MCMC and Bayesian modeling in R

Sarah Kaspar
Centre for Statistical Data Analysis
2022-11-18

EMBL

Tutorial overview

- **Markov chain Monte Carlo simulation**
  - And why we need it

- **Demonstration in R**
  - Bayesian GLM for survey data
  - Bayesian LMM (Partial pooling) for Bayesball data

- **(Open coding / Exercises)**

Bayesian approach takes prior knowledge into account

EMBL

Likelihood
How compatible is the evidence with the hypothesis?

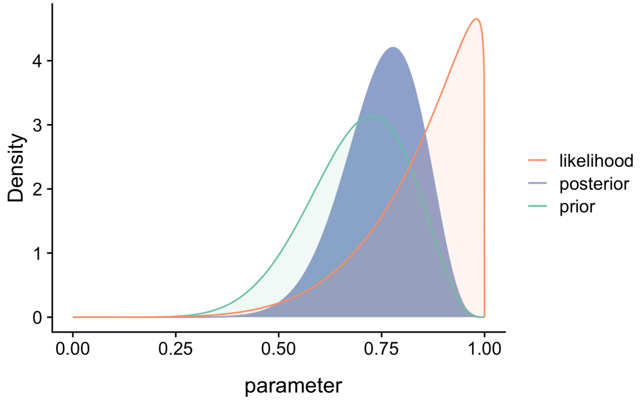
Prior distribution

Posterior distribution

$$P(\theta|data) = \frac{P(data|\theta) \cdot P(\theta)}{P(data)}$$

marginal likelihood
(Normalizing constant)

θ : The parameter we want to infer.



Methods overview: How to find a posterior

EMBL

- **Closed form solution**
  - Easy with conjugate priors
  - Difficult otherwise

- **Markov chain Monte Carlo simulation**

  Some implementations in R:

  - Established packages for specific data sets
    - DESeq2, edgeR, limma, …

  - rstanarm (wrapper for STAN programming)

Posterior distribution — Likelihood — Prior distribution

$$P(\theta|data) = \frac{P(data|\theta) \cdot P(\theta)}{P(data)}$$

marginal likelihood requires integrating over prior

In MCMC, we sample from this distribution:

$$P(\theta|data) \propto P(data|\theta) \cdot P(\theta)$$

This term is easy to solve for specific values of parameters and data.

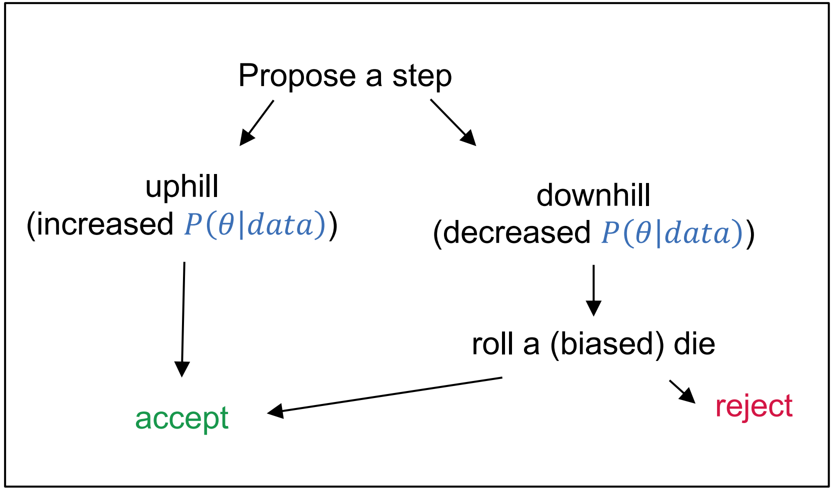
MCMC algorithm

EMBL

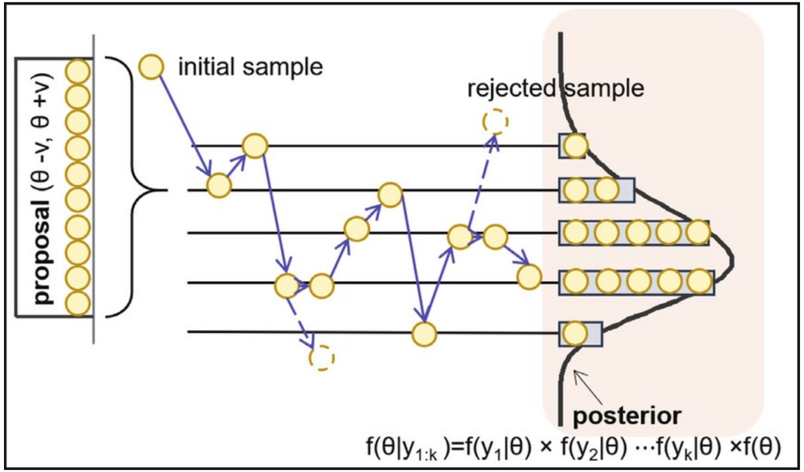
Sampling from the joint probability distribution:

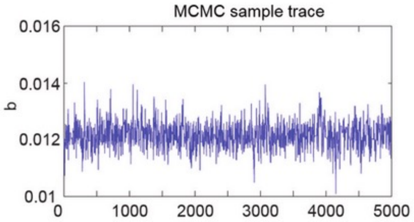
$$P(\theta|data) \propto P(data|\theta) \cdot P(\theta)$$

1. Randomly initialize parameter(s)
2. Random walk through parameter space:



The random walk visits locations with high $P(\theta|data)$ more often.





