have $F_j(x)(1 - F_j(x)) \geq 0$, $W(x) = 0$ if and only if $F_j(x) = 0$ or $F_j(x) = 1$ for $j \in \{1, \ldots, m\}$. Almost all combinations are avoided by the assumption that the supports must overlap, except $F_1(x) = \cdots = F_m(x) = 0$ and $\tilde{F}_1(x) = \cdots = \tilde{F}_m(x) = 1$. In these two latter cases, an extension by continuity is possible as $R(x) = 1$ from (iii).

**Proof of Proposition 2.3.**  
(i): Assume for the sake of simplicity that $F$ is continuous and strictly increasing. Let us show that the corresponding $\hat{R}_\infty$ has the same distribution as in the standard uniform case. In view of (2), $\hat{R}_\infty$ can be written as

$$
\hat{R}_\infty = \sup_{x \in \mathbb{R}} \hat{R}(x) = \sup_{x \in \mathbb{R}} \left[ 1 + \frac{\sum_{j=1}^{m} \sum_{k=j+1}^{m} (\hat{F}_j(x) - \hat{F}_k(x))^2}{m \sum_{j=1}^{m} \hat{F}_j(x)(1 - \hat{F}_j(x))} \right],
$$

where for any $j \in \{1, \ldots, m\}$,

$$
\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\theta^{(i,j)} \leq x\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{F(\theta^{(i,j)}) \leq F(x)\}},
$$

and where the random variables $F(\theta^{(i,j)})$ are standard uniformly distributed for any $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Finally, observing that

$$
\sup_{x \in \mathbb{R}} \hat{R}(x) = \sup_{y \in [0,1]} \hat{R}(F^{-1}(y))
$$

concludes the proof.

(ii): Let $x \in \mathbb{R}$ and remark that (4) implies that $\hat{F}_j(x) \xrightarrow{P} F(x)$ as $n \to \infty$ for all $j \in \{1, \ldots, m\}$. Then it follows from (13) in the proof of Proposition 2.1 that

$$
\hat{W}(x) = \frac{1}{m} \sum_{j=1}^{m} \hat{F}_j(x)(1 - \hat{F}_j(x)) = F(x)(1 - F(x)) + o_p(1).
$$

Besides, from (4) and (6), for $j = 1, \ldots, m$, $\hat{F}_j(x)$ can be expanded as

$$
\hat{F}_j(x) = F(x) + \sqrt{\frac{F(x)(1 - F(x))}{\text{ESS}(x)/m}} \xi_{j,n},
$$

26
with \( \xi_{j,n} \xrightarrow{d} \mathcal{N}(0,1) \) as \( n \to \infty \). Thus (15) in the proof of Proposition 2.1 entails
\[
\hat{B}(x) = \frac{1}{m^2} \sum_{j<k} \left( \hat{F}_j(x) - \hat{F}_k(x) \right)^2 = \frac{F(x)(1 - F(x))}{\text{ESS}(x)} \times \frac{1}{m} \sum_{j<k} (\xi_{j,n} - \xi_{k,n})^2.
\tag{17}
\]
Introducing the random vector \( \xi_n = (\xi_{1,n}, \ldots, \xi_{m,n})^\top \xrightarrow{d} \mathcal{N}(0, I_m) \) where \( I_m \) is the \( m \times m \) identity matrix, one has
\[
\frac{1}{m} \sum_{j<k} (\xi_{j,n} - \xi_{k,n})^2 = \frac{m-1}{m} \sum_{j=1}^m \xi_{j,n}^2 - \frac{1}{m} \sum_{j \neq k} \xi_{j,n} \xi_{k,n} = \xi_n^\top A \xi_n,
\]
with \( A = I_m - J_m/m \), and where \( J_m \) is the \( m \times m \) matrix filled with ones. The symmetric matrix \( A \) can be eigen-decomposed as \( A = Q^\top \Lambda Q \) with \( Q \) an orthogonal matrix and \( \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d) = \text{diag}(1, \ldots, 1, 0) \). Remark that \( U_n := Q \xi_n \xrightarrow{d} \mathcal{N}(0, I_m) \) so that
\[
\frac{1}{m} \sum_{j<k} (\xi_{j,n} - \xi_{k,n})^2 = (Q \xi_n)^\top \Lambda (Q \xi_n) = \sum_{j=1}^m \lambda_j U_{j,n}^2 = \sum_{j=1}^m U_{j,n}^2 \xrightarrow{d} \chi^2_{m-1},
\tag{18}
\]
as \( n \to \infty \). Collecting (16), (17) and (18) yields
\[
\text{ESS}(x)(\hat{R}^2(x) - 1) = \text{ESS}(x) \frac{\hat{B}(x)}{W(x)} \xrightarrow{d} \chi^2_{m-1},
\]
as \( n \to \infty \) and the result is proved.

### B.2 Proofs in the multivariate case

**Lemma B.1 (Standardization of margins)** Assume the assumptions of Proposition 2.1 hold. If the margins of \( F_1, \ldots, F_m \) coincide, then the multivariate \( R_\infty \) is the same as the one calculated on the associated copulas \( C_1, \ldots, C_m \).

