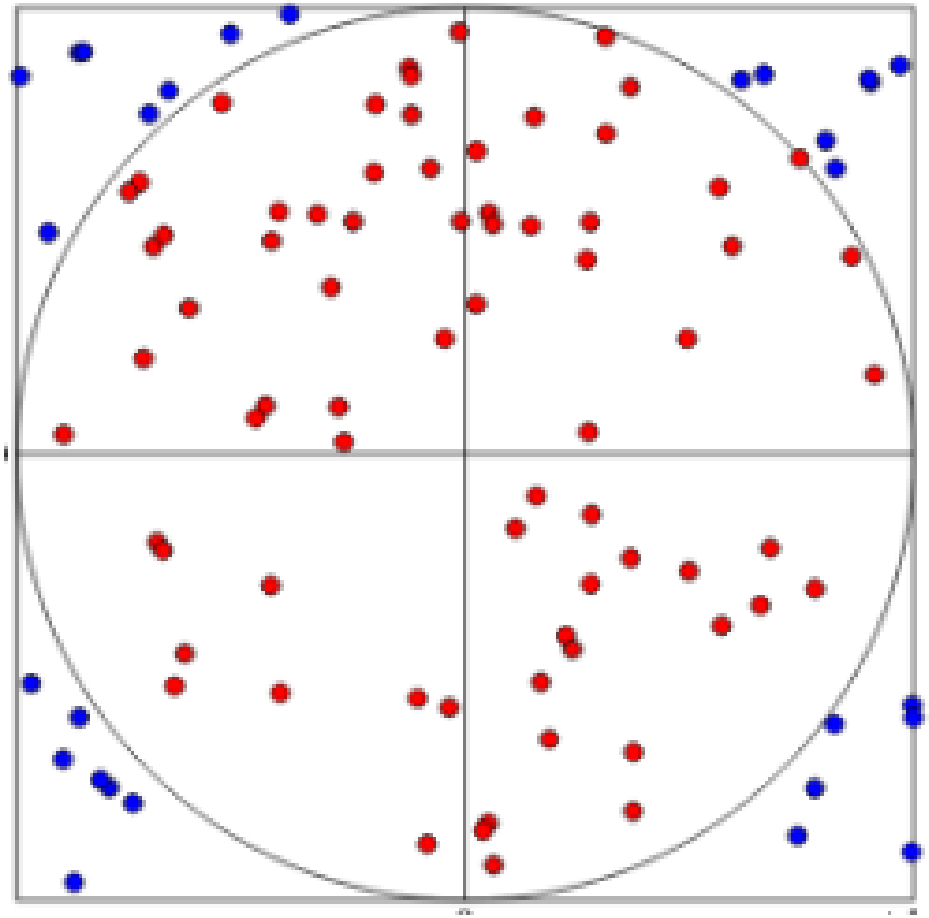
MCMC: The best bad method you have ever seen

FGV EMAp

Markov chain Monte Carlo (MCMC) methods are a broad class of stochastic algorithms to compute integrals.
Suppose you are confronted with the following question: what is the ratio between the circumference of inscribed circle and its diameter? You are **not** allowed to use any Geometry.



First, a warning

> *"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."*

Alan Sokal (1955-) in *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms* (1996, pg. 1).

MCMC is, in a way, like a captive tiger...

Also...

Repeat after me,

Idea 12 (Bayesian MCMC is not a thing)

There is no such thing as "Bayesian" MCMC.

MCMC is a numerical method for computing integrals. It does not care whether you are a Bayesian, frequentist, flamenguista or corintiana.

Computing integrals

Technically, for a probability space $(X, \mathcal{F}, P)$, for $f : X \to \mathbb{R}$, we want to compute

$$\mu_f = E_P[f] = \int_X f \, dP.$$

When P is absolutely continuous with respect to the Lebesgue measure, we have

$$\mu_f = \int_X f(x)p(x) \, dx,$$

as is usually written in introductory textbooks.
A "natural" approach to obtain an estimator of μ_f is

$$\hat{\mu}_{f,N}^{\text{MC}} = \frac{1}{N} \sum_{n=1}^{N} f(x_n),$$

with $x_1, \ldots, x_N \sim P$.

A central (limit) theorem

Define

$$\text{MC-SE}_N[f] = \sqrt{\frac{\text{Var}_P[f]}{N}}.$$

Then

$$\lim_{N\to\infty} \frac{\hat{\mu}^{\text{MC}}_{f,N} - \mathbb{E}_P[f]}{\text{MC-SE}_N[f]} \sim \text{Normal}(0,1),$$

Idea 13 (MCMC-CLT needs to hold)

A key insight is that MCMC only trustworthy when a central limit theorem holds. This means f needs to be $2+\epsilon$-integrable with respect to P. Look out for MC-SE, too. It is important to quantify "the probable error of the mean"[14], as it were.

[14] A "pun" with William Gosset's (1876–1937) paper: Student. (1908). The probable error of a mean. Biometrika, 1-25.

