STAT 547: Bayesian Workflow

Charles C. Margossian
University of British Columbia
Winter 2026
`https://charlesm93.github.io/stat_547/`

DRAFT

2 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a broad class of algorithms used in many fields, including Bayesian Statistics, Statistical Physics, Molecular Dynamics and more. It is often called the workhorse of Bayesian Inference and it is the default inference engine in Stan and other statistical software.

In this section, we review MCMC with the goal of understanding the algorithm's control parameters and how to check that it produces sufficiently accurate answers.

For excellent references which go into more details, I recommend:

- "General state space Markov chains and MCMC algorithms" [Roberts and Rosenthal, 2004]. Reading this paper requires familiarity with measure theory—but don't let that intimidate you!
- "Probabilistic Inference Using Markov Chain Monte Carlo Methods" [Neal, 1993].

2.1 Monte Carlo

Monte Carlo methods estimate properties of a target distribution π using samples.

Example (estimate of the mean of a function $f(z)$):

$$\hat{f}_N = \frac{1}{N} \sum_{n=1}^{N} f\left(z^{(n)}\right), \qquad z^{(1)}, z^{(2)}, \cdots z^{(N)} \overset{iid}{\sim} \pi. \tag{1}$$

<u>Question:</u> how good is this estimator?

Let's examine the expected squared error:

$$\mathbb{E}\left[(\hat{f}_N - \mathbb{E}f)^2\right] = \underbrace{\left(\mathbb{E}\hat{f}_N - \mathbb{E}f\right)^2}_{\text{squared bias}} + \underbrace{\text{Var}\hat{f}_N}_{\text{variance}}. \tag{2}$$

If we draw exact samples from $\pi(z)$, then $\hat{f}_N$ is unbiased,

$$\mathbb{E}\hat{f}_N = \mathbb{E}f, \tag{3}$$

and if the samples are independent, then the variance is

$$\text{Var}\hat{f}_N = \frac{1}{N}\text{Var}f. \tag{4}$$

In addition, we have:

(i) a strong law of large numbers,

$$P\left(\lim_{N\to\infty} \hat{f}_N = \mathbb{E}f\right) = 1. \tag{5}$$

(ii) a central limit theorem,

$$\frac{\sqrt{N}(\hat{f}_N - \mathbb{E}f)}{\sqrt{\mathbb{V}\mathrm{ar}f}} \xrightarrow{d} \mathrm{normal}(0,1). \tag{6}$$

Challenge: For many problems, we cannot draw exact samples from π. Instead we use MCMC to draw approximate samples from π. But MCMC produces samples which are neither independent nor identically distributed.

2.2 Asymptotic MCMC

It what follows, I'll consider distributions over $\mathcal{Z} \subseteq \mathbb{R}^d$ and I'll assume that these distributions admit a density in order to simplify a little bit the notation.

MCMC starts from an initial point z_0 and then applies a *transition kernel* from $z^{(i)}$ to $z^{(i+1)}$,

$$\Gamma\left(z^{(i)}, z^{(i+1)}\right) = p\left(z^{(i+1)} \mid z^{(i)}\right). \tag{7}$$

We denote P_n the distribution of the n^{th} element of the Markov chain, obtained by sequentially applying n times the transition kernel starting from z_0. We denote p_n the corresponding density. Similarly, we denote P_π the target distribution and π the corresponding density.

In general, we want to show that for any (measurable) set A and for a sufficiently large n,

$$P_n(z \in A) \approx P_\pi(z \in A). \tag{8}$$

Question: How should we define "$\approx$"?

We can start by checking convergence in distribution, that is

$$\lim_{n\to\infty} P_n(z \in A) = P_\pi(z \in A). \tag{9}$$

This can be shown for a broad class of transition kernels. The proof usually involves checking a property called *detailed balance* or *reversibility*.

