A local version of R-hat for MCMC convergence diagnostic

Théo Moins, Julyan Arbel, Anne Dutfoy, Stéphane Girard

A local version of R-hat for MCMC convergence diagnostic

Théo Moins[1], Julyan Arbel[1], Anne Dutfoy[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
{theo.moins, julyan.arbel, stephane.girard}@inria.fr
[2] *EDF R&D dept. Périclès, 91120 Palaiseau, France*
anne.dutfoy@edf.fr

Résumé. Diagnostiquer la convergence des chaînes de Markov est un enjeu crucial pour les méthodes de Monte Carlo par chaîne de Markov. Parmi les méthodes les plus populaires, le diagnostique de Gelman–Rubin, communément appelé $\hat{R}$, est un indicateur qui permet de vérifier la convergence en se basant sur une comparaison des variances inter- et intra-chaîne. Nous proposons ici une version localisée $\hat{R}(x)$ qui se concentre sur un quantile x de la distribution, et analysons certaines propriétés de la valeur théorique associée $R(x)$. Ceci conduit à proposer un nouvel indicateur $\hat{R}_\infty$, qui permet à la fois de localiser la convergence de la chaîne en différents quantiles de la distribution, et en même temps de traiter certains problèmes de convergence non détectés par d'autres versions de $\hat{R}$.

Mots-clés. Chaînes de Markov, Statistiques bayésienne, Diagnostique de convergence

Abstract. Diagnosing convergence of Markov chain is crucial for Markov Chain Monte Carlo methods and remains an essentially unsolved problem. Among the most popular methods, the Gelman–Rubin potential scale reduction factor, commonly named $\hat{R}$, is an indicator that monitors the convergence of output chains to a stationary distribution, based on a comparison of the between- and within-variances of the chains. Here, we propose a localized version $\hat{R}(x)$ that focuses on quantiles of the distribution and we analyse some properties of the associated population value $R(x)$. This 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 distribution, and at the same time for handling some convergence issues not detected by other $\hat{R}$ versions.

Keywords. Markov chains, Bayesian statistics, Convergence diagnostic

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. The fundamental idea behind those 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.

A property that is frequently mentioned to verify chain convergence is mixing (see Vats and Flegal (2021) for a discussion). It 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.

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

The use of the original version of Gelman and Rubin (1992) has some limitations, which can be found for example in Vats and Knudson (2021); Vehtari et al. (2021): lack of interpretability, lack of robustness for certain types of non-convergence, arbitrary choice of a threshold, etc. We hope to take a step forward in addressing these limitations with a localized version of $\hat{R}$ introduced in Moins et al. (2021) and developed here and in Moins et al. (2022): 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 suggest a new indicator $\hat{R}_\infty$, which in addition to being more interpretable, shows better results in terms of MCMC convergence diagnostic.

2 A local version of $\hat{R}$

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. The idea of our local convergence estimate is decidedly simple: we use I_x in place of θ in the original Gelman–Rubin construction. If $Z \in \{1, \ldots, m\}$ denotes the corresponding index of the chain, the population within-chain and between-chain variances at point x are then defined respectively as $W(x) = \mathbb{E}\left[\mathrm{Var}[I_x \mid Z]\right]$ and $B(x) = \mathrm{Var}\left[\mathbb{E}[I_x \mid Z]\right]$. Note that both quantities exist whatever the tail heaviness of θ distribution thanks to introduction of the indicator function, thus relaxing moment conditions of the original $\hat{R}$. We define the associated population $R(x)$

as

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

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

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

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

Thus, using I_x instead of θ defines a local convergence estimate at any point x which quantifies a distance between the F_j's. This allows for diagnostic convergence relatively to a quantile one wants to estimate (for a posterior credible interval for example).