**Proof.** Denote by \( \phi_1, \ldots, \phi_d \) the common margins of the cdfs \( F_1, \ldots, F_m \), so that \( F_j(x) = C_j(\phi_1(x_1), \ldots, \phi_d(x_d)) \) for any \( j \in \{1, \ldots, m\} \). Letting \( y = (\phi_1(x_1), \ldots, \phi_d(x_d)) \), we have
\[
\sup_{x \in \mathbb{R}^d} R(x) = \sup_{y \in [0,1]^d} R(\phi_1^{-1}(y_1), \ldots, \phi_d^{-1}(y_d)).
\]
Besides, in view of (8), \( m(R^2(\phi_1^{-1}(y_1), \ldots, \phi_d^{-1}(y_d)) - 1) \) can be written as
\[
\frac{\sum_{j=1}^m \sum_{k=j+1}^m (F_j(\phi_1^{-1}(y_1), \ldots, \phi_d^{-1}(y_d)) - F_k(\phi_1^{-1}(y_1), \ldots, \phi_d^{-1}(y_d)))^2}{\sum_{j=1}^m F_j(\phi_1^{-1}(y_1), \ldots, \phi_d^{-1}(y_d))(1 - F_j(\phi_1^{-1}(y_1), \ldots, \phi_d^{-1}(y_d)))}
\]
\[
= \frac{\sum_{j=1}^m \sum_{k=j+1}^m (C_j(y_1, \ldots, y_d) - C_k(y_1, \ldots, y_d))^2}{\sum_{j=1}^m C_j(y_1, \ldots, y_d)(1 - C_j(y_1, \ldots, y_d))},
\]
which proves the result.
Proof of Lemma 3.1. Let function $\hat{R}$ be defined on $[0,1]^2$ by

$$\hat{R}(c_1, c_2) = \sqrt{1 + \frac{(c_1 - c_2)^2}{2c_1(1-c_1) + c_2(1-c_2)}}.$$  

(19)

For any $c_1 \in [0,1]$, $c_2 \mapsto \hat{R}(c_1, c_2)$ is decreasing on $[0,c_1]$, and increasing on $[c_1,1]$. Let $(c_-, c_+) \in [0,1]^2$ and $(c_1, c_2) \in [c_-, c_+]$. Without loss of generality, assume that $c_1 \leq c_2$. Let $c \in [0,c_2]$. Since $c_2 \leq c_+$ and $\hat{R}(c, \cdot)$ is increasing on $[c,1]$, we have $\hat{R}(c, c_2) \leq \hat{R}(c, c_+)$. In particular, for $c = c_-$:

$$\hat{R}(c_-, c_2) \leq \hat{R}(c_-, c_+).$$  

(20)

Moreover, $\hat{R}(c_2, \cdot)$ is decreasing on $[0,c_2]$ and $c_- \leq c_1 \leq c_2$, we also have

$$\hat{R}(c_-, c_2) = \hat{R}(c_2, c_-) \geq \hat{R}(c_2, c_1).$$  

(21)

Combining (20) and (21), we finally obtain $\hat{R}(c_2, c_1) \leq \hat{R}(c_-, c_+)$, which concludes the proof.

Proof of Proposition 3.1. By definition of lower and upper Fréchet–Hoeffding bounds [Nelsen, 2006], $W_2(u) \leq C(u) \leq M_2(u)$ for any copula $C$ and $u \in [0,1]^d$. Thus in view of Lemma 3.1 it only remains to prove that $R_{\infty}(W_2, M_2) = (d + 1)/2$. To this end, let $p$ be the index such that $u_p = \min \{u_1, \ldots, u_d\}$. Two cases arise:

(i) If $1 - d + \sum_{i=1}^d u_i \leq 0$, then

$$f(u) := 2(R^2(u) - 1) = \frac{u_p^2}{u_p(1-u_p)} = \frac{1}{1/u_p - 1}.$$  

(22)

As a consequence, $R^2(u)$ is maximum when $u_p$ is maximum under the constraints

$$\begin{cases}
    u_p \leq u_i \quad \forall i \neq p, \\
    u_p \leq d - 1 - \sum_{i \neq p} u_i.
\end{cases}$$

It is easily seen that the maximum occurs in the equality case $u_1 = \cdots = u_d = (d - 1)/d$ and thus $2(R^2(u) - 1) = d - 1$.

(ii) Conversely, if $1 - d + \sum_{i=1}^d u_i \geq 0$, then

$$f(u) = \frac{(1 - d + \sum_{i \neq p} u_i)^2}{u_p(1-u_p) + (1-d + \sum_{i=1}^d u_i)(d - \sum_{i=1}^d u_i)}.$$  

The problem therefore amounts to maximising $f(u)$ under the constraints

$$\begin{cases}
    u \in [0,1]^d, \\
    u_p \leq u_i \quad \forall i \neq p, \\
    1 - d + \sum_{i=1}^d u_i \geq 0.
\end{cases}$$  

(23)

The associated Lagrangian can be written with $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}^d$ and $\lambda_4 \in \mathbb{R}$ as:

$$\mathcal{L}(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = f(u) + \sum_{i=1}^d \lambda_{1,i} u_i + \sum_{i=1}^d \lambda_{2,i}(1-u_i) + \sum_{i \neq p} \lambda_{3,i}(u_i-u_p) + \lambda_4 \left(1 - d + \sum_{i=1}^d u_i\right).$$