Definition 1. *(reversibility) We say Γ is reversible with respect to π if*

$$\pi(z)\Gamma(z,z') = \pi(z')\Gamma(z',z). \tag{10}$$

Proposition 2. *If Γ is reversible with respect to π, then π is a stationary distribution of the Markov chain generated by Γ.*

Proof. We need to show that the distribution of a new state z' given a previous state $z \sim \pi$ is also π, that is

$$\int \Gamma(z, z')\pi(z)\mathrm{d}z = \pi(z'). \tag{11}$$

Applying the assumption of reversibility,

$$\int \Gamma(z, z')\pi(z)\mathrm{d}z = \int \Gamma(z', z)\pi(z')\mathrm{d}z = \pi(z') \int \Gamma(z', z)\mathrm{d}z = \pi(z'), \tag{12}$$

where in the last step, we recalled that $\Gamma(z', z) = p(z \mid z')$ is a probability density and must therefore integrate to 1.

□

Interpretation: Once the Markov chain "reaches" π, it stays there.

Remark: Reversibility is a sufficient condition for π to be a stationary distribution but not a necessary one. What is an example of an MCMC algorithm which is not reversible but has the right stationary distribution?

Example (Metropolis-Hastings).

Given z, generate a proposal z^* from a proposal distribution $q(z^* \mid z)$. Then accept this proposal with a probability,

$$\alpha(z, z^*) = \min\left[1, \frac{\pi(z^*)q(z \mid z^*)}{\pi(z)q(z^* \mid z)}\right]. \tag{13}$$

That is, for $u \sim \text{uniform}(0, 1)$,

$$z' = \begin{cases} z^*, & \text{if } u \leq \alpha(z, z^*) \\ z, & \text{otherwise.} \end{cases} \tag{14}$$

Exercise: Show that the stationary distribution of a Markov chain using the Metropolis-Hastings transition has stationary distribution π. (Hint: show that it is reversible with respect to π.)

Once we establish that our Markov chain has the right stationary distribution, we need to make sure that it asymptotically gets there. For this, we need to verify two conditions: (i) ϕ-irreducibility and (ii) aperiodicity. The exact definitions of these terms requires measure theory, which is a bit beyond this course. Instead, we'll provide some intuition:

- **ϕ-irreducibility:** From any z, the Markov chain will eventually "explore" the entire space and reach any measurable set.
- **Aperiodicity:** The chain will not oscillate in a regular pattern between different states.

  Example (Periodic Markov chain with the right stationary distribution). Consider a state space $\mathcal{Z} = \{1, 2, 3\}$ and let π be uniform over $\mathcal{Z}$. Consider a Markov chain with

$$\Gamma(1, 2) = \Gamma(2, 3) = \Gamma(3, 1) = 1. \tag{15}$$

Then π is stationary. (Why?)
Intuitively, the chain is irreducible.
However, if $z_0 = 1$, then $p\left(z^{(n)}\right) \neq \pi$ for any n.

Equipped with these notions, we now have sufficient conditions to build a Markov chain which asymptotically reaches π. Specifically, we have two results: (i) convergence in distribution, which is a statement about the distribution of $z^{(n)}$ as $n \to \infty$; and (ii) a law of large numbers, which is statement about the Monte Carlo estimator $\hat{f}_N$ itself. For completeness, I'm providing the formal measure-theoretical statement of this result.

Theorem 3. *If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible, aperiodic, and has stationary distribution π, then for π-almost-everywhere $z \in \mathcal{Z}$ and for any measurable set A,*

$$\lim_{n \to \infty} P\left(z^{(n)} \in A\right) = P_\pi(z \in A). \tag{16}$$

Furthermore, for $f : \mathcal{Z} \to \mathbb{R}$ with $\mathbb{E}(|f|) < \infty$, we have a strong law of large numbers,

$$P\left(\lim_{N \to \infty} \hat{f}_N = \mathbb{E}f\right) = 1. \tag{17}$$

Question: How might you verify the conditions of Theorem 3 for the Metropolis-Hastings algorithm? What conditions on the proposal distribution q do we need?

