Introduction to rstanarm

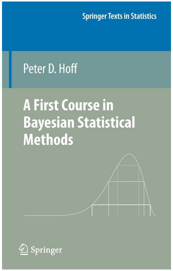
Bayesian Inference - Lab Sessions (1/3)

Marika D'Agostini
marika.dagostini2@unibo.it

University of Bologna

November-December 2023

References



Springer Texts in Statistics

Peter D. Hoff

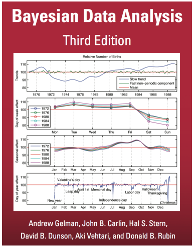
A First Course in Bayesian Statistical Methods

Springer

Basic Textbook:

Peter D. Hoff - *A First Course in Bayesian Statistical Methods* (2009)

`https://pdhoff.github.io/book/`



Bayesian Data Analysis

Third Edition

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin

Advanced Textbook:

Andrew Gelman et al. - *Bayesian Data Analysis (3rd Ed.)* (2020)

`http://www.stat.columbia.edu/~gelman/book/`

Bayesian Statistics: definition

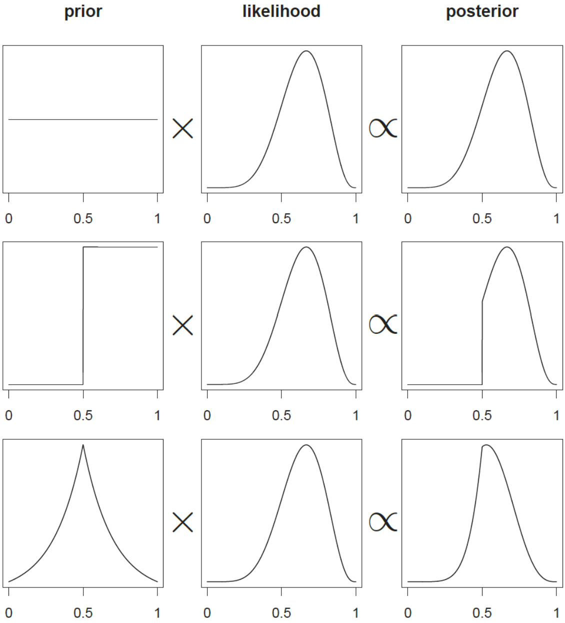
Suppose we observe data $\mathbf{y} = (y_1, ..., y_n)$ which we model as a realisation of random variable $\mathbf{Y} = (Y_1, ..., Y_n)|\theta \sim f(\mathbf{Y}|\theta), \theta \in \Theta$

1) Before using any information from data **y**, we assume there is a distribution over θ called the **prior distribution** with pdf $p(\theta)$
2) The parametric family of distributions with pdf $f(\mathbf{y}|\theta)$ we assume for data can be viewed as a **conditional distribution** of data **y** given θ
3) Can update our knowledge about θ using observed data **Y** from $p(\theta)$ to the conditional distribution of θ given observed data **Y**, **called posterior distribution** of θ, using Bayes theorem

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)p(\theta)}{\int_\Theta f(\mathbf{y}|\theta)p(\theta)d\theta} = \frac{f(\mathbf{y}|\theta)p(\theta)}{f(\mathbf{y})}$$

which is $\propto f(\mathbf{y}|\theta)p(\theta)$ as a function of θ.
Thus,

$$\textbf{likelihood} \times \textbf{prior pdf} \propto \textbf{posterior pdf}$$


prior
likelihood
posterior
×
∝
0
0.5
1

Bayesian vs classical (frequentist) approach

1) Unknown parameter θ:
   - *Frequentist*: a **fixed number**
   - *Bayesian*: a **random variable**
2) Inference about θ:
   - *Frequentist*: ad hoc (different types of estimators/tests are "best" for different problems, **no unique algorithm**)
   - *Bayesian*: given 3 choices (likelihood, prior, loss), there is a **unique inferential procedure**
3) Interval estimation of θ:
   - *Frequentist*: $(1-\alpha)100\%$ **confidence interval** of θ: among all such data sets **y**, in $(1-\alpha)100\%$ of them, θ belongs to this interval
   - *Bayesian*: $(1-\alpha)100\%$ **credible interval** of θ: for given data **y**, θ belongs to this interval with probability $(1-\alpha)$