28
The first-order conditions are given by (23) and \( \nabla_u \mathcal{L}(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = 0 \), and the Karush–Kuhn–Tucker conditions are

\[
\begin{align*}
\lambda_{1,i} & \geq 0 \quad \text{with} \quad \lambda_{1,i} u_i = 0, \quad \forall i \in \{1, \ldots, d\}, \\
\lambda_{2,i} & \geq 0 \quad \text{with} \quad \lambda_{2,i}(1 - u_i) = 0, \quad \forall i \in \{1, \ldots, d\}, \\
\lambda_{3,i} & \geq 0 \quad \text{with} \quad \lambda_{3,i}(u_i - u_p) = 0, \quad \forall i \in \{1, \ldots, d\} \quad \text{s.t.} \quad i \neq p, \\
\lambda_4 & \geq 0 \quad \text{with} \quad \lambda_4 \left(1 - d + \sum_{i=1}^d u_i\right) = 0.
\end{align*}
\]

We distinguish different cases:

- If \( \lambda_{1,i_0} \neq 0 \) for some \( i_0 \in \{1, \ldots, d\} \), then \( u_{i_0} = 0 \) and combined with (23) we obtain necessarily \( i_0 = p \) and \( 1 - d + \sum_{i=1}^d u_i = 0 \), leading to a non-optimal solution for \( f(u) \).

- If \( \lambda_{2,i_0} \neq 0 \) for some \( i_0 \in \{1, \ldots, d\} \), then \( u_{i_0} = 1 \), and the problem is exactly the same written in dimension \( d - 1 \), and a recurrence proves that the maximum is equal to \( \sqrt{(d+1)/2} \) in dimension \( d \). So the maximum is increasing with the dimension and therefore is not reached in this case.

One can thus assume \( \lambda_1 = 0 \) and \( \lambda_2 = 0 \). Moreover, for all \( (i, j) \) such that \( i \neq p \) and \( j \neq p \), we have \( \frac{\partial f}{\partial u_i} = \frac{\partial f}{\partial y} \). Combined with \( \nabla_u \mathcal{L} = 0 \), we obtain \( \lambda_{3,i} = \lambda_{3,j} \), which leads to considering the simplified Lagrangian:

\[
\mathcal{L}(u, \lambda_3, \lambda_4) = f(u) + \lambda_3 \left(\sum_{i=1}^d u_i - d u_p\right) + \lambda_4 \left(1 - d + \sum_{i=1}^d u_i\right).
\]

In that form, the function \( f \) and the constraints of the Lagrangian can be written only as a function of \( (1 - d + \sum_{i=1}^d u_i, u_p, \lambda_3, \lambda_4) \). Let \( \tilde{\mathcal{L}}(x = 1 - d + \sum_{i=1}^d u_i, y = u_p, \lambda_3, \lambda_4) = \mathcal{L}(u, \lambda_3, \lambda_4) \). Since for all \( i \)

\[
\frac{\partial \tilde{\mathcal{L}}}{\partial u_i} = \frac{\partial \tilde{\mathcal{L}}}{\partial x} \frac{\partial x}{\partial u_i} + \frac{\partial \tilde{\mathcal{L}}}{\partial y} \frac{\partial y}{\partial u_i} = \frac{\partial \tilde{\mathcal{L}}}{\partial x} + \frac{\partial \tilde{\mathcal{L}}}{\partial y} \sum_{i=1}^d u_i \{ i = p \},
\]

solving \( \nabla_u \mathcal{L} = 0 \) is equivalent to solving \( \nabla_{(x,y)} \tilde{\mathcal{L}} = 0 \). The corresponding problem is therefore

\[
\max \frac{(x - y)^2}{x(1 - x) + y(1 - y)}, \quad \text{under the constraints} \quad \begin{cases} x \geq \max\{0, 1 - d + dy\}, \\ x \leq y \leq 1. \end{cases}
\]

(25)

Combining the constraints with the study of function in the proof of Lemma 3.1 leads to the solution \( (x, y) = (0, \frac{d-1}{d}) \). So

\[
\sum_{i=1}^d u_i = d - 1 \quad \text{and} \quad \min\{u_1, \ldots, u_d\} = \frac{d - 1}{d},
\]

so necessarily \( u_1 = \cdots = u_d = \frac{d-1}{d} \), which concludes the proof.
Proof of Corollary 3.1 For all \( u \in [0,1]^d \), we have

\[
m( R^2(u) - 1 ) = \sum_{j=1}^{m} \sum_{k=j+1}^{m} \frac{(C_j(u) - C_k(u))^2}{\sum_{\ell=1}^{m} C_\ell(u)(1 - C_\ell(u))},
\]

so that an upper bound on the multivariate \( R_\infty \) can be expressed thanks to bivariate as follows:

\[
m( R_\infty^2(C_1, \ldots, C_m) - 1 ) \leq \sum_{j=1}^{m} \sum_{k=j+1}^{m} 2 \left( R_\infty^2(C_j, C_k) - 1 \right),
\]

using Lemma 3.1 and Proposition 3.1. The result is thus proved.