Graphics: Dong et al., 2019

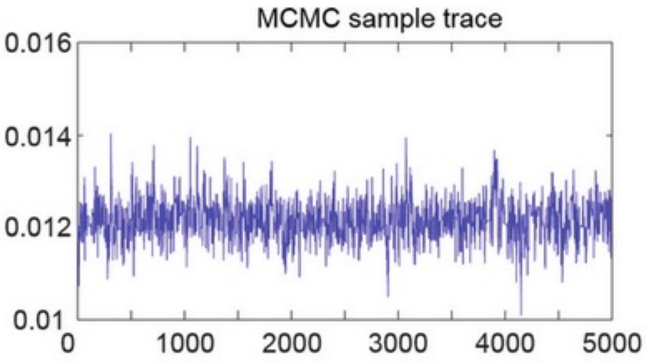
MCMC algorithm

The Markov Chain **approximates the posterior distribution**.

Posterior mode/mean and **credible intervals** can be approximated from the sample.

Practical aspects:

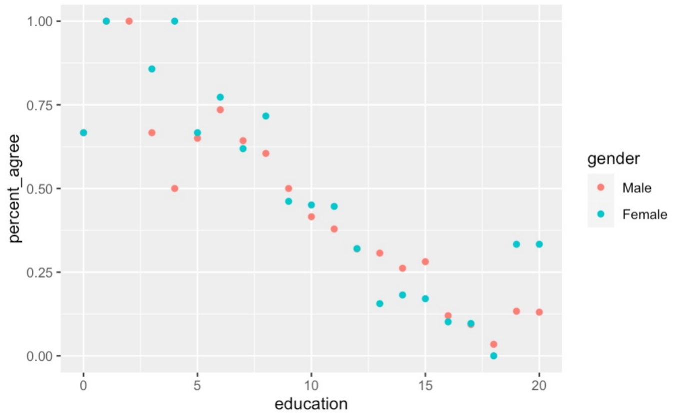
- There is a „burn in" period, before the chain converges.
- Not all chains converge.
- Initialize several chains and check whether they converge to the same parameters (otherwise they sample from local optima)



graphics: Dong et al., 2019

Example data – womens role

Survey: Who agrees to the statement: “Women should […] leave running the country up to men.”



education <int>	gender <fctr>	agree <dbl>	disagree <dbl>
0	Male	4	2
1	Male	2	0
2	Male	4	0
3	Male	6	3

```
glm(cbind(agree, disagree) ~ education + gender,
          family = binomial(link = "logit"))
```

Linear predictor:
$\eta = a + x^T b$

x: variables
a, : intercept
b_1, b_2: coefficients for education and gender

We put priors on these

Binomial likelihood for single observation:

$$\binom{n}{y}\left(\text{logit}(\eta)\right)^y \left(1 - \text{logit}(\eta)\right)^{n-y}$$

Joint distribution $P(data|\theta) \cdot P(\theta)$ for given parameters

$$f(a, b_1, b_2|y, X) \propto \underbrace{f(a) f(b_2) f(b_2)}_{\text{Joint distribution of the priors}} \times \prod_{i=1}^{J} \underbrace{\left(\text{logit}(\eta_i)\right)^{y_i}\left(1 - \text{logit}(\eta_i)\right)^{n_i - y_i}}_{\text{Binomial likelihood}}$$

Joint distribution of the priors

Binomial likelihood

Joint likelihood of all J data points

https://mc-stan.org/rstanarm/articles/rstanarm.html

Example data – what does rstanarm do?

EMBL

```
stan_glm(cbind(agree, disagree) ~ education + gender,
         family = binomial(link = "logit"),
         prior = student_t(df = 7, 0, 5),
         prior_intercept = student_t(df = 7, 0, 5))
```

→ Linear predictor η
→ logit link
→ prior distributions

Translate into joint distribution $P(data|\theta) \cdot P(\theta)$

$$f(a, b_1, b_2|y, X) \propto \underbrace{f(a)f(b_2)f(b_2)}_{\text{Joint distribution of the priors}} \times \underbrace{\prod_{i=1}^{J} \underbrace{(\text{logit}(\eta_i))^{y_i}(1 - \text{logit}(\eta_i))^{n_i - y_i}}_{\text{Binomial likelihood}}}_{\text{Joint likelihood of all J data points}}$$

Run a series of MCMC chains that sample from this distribution

https://mc-stan.org/rstanarm/articles/rstanarm.html

References

Dong, Ting & An, Dawn & Kim, Nam. (2019). Prognostics 102: Efficient Bayesian-Based Prognostics Algorithm in MATLAB. 10.5772/intechopen.82781.

Example from: https://mc-stan.org/rstanarm/articles/rstanarm.html