Steps of Bayesian Inference

1) **Identify/Collect the data** (general recommendation: data visualization)
2) Choose a statistical **model for the data** → $f(\mathbf{y}|\theta)$
3) **Specify prior distributions** for the model parameters → $p(\theta)$
4) Obtain the **posterior distributions** for the model parameters

   $$\rightarrow p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)p(\theta)$$

   4.1) For mathematical approximations, check the algorithms for convergence (**Post-run diagnostics**)
5) Conduct a **posterior predictive check** to examine if the fitted model is compatible with the observed data

   5.1) If the model does not fit the data, one should go back to step 2 to specify a different model
6) **Summarizing the Posterior Distribution**
   - Posterior Mean, Median, and Mode
   - Uncertainty Estimates
   - Credible Intervals

Bayesian computation (I)

4) Obtain the **posterior distributions** for the model parameters

$$\rightarrow p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)p(\theta)$$

Making inference in the Bayesian framework implies to deal with multidimensional integrals:

- Normalizing constants
- Marginal posterior distributions for the parameters of interest
- Expected values
- Posterior predictive distribution

Bayesian computation (II)

Different approaches might be distinguished:

- Conjugate Priors
- *Numerical integration*: feasible only with a regular function with a low-dimensional parameter space
- *Analytical approximation*: Normal or Laplace approximation (e.g. INLA) with Maximum A Posteriori (MAP) Estimation
- **Simulation methods**: numerical values obtained through random generator algorithms → Markov Chain Monte Carlo (**MCMC**)

Monte Carlo (MC) approximation: why does it work? (I)

Reference: *Hoff, 2009; chapter 4 [Gelman, 2020; chapter 11]*

Suppose we are interested in estimating the parameter θ, once the sample $\mathbf{y}$ is observed and the likelihood $f(\mathbf{y}|\theta)$ is assumed for data.

Since we are Bayesian statisticians, we are interested in the **posterior distribution** of θ: $p(\theta|\mathbf{y})$.

Let us suppose the analytic properties of $p(\theta|\mathbf{y})$ to be unknown but we are able to generate a random sample of size S from it:

$$\theta^{(1)}, ..., \theta^{(S)} \overset{iid}{\sim} p(\theta|\mathbf{y}).$$

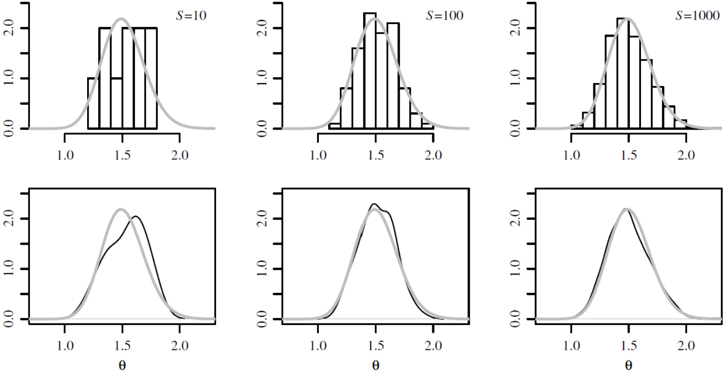
Thanks to rigorous mathematical results based on the *law of large numbers* it is possible to state that the empirical distribution of the generated sample $\{\theta^{(1)}, ..., \theta^{(S)}\}$ is an approximation of the true posterior distribution $p(\theta|\mathbf{y})$.

→ $\{\theta^{(1)}, ..., \theta^{(S)}\}$ is known as a **MC approximation** of $p(\theta|\mathbf{y})$.