Proof of Corollary 3.2. One can prove that the maximum of \( u \in [0,1]^d \mapsto R(\Pi_d(u), M_d(u)) \)

is reached at \( u_1 = \cdots = u_d := u \in [0,1] \), which leads to studying the maximum of

\[
f_d(u) := 2( R^2(u, \ldots, u) - 1 ) = \frac{(u^d - u)^2}{u^d(1 - u^d) + u(1 - u)}.
\]

The first derivative of \( f_d \) is proportional to

\[
g_d(u) = -2(d-1)u^{2d-1} + du^{2d-2} - 2(d-1)u^d + 3(d-1)u^{d-1} - 1.
\]

Routine calculations show the existence of a unique root in \([0,1]\), but finding the explicit value does not seem possible when \( d > 2 \) since \( g_d \) is a polynomial of order \( 2d - 1 \). We thus restrict ourselves to an asymptotic analysis when \( d \to \infty \). First, Lemma 3.1 and Proposition 3.1 entail that

\[
\max_{u \in [0,1]} f_d(u) \leq 2( R_\infty^2(W_d, M_d) - 1 ) = d - 1.
\]

Second, a lower bound can be obtained by letting \( u_d = 1 - (\log d)/d \). Indeed, \( u_d \to 1 \) and \( u_d^d = \exp(-\log d(1 + o(1))) \to 0 \) so that the numerator of \( f_d(u) \) satisfies \((u_d^d - u_d)^2 \to 1\) as \( d \to \infty \). Moreover, the denominator satisfies

\[
u_d^d(1 - u_d^d) + u_d(1 - u_d) = \exp(-\log d(1 + o(1)))(1 + o(1)) + \frac{\log d}{d}(1 + o(1)) = \frac{\log d}{d}(1 + o(1))
\]

as \( d \to \infty \). As a consequence, \( f_d(u_d) = \frac{d}{\log d}(1 + o(1)) \) and we have proved that

\[
\frac{d}{\log d}(1 + o(1)) \leq \max_{u \in [0,1]} f_d(u) \leq d - 1.
\]

The result follows.
Proof of Corollary 3.3. From Proposition 3.1 and the definition of a NLOD copula, the proof reduces to calculating $R_{\infty}(W_d, \Pi_d)$. Two cases are considered.

(i) First, if $1 - d + \sum_{i=1}^{d} u_i \leq 0$, then $W_d(u) = 0$ and
\[
2(R^2(u) - 1) = \frac{1}{\prod_{i=1}^{d} u_i} - 1.
\]
The maximisation of $2(R^2(u) - 1)$ is then equivalent to solving:
\[
\max \prod_{i=1}^{d} u_i, \quad \text{under the constraints} \quad \left\{ \begin{array}{l}
u \in [0, 1]^d, \\ 1 - d + \sum_{i=1}^{d} u_i \leq 0. \end{array} \right.
\]
(26)
Since the constraints are linear and the objective function is convex, the above optimization problem is convex. The Lagrangian associated with (26) can be written with $\lambda_1 \in \mathbb{R}$, $\lambda_2, \lambda_3 \in \mathbb{R}^d$, as:
\[
\mathcal{L}(u, \lambda_1, \lambda_2, \lambda_3) = \prod_{i=1}^{d} u_i - \lambda_1 \left(1 - d + \sum_{i=1}^{d} u_i\right) - \sum_{i=1}^{d} \lambda_{2,i}(u_i - 1) + \sum_{i=1}^{d} \lambda_{3,i} u_i,
\]
The first-order conditions are
\[
\left\{ \begin{array}{l}
\nabla_u \mathcal{L}(u, \lambda_1, \lambda_2, \lambda_3) = 0, \\
0 \leq u_i \leq 1 \quad \forall i \in \{1, \ldots, d\}, \\
1 - d + \sum_{i=1}^{d} u_i \leq 0, \end{array} \right.
\]
(27)
and the Karush–Kuhn–Tucker conditions are
\[
\left\{ \begin{array}{l}
\lambda_1 \geq 0 \quad \text{with} \quad \lambda_1(1 - d + \sum_{i=1}^{d} u_i) = 0, \\
\lambda_{2,i} \geq 0 \quad \text{with} \quad \lambda_{2,i}(u_i - 1) = 0, \quad \forall i \in \{1, \ldots, d\}, \\
\lambda_{3,i} \geq 0 \quad \text{with} \quad \lambda_{3,i} u_i = 0, \quad \forall i \in \{1, \ldots, d\}. \end{array} \right.
\]
(28)
If there exists $i_0$ such that $u_{i_0} = 0$, then $\prod_{i=1}^{d} u_i = 0$, which is clearly non-optimal. Thus, (28) implies $\lambda_{3,i} = 0$ for all $i \in \{1, \ldots, d\}$. Moreover, for all $i \in \{1, \ldots, d\}$,
\[
\frac{\partial \mathcal{L}}{\partial u_i} = 0 \quad \Rightarrow \quad \lambda_1 + \lambda_{2,i} = \prod_{j \neq i} u_j,
\]
so that, for all $(j, k) \in \{1, \ldots, d\}^2$,
\[
(\lambda_1 + \lambda_{2,j})u_j = (\lambda_1 + \lambda_{2,k})u_k.
\]
(29)
If there exists $i_0$ such that $\lambda_{2,i_0} \neq 0$, then $u_{i_0} = 1$ from (28) and consequently, for all $k \neq i_0$,
\[
u_k = \frac{\lambda_1 + \lambda_{2,i_0}}{\lambda_1 + \lambda_{2,k}}.
\]
Taking account of (27) yields $u_k \leq 1$ which implies in turn $\lambda_{2,k} \geq \lambda_{2,i_0} \neq 0$ and $u_k = 1$ for all $k \in \{1, \ldots, p\}$ from (28). The resulting $u$ does not fulfill the third constraint in (27).
Then, necessarily, \( \lambda_{2,i} = 0 \) for all \( i \in \{1, \ldots, d\} \), and combining with (29), it follows that the optimum is reached when \( u_1 = \cdots = u_d = u \). Replacing in (26) yields the optimization problem

\[
\max u^d, \quad \text{under the constraints}\quad \begin{cases}
0 \leq u \leq 1, \\
1 + d(u - 1) \leq 0,
\end{cases}
\]  

whose solution is \( u = (d - 1)/d \).