FGV EMAp

Idea 14 (Diagnose your MCMC!)

Perhaps as important as learning how to run an MCMC is to learn to ***diagnose*** *it. This means detecting failure to converge to P and/or poor statistical performance.*

When running K chains, the between sample variance can be written as

$$B = \frac{N}{K-1} \sum_{k=1}^{K} \left(\bar{x}_k - \bar{\bar{x}}\right)^2,$$

where $\bar{x}_k = N^{-1} \sum_{n=1}^{N} x_k^{(n)}$ and $\bar{\bar{x}} = K^{-1} \sum_{k=1}^{K} \bar{x}_k$. Now we can define the within variance as

$$W = K^{-1} \sum_{k=1}^{K} s_k^2 \text{ and } s_k^2 = (N-1)^{-1} \sum_{n=1}^{N} \left(x_k^{(n)} - \bar{x}_k\right)^2$$

Finally we can define the **potential scale reduction factor** (PSRF) (Gelman and Rubin, 1992):

$$\hat{R} = \sqrt{\frac{(N-1)W + B}{NW}}.$$

At convergence, $\hat{R} < 1.1$, providing a univariate measure of convergence across chains (for a given parameter).

More diagnostics

One of the things we are interested in is *statistical* performance, i.e., how precise the estimator $\hat{\mu}_{f,N}^{\text{MC}}$ is. To measure that, we can compute the **effective sample size**:

$$\text{ESS} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t},$$

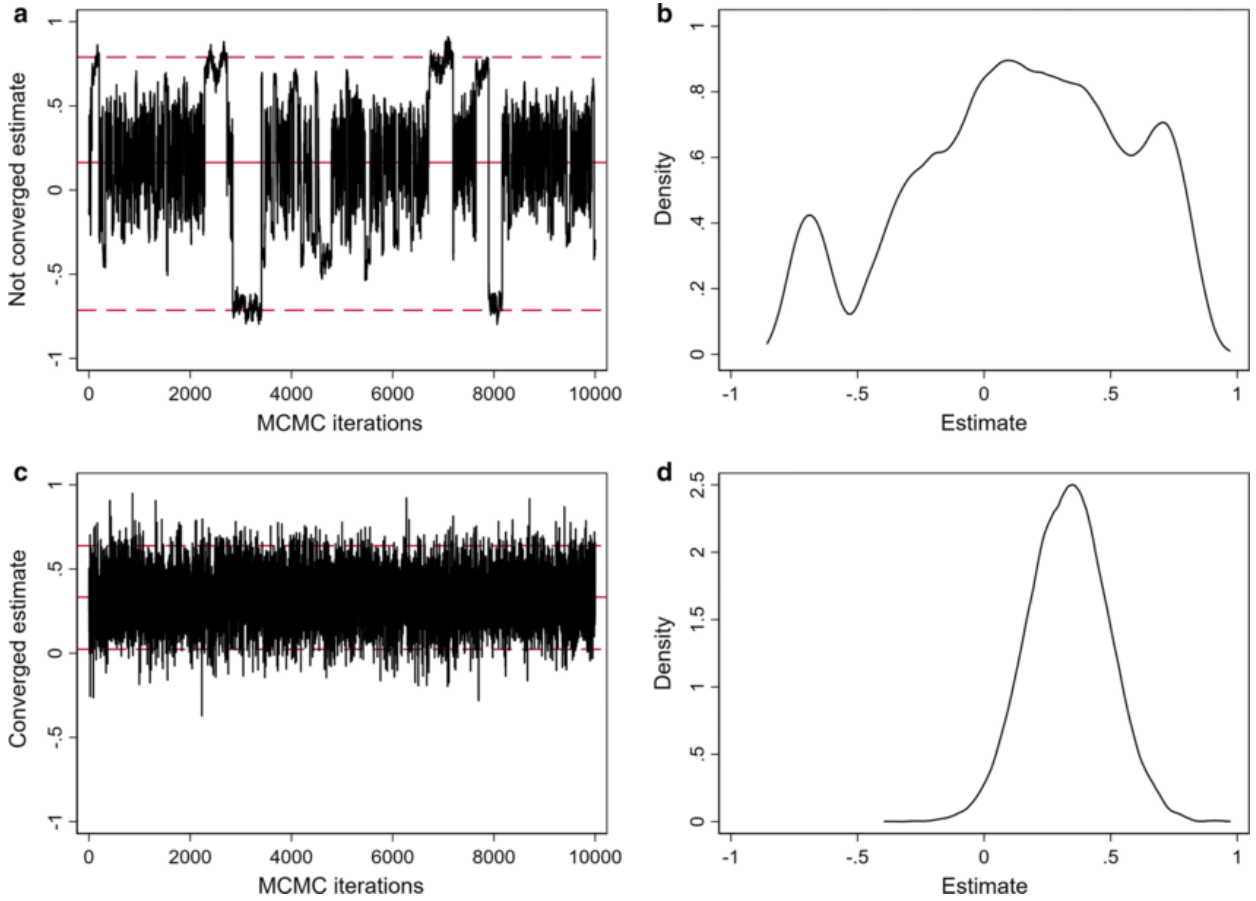
where ρ_t is the **autocorrelation** at lag t, $t = 1, 2, \ldots$. A good rule of thumb[15] is that if one wants to have a an standard error which is 1% of the width of the 95% interval of the true distribution is to have ESS ≥ 625:

$$\begin{aligned} \frac{\sigma}{\sqrt{N}} &\leq \frac{\sigma}{\sqrt{\text{ESS}}}, \\ 0.01 \times 4 \times \sigma &\leq \frac{\sigma}{\sqrt{\text{ESS}}}, \\ &\Longrightarrow \\ \text{ESS} &\geq 625, \end{aligned}$$

where $\sigma = \sqrt{\text{Var}_P[f]}$.

[15] Assuming approximate normality. Calculation stolen from https://www.biorxiv.org/content/10.1101/2021.05.04.442586v1.full.pdf

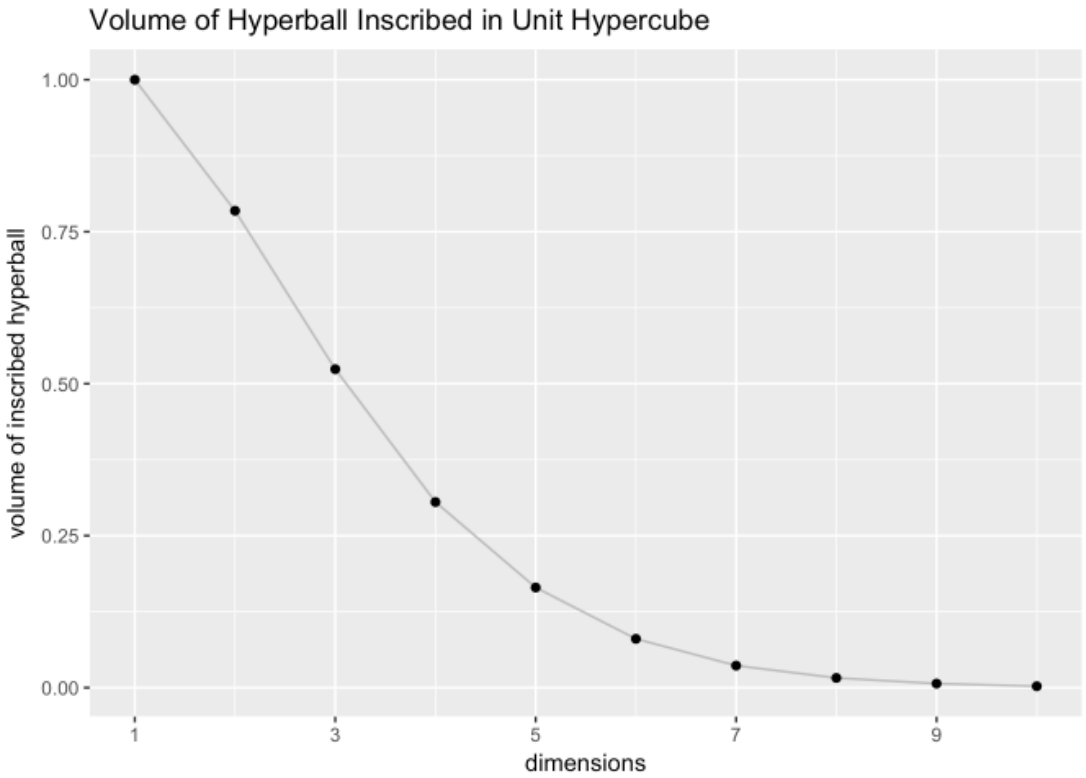
Even more diagnostics



Idea 15 (No one diagnostic is enough)

Use multiple diagnostic metrics, always. Every MCMC diagnostic out there has blind spots; using multiple simultaneously increases the chances those blind spots are covered.

Scaling with dimension



Taken from https://mc-stan.org/users/documentation/case-studies/curse-dims.html.

Idea 16 (The higher the dimension, the more structure you need)

As dimension increases, things start to get pretty lonely pretty fast for a particle. The only way to counteract this "thinning" is to introduce more structure. This is the intuitive basis for the success of gradient-based methods such as MALA[16] and HMC[17].

[16] Metropolis-adjusted Langevin algorithm

Take home

FGV EMAp

- MCMC allows us to make inferences about huge models in Science and Engineering;
- MCMC is a terrible method, which nevertheless is our best shot at computing high-dimensional integrals;
- One has to make sure a CLT holds;
- One has to verify diagnostics to ensure no convergence/performance problems are present;
- No one diagnostic is enough.

Recommended reading

FGV EMAp

- Robert (2007), Ch. 6[18].
- https://betanalpha.github.io/assets/case_studies/markov_chain_monte_carlo.html

[18] The Bayesian Choice by Christian Robert (2007, 2nd edition).