In order to summarize this continuous index into a scalar one, we may also consider its supremum over $\mathbb{R}$:

$$R_\infty = \sup_{x \in \mathbb{R}} R(x). \tag{2}$$

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

Sample version. Population version $R(x)$ can be estimated by replacing $F_j(x)$ in (1) by its empirical counterpart $\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\theta^{(i,j)} \leq x\}$. This is equivalent to computing the original version of $\hat{R}$ on indicator variables $I_x^{(i,j)} = \mathbb{I}\{\theta^{(i,j)} \leq x\}$ instead of $\theta^{(i,j)}$. This connects with the Raftery–Lewis diagnostic (Raftery and Lewis, 1992) and more recently with Vehtari et al. (2021) who suggest this transformation for effective sample size (ESS) to construct graphical diagnostics or "tail-versions" of this diagnostic. Moreover, a rank-normalization step is added in Vehtari et al. (2021)'s versions 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 (1). This makes the diagnostic more interpretable and allows us to obtain key theoretical results for the associated theoretical R and R_∞.

3 Illustrations

In this section, we consider toy distributions for the chains, where the computation of R_∞ can be done explicitly. In particular, we first focus on two cases raised by Vehtari

et al. (2021) of deficient behaviour of the traditional $\hat{R}$. Then, we exhibit a situation of discrepancy of rank-$\hat{R}$, the updated version of Vehtari et al. (2021). All these theoretical behaviours are illustrated on a small simulation study.

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 $\mathcal{U}(-\sigma, \sigma)$ while the last chain has the cdf F_m of the uniform distribution $\mathcal{U}(-\sigma_m, \sigma_m)$ with $0 < \sigma \leq \sigma_m$. In such a case, the between-variance is zero and it is thus expected that $\hat{R} \approx 1$. In contrast, R_∞ can be written

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

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

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

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

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 a finite first moment when $\alpha > 1$. We focus on the case where one chain is shifted from the other ones: $F_1(x) = \cdots = F_{m-1}(x) = F(x \mid \alpha, \eta)$ and $F_m(x) = F(x \mid \alpha, \eta_m)$ with $0 < \eta \leq \eta_m$ and $0 < \alpha \leq 1$. Here, the within- and between-variances do not exist and it is expected in practice that $\hat{R} \approx 1$. In contrast, R_∞ can be written as

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

Clearly, R_∞ is an increasing function of η_m/η starting from $R_\infty = 1$ when $\eta_m = \eta$ and such that $R_\infty \to \infty$ as $\eta_m/\eta \to \infty$. Results are shown in the second column of Figure 1. 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 more prone to indicating a convergence issue than rank-$\hat{R}$.