(ii) Second, if \( 1 - d + \sum_{i=1}^d u_i \geq 0 \), the problem to solve is

\[
\max f_d(u) := \frac{\left(\prod_{i=1}^d u_i - 1 + d - \sum_{i=1}^d u_i\right)^2}{\left(1 - \prod_{i=1}^d u_i\right) \left(1 - d + \sum_{i=1}^d u_i\right) \left(d - \sum_{i=1}^d u_i\right)},
\]

under the constraints \( \begin{cases} u \in [0,1]^d, \\
1 - d + \sum_{i=1}^d u_i \geq 0. \end{cases} \)

The Lagrangian can be written, with \( \lambda_1 \in \mathbb{R}, \lambda_2, \lambda_3 \in \mathbb{R}^d \):

\[
\mathcal{L}(u, \lambda_1, \lambda_2, \lambda_3) = f_d(u) + \lambda_1 \left(1 - d + \sum_{i=1}^d u_i\right) + \sum_{i=1}^d \lambda_{2,i}(1 - u_i) + \sum_{i=1}^d \lambda_{3,i} u_i.
\]

In the same way as in the proof of Proposition 3.1, we can focus on the solution such that \( \lambda_2 = 0 \) and \( \lambda_3 = 0 \). The first order condition \( \nabla_u \mathcal{L} = 0 \) leads to the solution

\[
u_j = \prod_{i=1}^d u_i \times \frac{2 \lambda_1(1 - d + \sum_{i=1}^d u_i)}{2 \lambda_1 + \lambda_1 - \lambda_1(1 - d + \sum_{i=1}^d u_i)(d - \sum_{i=1}^d u_i)} - \lambda_1,
\]

for \( j \in \{1, \ldots, d\} \) with \( \lambda_1 = \prod_{i=1}^d u_i (1 - \prod_{i=1}^d u_i) + (1 - d + \sum_{i=1}^d u_i)(d - \sum_{i=1}^d u_i) \). Note that from this expression, the maximum verifies \( u_1 = \cdots = u_d \), and so the initial \( d \)-dimensional optimization problem amounts to the one-dimensional problem:

\[
\max f_d(u) := \frac{(u^d - 1 + d - du)^2}{u^d(1 - u^d) + (1 - d + du)(d - du)}, \quad \text{under the constraints}\quad \begin{cases} u \in [0,1]^d, \\
u \geq \frac{d-1}{d}. \end{cases}
\]

Iterated derivative computations allow to show that \( f_d \) is a decreasing function on \( [\frac{d-1}{d},1] \), so the maximum is reached at \( u = \frac{d-1}{d} \), which concludes the proof.

**Proof of Corollary 3.4.** Let us denote by \( R(\mathbb{I}\{\theta_1^{(c)} \leq x_1, \theta_2^{(c)} \leq x_2\}) \) and \( R(\mathbb{I}\{\theta_1^{(c)} \leq x_1, \theta_2^{(c)} \geq x_2\}) \) the two considered versions of \( R(x) \) in the bivariate case, with standard uniform margins. We clearly have:

\[
R^- = \max_{(x_1, x_2) \in [0,1]^2} R \left( \mathbb{I}\{\theta_1^{(c)} \leq x_1, \theta_2^{(c)} \geq x_2\} \right) = \max_{(x_1, x_2) \in [0,1]^2} R \left( \mathbb{I}\{\theta_1^{(c)} \leq x_1, 1 - \theta_2^{(c)} \leq 1 - x_2\} \right) = \max_{(x_1, x_2) \in [0,1]^2} R \left( \mathbb{I}\{\theta_1^{(c)} \leq x_1, 1 - \theta_2^{(c)} \leq x_2\} \right).
\]

32
Then, remarking that

\[
(\theta_1^{(\cdot)}, \theta_2^{(\cdot)}) \sim M_2 \implies (\theta_1^{(\cdot)}, 1 - \theta_2^{(\cdot)}) \sim W_2,
\]

\[
(\theta_1^{(\cdot)}, \theta_2^{(\cdot)}) \sim W_2 \implies (\theta_1^{(\cdot)}, 1 - \theta_2^{(\cdot)}) \sim M_2,
\]

\[
(\theta_1^{(\cdot)}, \theta_2^{(\cdot)}) \sim \Pi_2 \implies (\theta_1^{(\cdot)}, 1 - \theta_2^{(\cdot)}) \sim \Pi_2
\]

proves the result.

Appendix C  Threshold estimation for \( \hat{R}_\infty \)

This section details the computation of the empirical quantiles of \( \hat{R}_\infty \) that is done to obtain Table 2.