2.3 Pre-asymptotic MCMC

Theorem 3 is a remarkable result: under fairly weak conditions, we can construct a Markov chain that has the right stationary distribution and will asymptotically get there. But this does not tell us how our Markov chain behaves over a finite number of iterations. To answer this question, we need to study two properties of the Markov chain:

(i) how quickly does P_n approach P_π? This, as we will see, is a statement about the *bias* of $\hat{f}_N$.

(ii) If the Markov chain is stationary, how quickly does the variance of $\hat{f}_N$ decrease?

2.3.1 How quickly does P_n approach P_π?

To tackle the question of convergence speed, it is common to bound a distance between P_n and P_π by a function of N, the number of iterations. There are many ways to compare distributions and none of them seem to be perfect. Which way to use remains, in my view, a somewhat open question (we'll revisit this topic when talking about variational inference). In the MCMC literature, it is common to bound the *total variation distance*.

Definition 4. *The total variation distance between two probability distributions P_{ν_1} and P_{ν_2}*

is:

$$||P_{\nu_1} - P_{\nu_2}|| = sup_A |P_{\nu_1}(z \in A) - P_{\nu_2}(z \in A)|. \tag{18}$$

In words, if we consider all possible measurable sets A and pick the one that "maximizes" (supremizes?) the difference between P_{ν_1} and P_{ν_2}, how big is this difference?

The total variation distance can be written in a way that more clearly relates to our goal of constructing Monte Carlo estimators.

Proposition 5. *The total variation distance between two probability measures P_{ν_1} and P_{ν_2}, respectively with density ν_1 and ν_2, is, for $a < b$,*

$$||P_{\nu_1} - P_{\nu_2}|| = \frac{1}{b-a} sup_{f:\mathcal{Z}\to[a,b]} \left| \int f(z)\nu_1(z)dz - \int f(z)\nu_2(z)dz \right|. \tag{19}$$

Proof. Left as an exercise. □

In words, if we consider all bounded functions f, the total variation distance provides an upper-bound on how much the *expectation value* of f with respect to ν_1 and ν_2 can disagree. Applying this to the setting of MCMC, we obtain,

$$\frac{1}{b-a} \left| \mathbb{E}\hat{f}_N - \mathbb{E}f \right| \leq ||P_n - P_\pi||. \tag{20}$$

This is a bound on the bias of our Monte Carlo estimator but only for bounded functions.

Another useful property of the total variation distance is that it provides bounds on the bias of quantile estimates. For what follows, let's take $\mathcal{Z}$ to be $\mathbb{R}$ since quantiles are defined for univariate quantities. Recall that the α^{th} quantile of measure ν_1, $a_{\nu_1}(\alpha)$, is implicitly defined as

$$P_{\nu_1}(z \leq a_{\nu_1}(\alpha)) = \alpha. \tag{21}$$

Proposition 6. *Suppose $||\nu_1 - \nu_2|| = \varepsilon$ and assume that the cumulative distribution functions (CDFs) of ν_1 and ν_2 are strictly increasing. Then,*

$$a_{\nu_1}(\alpha - \varepsilon) \leq a_{\nu_2}(\alpha) \leq a_{\nu_1}(\alpha + \varepsilon). \tag{22}$$

Proof. Recall the CDF, $F_{\nu_1} : \mathcal{Z} \to [0, 1]$, is,

$$F_{\nu_1}(a) = P_{\nu_1}(z \leq a). \tag{23}$$

Since we assume that F_{ν_1} is strictly increasing, we have that F_{ν_1} is invertible and so for $\alpha = F_{\nu_1}(a)$, we have that $a = F_{\nu_1}^{-1}(\alpha)$.