MC approximation: why does it work? (II)

More formally:

$$\frac{1}{S}\sum_{s=1}^{S} f(\theta_s^*) \to \mathbb{E}\left[f(\theta)|\mathbf{y}\right] = \int_{\Theta} f(\theta)p(\theta|\mathbf{y})\mathrm{d}\theta, \quad \text{as } S \to +\infty.$$



MC approximation: why does it work? (III)

As a consequence all the empirical evaluations of the following useful characteristics of the distribution can be considered as reliable approximation of the true values:

- Mean and variance,
- Quantiles,
- Probabilities.

Since we are dealing with approximations, it is possible to provide a measure of the **accuracy**:

$$SE_{MC} = \sqrt{\frac{\hat{\sigma}^2}{S}},$$

and it is named **Monte Carlo Standard Error**.

MC approximation: example

Let us consider the simple **Beta-Binomial model**.

Beta prior for the proportion parameter θ:

$$\theta|a, b \sim \mathcal{B}\left(a, b\right)$$

Binomial data model & likelihood function:

$$\mathbf{y}|\theta \sim Bin(n, \theta)$$

Then, given that r successes are observed in n trials (i.e. $\mathbf{y} = r$),

$$\theta|(\mathbf{y} = r) \sim \mathcal{B}\left(a + r, n - r + b\right)$$

→ To describe the posterior it is possible to use Monte Carlo simulations.

See R script: `example_MC.R`

Markov Chains (I)

Simple MC simulation alone is not enough in case of **high dimensional parameters** problems
→ It is required to support it with the concept of **Markov chain**, in order to generate the desired sample $\{\theta^{(1)}, ..., \theta^{(S)}\}$ from $p(\theta|\mathbf{y})$.

A discrete-time Markov chain (or Markov process) is a discrete-time stochastic process such that the **Markovian property** holds

$$\mathbb{P}\left[\theta_t^*|\theta_0^*, ..., \theta_{t-1}^*\right] = \mathbb{P}\left[\theta_t^*|\theta_{t-1}^*\right]$$

i.e.,

it is a discrete-time stochastic model describing a sequence of possible events in which **the probability of each event depends only on the state attained in the previous event**

Markov Chains (II)

If a Markov Chain possesses all these three properties

- **Irreducibility**: each set of states can be reached staring from each state with a finite number of steps
- **Positively recurrent (or persistent)**: the probability of returning to the current state in a finite number of steps is 1.
- **Aperiodic**: there is no periodic oscillation among the states

then the **ergodic theorem** holds and

$$\frac{1}{S}\sum_{b=1}^{B} f(\theta_s^*) \to \mathbb{E}\left[f(\theta)|\mathbf{y}\right], \quad \text{as } S \to +\infty.$$

It is a parallel result of the one for the Monte Carlo integration. These chains converges to the **stationary distribution**, that is unique, independently from the initial value θ_0^*.

Why is this useful in Bayesian inference?

Reference: *Hoff, 2009; chapter 6 [Gelman, 2020; chapter 13]*

Markov Chains Monte Carlo (MCMC) algorithm is mostly used to sample from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ if we are dealing with a multidimensional estimation problem.

Main steps:

- Define Markov chains with the same parameter space of $\boldsymbol{\theta}$.
- Choose a Proposal Distribution: select a simple distribution that can be easily sampled from (e.g. univariate distributions). This distribution is used to propose new candidate points in the parameter space.
- Gradually move in the chain converging towards stationary distribution. $\rightarrow$ The stationary distribution is $p(\boldsymbol{\theta}|\mathbf{y})$.

Once the conditions for the validity of the ergodic theorem are verified, a sequence of **dependent** realizations from $p(\boldsymbol{\theta}|\mathbf{y})$ is obtained.

Thus, on reaching stationary distribution we have approximated posterior probability distribution.

Examples of MCMC algorithms

- **Metropolis-Hastings algorithm**: general framework which includes
  - **Gibbs Sampler** → special case of Metropolis–Hastings algorithm with acceptance rate uniformly equal to 1
  - **Metropolis algorithm** → special case of Metropolis–Hastings algorithm with symmetric proposal distribution.
- **Hamiltonian Monte Carlo (HMC)** → it allows to sample from the posterior of the target parameters more efficiently than basic MCMC algorithms.