C.1 Univariate case

Invariance on the underlying distribution under the null hypothesis (all chains have the same distribution). A primary step to compute the quantiles of \( \hat{R}_\infty \) when all the chain distributions are identical is to verify that such quantiles are well-defined, in the sense that they do not depend on the choice of chain distribution. This property is expected as using a supremum over the quantiles provides invariance to bijective transformations (see Proposition 2.3(i)). The first row of Figure C.10 illustrates the behaviour of \( \hat{R}, \text{rank-}\hat{R}, \) and \( \hat{R}_\infty \) on two cases:

- all chains are uniform;
- all chains are Pareto distributed.

The QQ-plots seem to confirm the invariance in distribution of rank-\( \hat{R} \) and \( \hat{R}_\infty \) for various choices of \( m \) (see the yellow and violet dots in the first row of Figure C.10). This was also expected for rank-\( \hat{R} \) because of the use of a rank-normalization step. For the traditional \( \hat{R} \), the QQ-plots show a difference of distribution (see the red dots in the first row of Figure C.10), which makes the quantile estimation ill-defined if the chains distribution is not provided. The rest of Figure C.10 shows histograms of replications for \( \hat{R} \) (second row), rank-\( \hat{R} \) (third row) and \( \hat{R}_\infty \) (fourth row), when the chains are distributed according to the same uniform and Pareto distributions. Histograms for \( \hat{R} \) confirm the difference of behaviour when the chain distribution is uniform or Pareto, while the histograms overlap much more for rank-\( \hat{R} \) and \( \hat{R}_\infty \).

Threshold elicitation. As a direct consequence of Proposition 2.3(i), the empirical quantiles of \( \hat{R}_\infty \) only depend on the number of chains \( m \) and the length \( n \). The same holds true for rank-\( \hat{R} \). Therefore, we are able to estimate empirical quantiles with replications of \( \hat{R}_\infty \) using any chain distribution (typically uniform); results are reported in Table 2 for a fixed value of \( mn = 400 \), which is the effective sample size as we are in the i.i.d case. The same estimation can be done for rank-\( \hat{R} \), in order to associate a choice of threshold to a confidence level \( 1 - \alpha \). Consequently, this viewpoint implies a threshold that depends on \( m \) for rank-\( \hat{R} \) too, as the empirical quantiles change with \( m \) (third row of Figure C.10).
Figure C.10: Study of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$ when all the distributions are the same, for a number of chains $m \in \{2, 4, 8\}$. First row: Q-Q plot that compares 500 replications computed on $m$ uniform chains $U(-1, 1)$ with $m$ Pareto($\alpha = 0.8, \eta = 1$) ones. Second, third and fourth rows: Histograms of 500 replications with $n \in \{200, 100, 50\}$ for $\hat{R}$ (second row), rank-$\hat{R}$ (third row), and $\hat{R}_\infty$ (fourth row) in the case of uniform and Pareto chains. The dashed lines represent the suggested thresholds for rank-$\hat{R}$ and $\hat{R}_\infty$, corresponding to a confidence level of approximately 95%.
Although Vehtari et al. [2021] do not suggest any tuning with respect to the number of chains, using a threshold of 1.01 can lead to large type I error: for example, a threshold of 1.01 leads to approximately 5% when \( m = 2 \), but to 21% when \( m = 8 \) (third row of Figure C.10). Therefore a threshold of 1.01 seems too strong for rank-\( \hat{R} \) when \( m \) increases. The empirical quantile at \( \alpha = 0.05 \) suggests to use 1.01 when \( m = 2 \), 1.013 when \( m = 4 \) and 1.018 when \( m = 8 \). The same observation holds for \( \hat{R}_\infty \), and looking at Table 2 and the fourth row of Figure C.10 leads to a threshold of 1.01 for \( m = 2 \), 1.02 for \( m = 4 \) and 1.03 for \( m = 8 \) in order to keep a type I error of approximately 5%.

C.2 Multivariate case

In the multivariate extension (see Section 3), two thresholds have to be elicited: \( R_{\infty, \text{lim}}^{(M)} \) for the convergence of margins, and \( R_{\infty, \text{lim}}^{(C)} \) for the convergence of the copula. Focusing on \( R_{\infty, \text{lim}}^{(M)} \), the quantiles of \( \hat{R}_\infty \) for the margins are the same as in the univariate case (given in Table 2), but a Bonferroni correction is necessary to take into account the multiplicity of tests. For the convergence of margins, and \( \hat{R}_\infty \) for the copula, \( \hat{R}_\infty^{(\max)} \) is a maximum of multiple versions of \( \hat{R}_\infty \) computed on different directions of dependence, so its quantiles are different from \( \hat{R}_\infty \) ones. However, to reduce the calculation cost, one can compute the quantiles of \( \hat{R}_\infty \) under the null hypothesis for the chains, with a Bonferroni correction with \( 2^{d-1} \) hypotheses. We estimate the corresponding quantiles using replications to determine the two thresholds \( R_{\infty, \text{lim}}^{(M)} \) and \( R_{\infty, \text{lim}}^{(C)} \). Here, the computation is done for several values of \( d \), with \( d \) relatively small. Results are reported in Table C.3. Values in bold confirm the rule of thumb given in Section 3: \((R_{\infty, \text{lim}}^{(M)}), R_{\infty, \text{lim}}^{(C)}) = (1.03, 1.03)\) for \( m = 4 \) and \((R_{\infty, \text{lim}}^{(M)}), R_{\infty, \text{lim}}^{(C)}) = (1.04, 1.05)\) for \( m = 8 \).