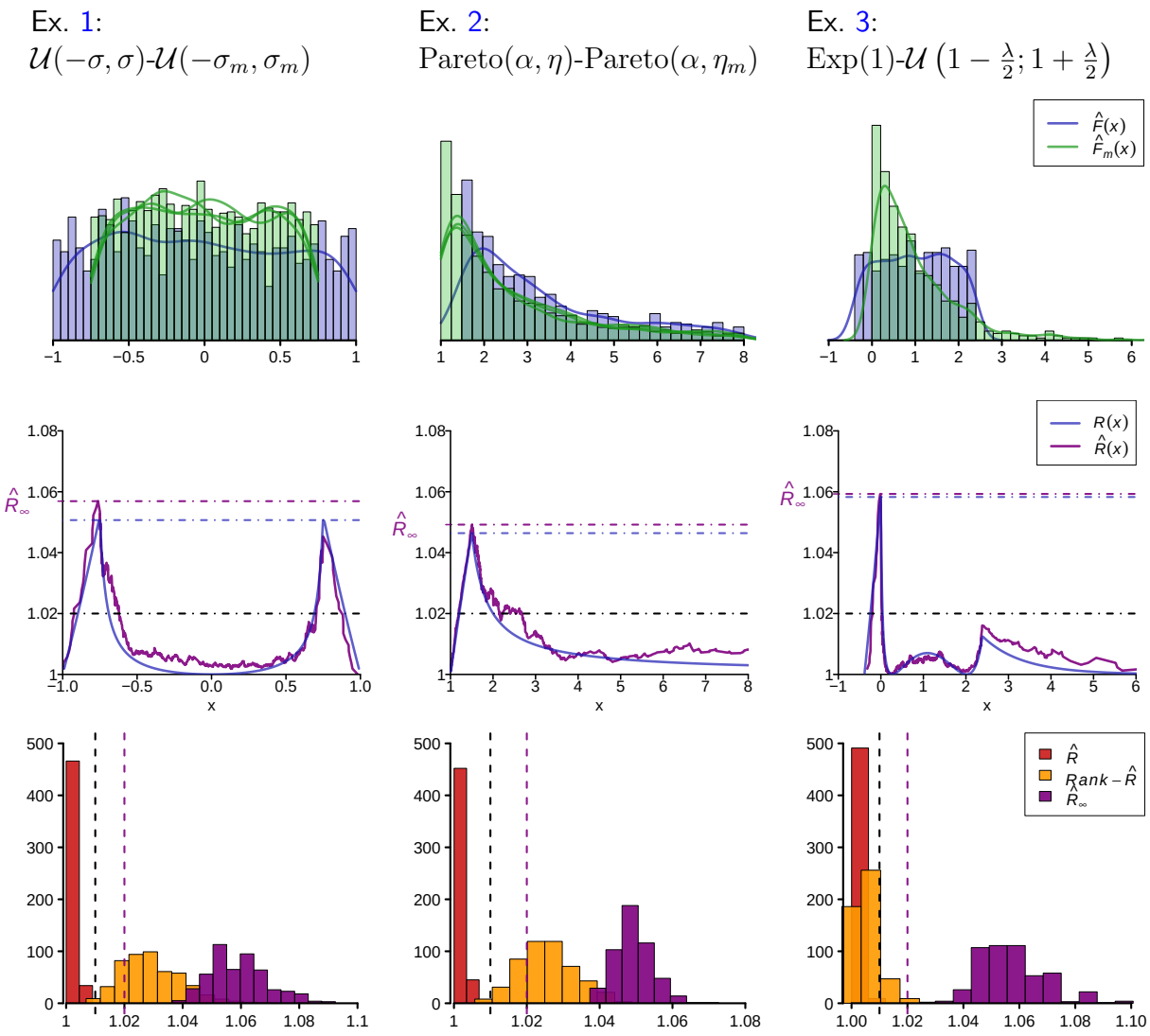

Figure 1: Illustrations with $m = 4$ chains, $n = 200$ independent iterations each. Top row: Empirical distributions $F_1 = \cdots = F_{m-1}$ in green distinct from F_m in blue. For the uniform example (left), $\sigma = 3/4$ and $\sigma_m = 1$, for the Pareto (middle) $\eta = 1$ and $\eta_m = 1.5$, and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

Example 3: chains with same mean and mean over the median. Finally, we look at an example where both $\hat{R}$ and rank-$\hat{R}$ fail to detect non-convergence. Consider $m-1$ exponential chains Exp(1) and one uniform $\mathcal{U}(1-2\log 2, 1+2\log 2)$. This results in chains with same mean and mean over the median, which by construction should fool rank-$\hat{R}$. Results are illustrated in the third row of Figure 1: the histograms of replications confirm that $\hat{R}_\infty$ is able to spot the convergence issue that both $\hat{R}$ and rank-$\hat{R}$ do not.

In the communication, we shall also present applications to other models in a more practical case for Bayesian inference.

References

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events*, volume 33. Springer Science & Business Media.

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

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

— (2022). "On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic." *Preprint*. URL https://hal.inria.fr/hal-03600407

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

Roy, V. (2020). "Convergence diagnostics for Markov chain Monte Carlo." *Annual Review of Statistics and Its Application*, 7, 387–412.

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

Vats, D. and Knudson, C. (2021). "Revisiting the Gelman–Rubin Diagnostic." *Statistical Science*, 36(4), 518 – 529.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion)." *Bayesian Analysis*, 16(2), 667–718.