Denote $A = (-\infty, a]$. Then, from the bound provided by the total variation distance,

$$|P_{\nu_1}(A) - P_{\nu_2}(A)| \leq \varepsilon, \tag{24}$$

or using a different notation,

$$|F_{\nu_1}(a) - F_{\nu_2}(a)| \leq \varepsilon. \tag{25}$$

By assumption, F_{ν_2} is strictly increasing and therefore invertible. Pick $a = a_{\nu_2}(\alpha) = F_{\nu_2}^{-1}(\alpha)$. Then, plugging this in,

$$\begin{aligned} & |F_{\nu_1} \circ F_{\nu_2}^{-1}(\alpha) - F_{\nu_2} \circ F_{\nu_2}^{-1}(\alpha)| && \leq \varepsilon && (26) \\ \iff & |F_{\nu_1} \circ F_{\nu_2}^{-1}(\alpha) - \alpha| && \leq \varepsilon && (27) \end{aligned}$$

Then,

$$\alpha - \epsilon \leq F_{\nu_1} \circ F_{\nu_2}^{-1}(\alpha) \leq \alpha + \epsilon. \tag{28}$$

Applying $F_{\nu_1}^{-1}$ on each side (and noting that $F_{\nu_1}^{-1}$ must also be strictly increasing),

$$F_{\nu_1}^{-1}(\alpha - \epsilon) \leq F_{\nu_2}^{-1}(\alpha) \leq F_{\nu_1}^{-1}(\alpha + \epsilon), \tag{29}$$

which is the wanted result.

□

Now that we have some motivation for bounding the total variation distance, we can try to understand how quickly this distance decreases. We'll start with an elementary property that states that applying a transition kernel must decrease the total variation distance.

Proposition 7. *If P_π is a stationary distribution for a Markov chain kernel, then for any $n \in \mathbb{N}$,*

$$||P_{n+1} - P_\pi|| \leq ||P_n - P_\pi||. \tag{30}$$

Proof. For any event A,

$$|P_{n+1}(z \in A) - P_\pi(z \in A)| = \left| \int P_1(z \in A \mid z')p_n(z')\mathrm{d}z' - \int P_1(z \in A \mid z')\pi(z')\mathrm{d}z' \right|, \tag{31}$$

where we used the fact that, by assumption, applying a transition kernel to $z' \sim \pi$ still generates a sample $z \sim \pi$.

Let $f(z') = P_1(z \in A \mid z')$ and notice f is a bounded function, specifically $f : \mathcal{Z} \to [0, 1]$. Then eq. (32) can be rewritten as,

$$|P_{n+1}(z \in A) - P_\pi(z \in A)| = \left| \int f(z')p_n(z')\mathrm{d}z' - \int f(z')\pi(z')\mathrm{d}z' \right| \leq ||P_n - P_\pi||, \tag{32}$$

where the inequality follows from Proposition 5 with $a = 0$ and $b = 1$. Taking the supremum with respect to A on both sides of eq. (32), we obtain the wanted inequality. □

With a more detailed analysis, we can derive a bound on the total variation distance of the form,

$$||P_n - P_\pi|| \leq b(z_0)h(N), \tag{33}$$

where by convention $h(0) = 1$ and $b(z_0)$ is the initial bias $|f(z_0) - \mathbb{E}f|$. Under stronger assumptions, it is possible to characterize $h(N)$. One regime of particular interest arises when h is a geometric function, $h(N) = \lambda^N$ for $\lambda \in (0, 1)$. This regime is called *geometric ergodicity*. In this regime:

(i) the bias of $f\left(z^{(n)}\right)$ decreases exponentially and so does the bias of $\hat{f}_N$ if we discard early samples.

Exercise: how quickly does the bias of $\hat{f}_N$ decrease if we don't discard early samples?

(ii) we have a central limit theorem for $\hat{f}_N$, which is the topic of the next section.

Proposition 8. *When the state space $\mathcal{Z}$ is finite, then any irreducible and aperiodic Markov chain with stationary distribution π is geometrically ergodic.*

Not all Markov chains on discrete spaces are irreducible and aperiodic and furthermore, not all irreducible and aperiodic Markov chains on continuous state spaces are geometrically ergodic. Ensuring geometric ergodicity over continuous spaces requires additional technical requirements.

In general, it is difficult to calculate h and even then, bounds on the total variation distance tend to be conservative, since they need to account for worst case scenarios, rather than focusing on the particular functions we might be interested in. This will motivate us to look at empirical measures of convergence which, while imperfect, can be deployed in practice.