Appendix D Examples of closed-form \( R(x) \) and \( R_\infty \)

We start by a lemma providing useful tools to simplify the calculation of \( R(x) \) and \( R_\infty \) in the univariate case and when all chains but one have converged. We then review families of distributions for which \( R(x) \) and \( R_\infty \) can be computed in closed-form.

Lemma D.1 Assume the assumptions of Proposition 2.1 hold with \( F := F_1 = \cdots = F_{m-1} \neq F_m \).

(i) Then (2) can be simplified as

\[
R(x) = \left[ \frac{1 + \frac{(m-1)(F(x) - F_m(x))^2}{m ((m-1)F(x)(1 - F(x)) + F_m(x)(1 - F_m(x)) '}}}{\right]^{1/2}.
\]

(ii) If \( F \) and \( F_m \) are symmetrical distributions wrt 0, then \( R \) is an even function.

(iii) Let \( a \) and \( a_m \in \mathbb{R} \cup \{-\infty\} \) be the starting points of \( F \) and \( F_m \) respectively, and assume \( a \leq a_m \). Then, \( R(\cdot) \) reaches its supremum on \([a_m, \infty)\):

\[
R_\infty \geq \sqrt{1 + \frac{F(a_m)}{m(1 - F(a_m)))}}.
\]
Table C.3: Left: Empirical quantiles of $\hat{R}_\infty$ for the margins with a Bonferroni correction, under the null hypothesis that all chains have the same distribution, for a fixed value of $mn = 400$. Right: Empirical quantiles of $\hat{R}_{\infty, \text{lim}}^{(C)}$ for the copula, for a fixed value of $mn = 400$. We have used 500 replications for estimation. Values in bold justify the rule of thumb proposed in Section 3.1 for $m \in \{4, 8\}$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$R_{\infty, \text{lim}}^{(M)}$</th>
<th>$R_{\infty, \text{lim}}^{(C)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d$</td>
<td>$m$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.017</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.024</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.030</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>1.041</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.022</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.032</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.032</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1.045</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.025</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.027</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.027</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>1.047</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.025</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.027</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.030</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>1.053</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.022</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.024</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.030</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1.047</td>
</tr>
</tbody>
</table>
(iv) Let \( b \) and \( b_m \in \mathbb{R} \cup \{+\infty\} \) be the endpoints of \( F \) and \( F_m \) respectively, with \( b \leq b_m \). Then, \( R(\cdot) \) reaches its supremum on \((-\infty, b]\) and

\[
R_\infty \geq \sqrt{1 + \frac{(m-1)(1-F_m(b))}{mF_m(b)}},
\]

**Proof.** (i) and (ii) are straightforward. For (iii) remark that, when \( x \leq a \), \( F(x) = F_m(x) = 0 \) so that \( R(x) = 1 \). Besides, for all \( x \in [a, a_m] \), one has \( F_m(x) = 0 \) and thus

\[
R^2(x) = 1 + \frac{F(x)}{m(1-F(x))}.
\]

The above defined function is increasing so that the supremum of \( R^2(\cdot) \) is reached for \( x \geq a_m \) and therefore \( R_\infty \geq R(a_m) \). Similarly for (iv), when \( x \geq b_m \), \( F(x) = F_m(x) = 1 \) so that \( R(x) = 1 \). Besides, for all \( x \in [b, b_m] \), one has \( F(x) = 0 \) and thus

\[
R^2(x) = 1 + \frac{(m-1)(1-F_m(x))}{mF_m(x)}.
\]

The above defined function is decreasing so that the supremum of \( R(\cdot) \) is reached for \( x \leq b \) and therefore \( R_\infty \geq R(b) \).

**Lemma D.2 (Uniform distribution)** Assume that \( F_1 = \cdots = F_{m-1} \) are the cdf of the uniform distribution \( \mathcal{U}(-\sigma, \sigma) \) while \( F_m \) is the cdf of the uniform distribution \( \mathcal{U}(-\sigma_m, \sigma_m) \) with \( 0 < \sigma \leq \sigma_m \). Then,

\[
R^2(x) = \begin{cases} 
1 + \frac{(1 - \frac{1}{2\sigma_m})^2}{\frac{1}{m} - \frac{1}{2\sigma} + \frac{1}{m} - \frac{1}{2\sigma_m}} & \text{if } |x| \leq \sigma, \\
1 + \frac{m-1}{m} \left(1 - \frac{2}{1+\sigma_m/|x|}\right) & \text{if } \sigma \leq |x| \leq \sigma_m, \\
1 & \text{if } |x| \geq \sigma_m.
\end{cases}
\]

Moreover,

\[
R_\infty = R(\pm \sigma) = \sqrt{1 + \frac{m-1}{m} \left(1 - \frac{2}{1+\sigma_m}\right)}.
\]

**Proof.** Recall that

\[
F_1(x) = \cdots = F_{m-1}(x) = \frac{x}{2\sigma} + \frac{1}{2}, \quad \forall x \in [-\sigma; \sigma],
\]

and \( F_m(x) = \frac{x}{2\sigma_m} + \frac{1}{2}, \quad \forall x \in [-\sigma_m; \sigma_m] \).