The Gibbs sampler (I)

- The **Gibbs sampler** is the easiest MCMC algorithm and it is based on the **full conditionals distributions**.
- Gibbs sampling is attractive because it can sample from high-dimensional posteriors
- The main idea is to break the problem of sampling from the high-dimensional joint distribution into a series of samples from low-dimensional conditional distributions
- Updates can also be done in blocks (groups of parameters)
- Because the low-dimensional updates are done in a loop, samples are not independent → the dependence turns out to be a Markov distribution → MCMC

The Gibbs sampler (II)

If a m-dimensional estimation problem is faced: $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$, the m full conditionals posterior distributions are:

$$p(\theta_j|\boldsymbol{\theta}_{-j}, y), \quad j = 1, ..., m.$$

The algorithm is constituted by the following steps:

- Fixing the initial state at $(\theta^*_{1,(0)}, ..., \theta^*_{m,(0)})$
- For each step b generate:

$$\begin{aligned}
\theta^*_{1,(b)} &\sim p(\theta_1|\theta^*_{2,(b-1)}, ..., \theta^*_{m,(b-1)}, y),\\
\theta^*_{2,(b)} &\sim p(\theta_2|\theta^*_{1,(b)}, \theta^*_{3,(b-1)}, ..., \theta^*_{m,(b-1)}, y),\\
&\dots\\
\theta^*_{m,(b)} &\sim p(\theta_m|\theta^*_{1,(b)}, ..., \theta^*_{m-1,(b)}, y).
\end{aligned}$$

- Repeat B times.

The Gibbs sampler (III)

$$\mathbb{P}(\boldsymbol{\theta}^{(b)} \in A) \rightarrow \int_A p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad as \quad b \rightarrow \infty$$

In words, the sampling distribution of $\boldsymbol{\theta}^{(b)}$ approaches the target distribution as $b \rightarrow \infty$, no matter what the starting value $\boldsymbol{\theta}^{(0)}$ is (although some starting values will get you to the target sooner than others).

More importantly, for most functions g of interest,

$$\frac{1}{B} \sum_{b=1}^{B} g(\boldsymbol{\theta}^{(b)}) \rightarrow \mathbb{E}\left[g(\boldsymbol{\theta})\right] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad as \quad B \rightarrow \infty$$

This means we can approximate $\mathbb{E}\left[g(\boldsymbol{\theta})\right]$ with the sample average of $g(\boldsymbol{\theta}^{(1)}), \ldots, g(\boldsymbol{\theta}^{(B)})$, just as in Monte Carlo approximation. → That's why we call such approximations Markov Chain Monte Carlo (MCMC) approximations, and the procedure an MCMC algorithm.

The Gibbs sampler: an example

Normal model $y_i|\theta, \phi \sim \mathcal{N}(\theta, \phi) \forall i$, with the semi-conjugate prior distributions:

$$\theta|\theta_0, \phi_0 \sim \mathcal{N}(\theta_0, \phi_0), \qquad \phi|\nu_0, S_0 \sim \mathcal{IG}(\nu_0/2, S_0/2).$$

If a sample $\mathbf{y}$ is observed, the full conditionals of the model parameters are:

$$\theta|\phi, \mathbf{y} \sim \mathcal{N}(\theta_1, \phi_1), \qquad \phi|\theta, \mathbf{y} \sim \mathcal{IG}(a_1, b_1);$$

where

$$\theta_1 = \frac{\frac{\theta_0}{\phi_0} + \frac{n\bar{y}}{\phi}}{\frac{1}{\phi_0} + \frac{n}{\phi}}, \qquad \phi_1 = \frac{1}{\frac{1}{\phi_0} + \frac{n}{\phi}};$$

and

$$a_1 = \frac{\nu_0}{2} + \frac{n}{2}, \qquad b_1 = \frac{S_0}{2} + \frac{\sum_{i=1}^{n}(y_i - \theta)^2}{2}.$$

The posterior distributions can be easily obtained by MCMC methods.

See R script: `example_MCMC.R`