2.3.2 Variance of MCMC at stationarity

A question of interest is how quickly does the variance of $\hat{f}_N$ decrease if we start from the stationary distribution? This idealized case approximates the behavior of Markov chains which have been warmed up for a sufficiently long time and which are *nearly* stationary. In this ideal setting, the Monte Carlo estimator is unbiased but we still need to handle its variance.

Here, a quantity of interest is the *effective sample size* (ESS), defined as follows,

$$\text{ESS} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho(t)}, \tag{34}$$

where $\rho(t)$ is the autocorrelation between two samples separated by t steps. Notice that if $\rho(t) = 0$ for all t, ESS $= N$.

The ESS plays a key role in the MCMC central limit theorem.

Theorem 9. *Consider the Monte Carlo estimator $\hat{f}_N$ obtained from a geometrically ergodic MCMC algorithm. Then*

$$\frac{\sqrt{ESS}(\hat{f}_N - \mathbb{E}f)}{\sqrt{Var(f)}} \xrightarrow{d} normal(0, 1). \tag{35}$$

This central limit theorem is almost identical to the one we have for independent samples (eq. (6)), except that the rate of convergence is $\sqrt{\text{ESS}}$ rather than $\sqrt{N}$. Moreover, for a large enough sample, we have

$$\hat{f}_N \overset{\text{approx.}}{\sim} \text{normal}\left(\mathbb{E}f, \frac{\text{Var}f}{\text{ESS}}\right). \tag{36}$$

This suggests another interpretation of the ESS for stationary Markov chains,

$$\text{ESS} \approx \frac{\text{Var}f}{\text{Var}\hat{f}_N} \iff \text{Var}\hat{f}_N \approx \frac{\text{Var}f}{\text{ESS}}, \tag{37}$$

and one can once again check that for i.i.d samples, ESS $= N$. Eq. (37) drives home the point that for stationary Markov chains and for large N, the ESS monitors the variance of $\hat{f}_N$, scaled by the posterior variance.

Based on eq. (37), the ESS can further be interpreted as the number of independent samples we would require to achieve the same expected squared error as our Monte Carlo estimator.

2.4 Practical diagnostics for MCMC

Remember that our goal is to control the expected squared error of the Monte Carlo estimator $\hat{f}_N$. For this we resort to a *warmup* phase and discard early samples to reduce the bias. We then run a *sampling* phase to control the variance. By default, Stan runs 1000 warmup iterations and 1000 sampling iterations.[1]

The bias and variance hint at two sources of error:

- The bias is due to the nonstationary initialization.
- The variance is due to the randomness in the MCMC (and is increased by the fact that the samples tend to be autocorrelated).

A general strategy to check the influence of these two sources of error on $\hat{f}_N$ is to run multiple chains with distinct initializations and seeds, and make sure that they still produce results which are in “good agreement” with one another.

They are multiple ways to measure agreement. First, we can perform visual checks using trace plots and density plots (Figures 1 and 2).

Another way to check for agreement is to compute the sample variance of the Monte Carlo estimator generated by individual chains. Moving forward, we denote with a superscript the chain identity of each Monte Carlo estimator: for example, $\hat{f}_N^{(m)}$ is the Monte Carlo estimator generated by the m^{th} chain, and,

$$\bar{f}_N^{(\cdot)} = \frac{1}{M}\sum_{m=1}^{M} \hat{f}_N^{(m)}, \tag{38}$$

is the Monte Carlo estimator obtained by averaging across all chains. The sample variance of the per chain Monte Carlo estimator is then

$$\widehat{B} := \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{f}_N^{(m)} - \bar{f}_N^{(\cdot)}\right)^2, \tag{39}$$

and we can check that $\widehat{B} \approx 0$ as a measure of how well the Markov chains agree with one another.

[1] We will see that the warmup phase in Stan is not only used to reduce the bias of $\hat{f}_N$ but also to tune the MCMC algorithm so that it performs better during the sampling phase.