The case \( |x| \geq \sigma_m \) is clear, we investigate the two other ones. First, if \( \sigma \leq x \leq \sigma_m \), then \( F_j(x) = 1 \) for \( j = 1, \ldots, m-1 \) and Lemma D.1(i) yields

\[
R^2(x) = 1 + \frac{(m-1)(1 - \frac{x}{2\sigma_m})^2}{m \left(\frac{1}{2} + \frac{x}{2\sigma_m}\right) \left(\frac{1}{2} - \frac{x}{2\sigma_m}\right)} = 1 + \frac{m-1}{m} \left(1 - \frac{2}{1 + \frac{\sigma_m}{x}}\right).
\]
Using, Lemma D.1(ii), allows concluding for $\sigma \leq |x| \leq \sigma_m$. Finally, if $|x| \leq \sigma$, then:

$$R^2(x) = 1 + \frac{x^2(\frac{1}{2\sigma} - \frac{1}{2\sigma_m})^2}{m \left( \frac{1}{2} + \frac{x}{2\sigma} \right) \left( \frac{1}{2} - \frac{x}{2\sigma} \right) + \frac{m-1}{m} \left( \frac{1}{2} + \frac{x}{2\sigma_m} \right) \left( \frac{1}{2} - \frac{x}{2\sigma_m} \right)}$$

$$= 1 + \frac{\left( \frac{1}{\sigma} - \frac{1}{\sigma_m} \right)^2}{(m-1)x^2 - m \left( \frac{1}{\sigma^2} + \frac{1}{(m-1)\sigma_m^2} \right)}.$$ 

Lemma D.1(iv) entails that the maximum is located at $x = \pm \sigma$, which gives the result.

**Lemma D.3 (Pareto distribution)** Assume that $F_1 = \cdots = F_{m-1}$ are the cdf of the Pareto($\alpha, \eta$) distribution with $\alpha > 0$ the shape parameter and $\eta > 0$ the position parameter. Let $F_m$ be the cdf of the Pareto($\alpha, \eta_m$) distribution with $0 < \eta_m \leq \eta$. Then,

$$R^2(x) = \begin{cases} 
1 + \frac{1}{m} \left( \left( \frac{x}{\eta} \right)^\alpha - 1 \right) & \text{if } \eta \leq x \leq \eta_m, \\
1 + \frac{1}{m} \left( \frac{\eta - \eta_m}{\eta} - \frac{\eta - \eta_m}{\eta} \right)^\alpha & \text{if } \eta_m \leq x, \\
1 & \text{if } x \leq \eta.
\end{cases}$$

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

**Proof.** Recall that $F_1(x) = \cdots = F_{m-1}(x) = 1 - (x/\eta)^{-\alpha}$, $\forall x \in [\eta, +\infty)$ and $F_m(x) = 1 - (x/\eta_m)^{-\alpha}$, $\forall x \in [\eta_m, +\infty)$. In the case where $\eta \leq x \leq \eta_m$, $F_m(x) = 0$, and using Lemma D.1(iii) entails that $R(\cdot)$ is increasing on $[\eta, \eta_m]$ and

$$R^2(x) = 1 + \frac{1}{m} \frac{F(x)}{1 - F(x)} = 1 + \frac{1}{m} \left( \left( \frac{x}{\eta} \right)^\alpha - 1 \right).$$

Moreover, for $\eta_m \leq x$, replacing the Pareto cdf in (32) yields

$$R^2(x) = 1 + \frac{1}{m} \frac{x^{-2\alpha} (\eta^\alpha - \eta_m^\alpha)^2}{x^{-\alpha} (1 - x^{-\alpha} \eta^\alpha) + \frac{1}{m-1} x^{-\alpha} \eta_m^\alpha (1 - x^{-\alpha} \eta_m^\alpha)}$$

$$= 1 + \frac{1}{m} \left( \eta^\alpha - \eta_m^\alpha \right)^2 \left( \eta^\alpha + \frac{\eta_m^\alpha}{m-1} \right)^{-\alpha}.$$ 

Clearly, $R^2(\cdot)$ is decreasing on $[\eta_m, +\infty)$ and is extended by continuity at $x = \eta_m$. In conclusion, $R^2(\cdot)$ is maximum at $x = \eta_m$, and the result is proved.

**Lemma D.4 (Uniform vs Laplace distribution)** Assume that $m = 2$, $F_1$ is the cdf of the uniform distribution $\mathcal{U}(-\sigma, \sigma)$ and $F_2$ is the cdf of the centred Laplace distribution $\mathcal{L}(0, \sigma/2)$ with $\sigma > 0$. Then for any $x$, $R(x) = R_1(x/\sigma)$ with

$$R_1^2(x) = \begin{cases} 
1 + \frac{\exp(-|x|)}{2 - \exp(-|x|)} & \text{if } |x| \geq 2, \\
1 + \frac{(|x|/2 + \exp(-|x|))^2}{2 - \exp(-|x|)} & \text{if } |x| \leq 2.
\end{cases}$$
Moreover, \( R_\infty = R(\pm \sigma) = R_1(\pm 1) = \sqrt{1 + \frac{1}{2e^2 - 1}} \).

**Proof.** In view of Lemma D.1(ii,iii), it is sufficient to compare the values of \( R_1(x) \) on \((-2, 0]\) with \( R_1(-2) \). Then, the derivation of \( R_1(x) \) is similar to the ones done in Lemma D.2 and Lemma D.3. Routine calculations show that \( R_1 \) has indeed a local maximum on \((-2, 0]\), but it remains lower than \( R_1(-2) \) (see last column of Figure 2 for an illustration), which is therefore the value of \( R_\infty \).

**References**