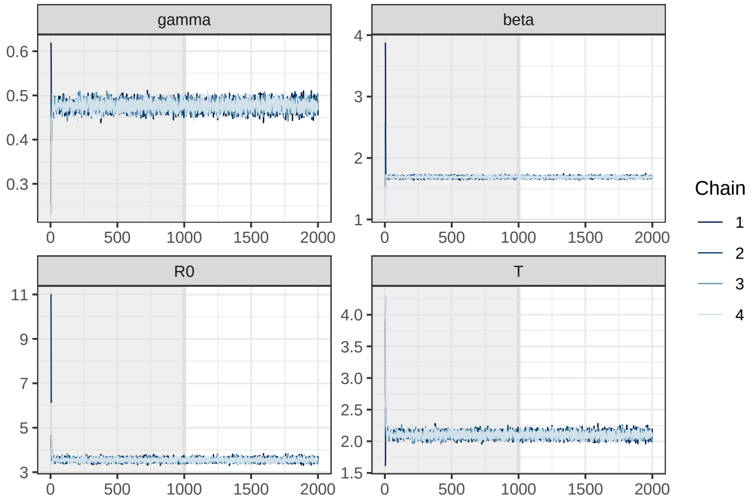

Figure 1: Trace plots for MCMC. *Value for each parameter across iterations. The shaded area corresponds to the warmup phase (first 1000 iterations). Here, we want to check that all the Markov chains have converged to the same "area" despite their distinct initialization and seed. For this model, it only takes a few iterations for all the chains to converge in distribution. The "fuzzy caterpillar" shape indicates that there is little autocorrelation between successive samples.*

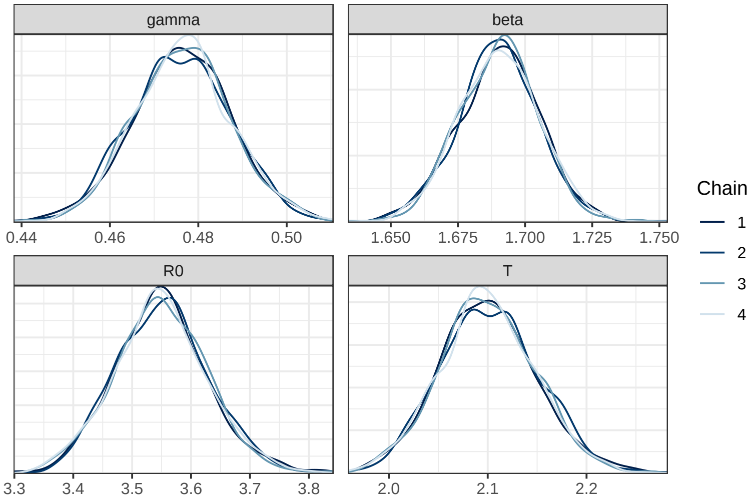

Figure 2: Density plots for MCMC. *Marginal density estimation using samples from the sampling phase. Here too we can check that the marginal densities returned by different chains are in good agreement.*

You might find this approach at best somewhat convincing. The sample variance lets us measure the variance but can it actually tell us something about the bias?[2] Intuitively, we expect that Markov chains which have not converged to π will generate Monte Carlo estimators with higher variance (why?).

We can formalize this intuition. Suppose we draw the Markov chain's initial point $z_0^{(k)}$ from an initial distribution P_0. Then, applying the law of total variance,

$$\text{Var}\hat{f}_N^{(k)} = \underbrace{\text{Var}\mathbb{E}\left(\hat{f}_N^{(k)} \mid z_0^{(k)}\right)}_{\text{nonstationary}} + \underbrace{\mathbb{E}\text{Var}\left(\hat{f}_N^{(k)} \mid z_0^{(k)}\right)}_{\text{persistent}}. \tag{40}$$

The variance term decomposes into a *nonstationary* and a *persistent* variance:

- The nonstationary variance measures how much the expectation value of $\hat{f}_N^{(k)}$ varies with the initial point $z^{(k)}$ and vanishes as the Markov chain "forgets" its starting point. Hence, it is an indirect measure of the squared bias and how close to convergence the Markov chain is. **Formalizing this connection is a open research problem!**
- The persistent variance eventually dominates the total variance and for stationary Markov chains measures the asymptotic variance.[3]

For many problems, we care about the squared error of $\hat{f}_N$ relative to the posterior variance and so $\widehat{B}$ is scaled by a measure of the posterior variance. For each Markov chain, we compute a sample variance, and then average across all chains,

$$\widehat{W} := \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N-1}\sum_{n=1}^{N}\left(f\left(z^{(nk)}\right) - \hat{f}_N^{(k)}\right)^2. \tag{41}$$

Then we examine the ratio $\widehat{B}/\widehat{W}$ and check that it is close enough to 0 (meaning the variance of the perchain Monte Carlo estimator is small relative to the estimated posterior variance).

For historical reasons, the quantity we typically measure is,

$$\widehat{R} = \sqrt{\frac{N-1}{N} + \frac{\widehat{B}}{\widehat{W}}}, \tag{42}$$

a quantity known as either the $\widehat{R}$ statistic, the *potential scale reduction factor* or the *Gelman-Rubin* statistic, and whose initial motivation is a bit different than what I'm presenting here. Nonetheless, $\widehat{R}$ is a 1-to-1 map with $\widehat{B}/\widehat{W}$.

A recent recommendation is to check that $\widehat{R} \leq 1.01$ [Vehtari et al., 2021]. This recommendation is battle-tested and seems to work well, **however finding a principled threshold for $\widehat{R}$ is an open research question.**

[2] This was a question I started working on during the final year of my PhD!

[3] One can also show that the stationary variance is inflated before the Markov chain convergences due to a "drift" phenomenon: that is, as long as the Markov chain is not stationary, we're likely averaging samples with a different mean and this increases the variance.

In addition to $\widehat{R}$, we also estimate the ESS as in eq. (34), based on estimates of the autocorrelation ρ_t [Geyer, 1992]. Now, from eq. (37), we have that for a stationary Markov chain,

$$\text{ESS} \approx \frac{\text{Var} f}{\text{Var} f_N} \approx \frac{\widehat{W}}{\widehat{B}}. \tag{43}$$

However, the above estimate tends to be less stable than the one obtained using autocorrelation functions, and so it is useful to compute the ESS, once we establish that the Markov chains have approximately converged to the stationary distribution with $\widehat{R}$.

Let's return now to the table of output from fitting the SIR model.

variable	**mean**	**median**	**sd**	**mad**	**q5**	**q95**	**rhat**	**ess_bulk**
gamma	0.476	0.476	0.0110	0.0108	0.459	0.495	1.00	2362
beta	1.69	1.69	0.0149	0.0146	1.67	1.71	1.00	2794
R0	3.55	3.55	0.0786	0.0766	3.42	3.68	1.00	2962
T	2.10	2.10	0.0484	0.0474	2.02	2.18	1.00	2362

For all quantities of interest, we achieve $\widehat{R} \leq 1.01$ which suggests good convergence and we have ESS$>$ 2000 (labeled ess_bulk).

A somewhat **open research question** is to determine what is a useful ESS and more generally an acceptable expected squared error. While this of course depends on the problem at hand, in my experience, field practitioners do not themselves quite know how precise they need their Monte Carlo estimators to be. But this is crucial to determine how long the Markov chains need to be!! See Margossian and Gelman [2024] for further discussion.

For an example where the Markov chains do not converge and disagree, and where $\widehat{R} \gg 1$, see Chapter 29 of *Bayesian Workflow*.

Open discussion: Are Stan's default control parameters (4 Markov chains with a warmup phase of 1000 iterations and a sampling phase of 1000 iterations) optimal for doing Bayesian inference on the SIR model?

References

C. J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992. doi: 10.1214/ss/1177011137.

C. C. Margossian and A. Gelman. For how many iterations should we run Markov chain Monte Carlo? 2024.

R. M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993. URL `https://www.cs.toronto.edu/~radford/review.abstract.html`.

G. O. Roberts and J. S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004. ISSN 1549-5787. doi: 10.1214/154957804100000024.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved rhat for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi: **10.1214/20-BA1221**.