Applied Bayesian Data Analysis — Chapter 7

Kim Albertsson

CERN and Luleå University of Technology

kim.albertsson@ltu.se

October 23, 2019

Chapter 7

Markov Chain Monte Carlo:
Weighted random walks, how to, and their properies.

Personal reflections

- The politician island thought experiment felt convoluted. Took some time to understand.
- Sample representation starts making sense. Still unclear, to me, why the problems of grid based approaches do not apply here.
- Working through the math here was fun and rewarding.
- Did not have time to go into details on MCMC accuracy etc. Seems interesting!

Today

- Unnormalised distributions,
- Representative Sample, and
- Two sampling techniques (Metropolis, Gibbs).

Introduction

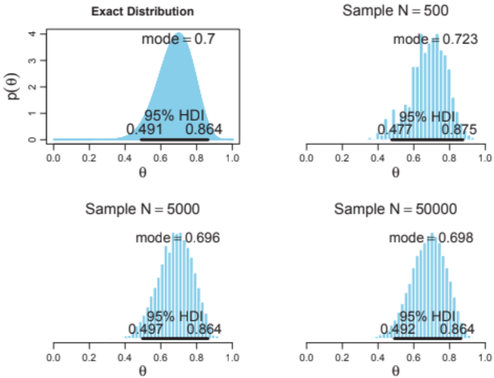
- Analytical solution (ch. 6):
  Feasible for limited set of prior-likelihood combinations (e.g. conjugate priors).
- Discretisation on grid (ch. 5):
  Curse of dimensionality. (E.g. 6 params with 1000 possibilities each $\rightarrow 1000^6 = 10^{18}$ cells in matrix).
- Representative sample (curr. ch.):
  Distribution approximated with samples drawn from distribution.

Representative Sample

Representative sample:

- Assumes: $p(\theta)$, $p(D|\theta)$ calculable up to multiplicative constant.
- Output: $p(\theta|D)$ as collection of samples to calculate e.g. central tendency and HDI.

For intuition: Compare representative sample to polling in politics.



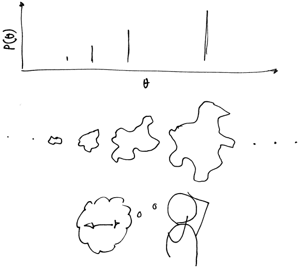
A Simple Case of the Metropolis Algorithm (I)

Example: Politician of Island chain

- Sequential islands
- Goal: Spend time on island proportional to population
- Heueristic:

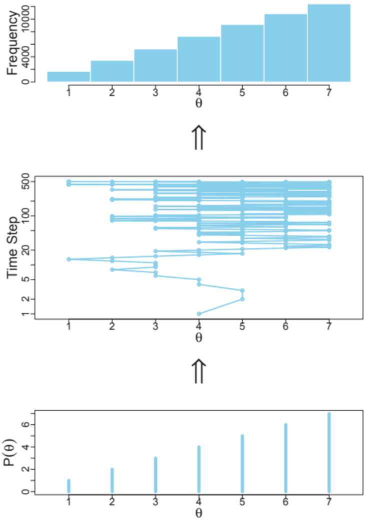
$$p_{move} = \begin{cases} 1, & \text{if } P_{propsed} > P_{current} \\ \frac{P_{proposed}}{P_{current}}, & \text{otherwise} \end{cases}$$

- Metropolis algorithm



A Simple Case of the Metropolis Algorithm (II)

- $P(\cdot)$, relative population. Note: Not normalised!
- Top: Relative frequency of visit *after long time*.
- Middle: One possible trajectory
- Bottom: True distribution



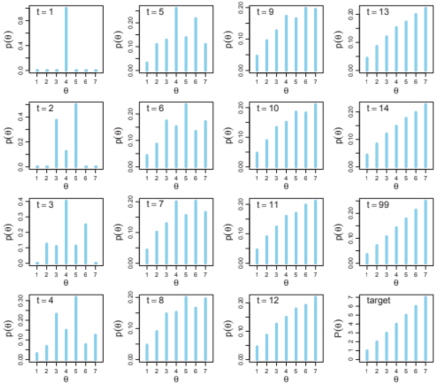
A Simple Case of the Metropolis Algorithm (III)

Analysis for this, *simple*, case:

- Proposal distribution: Moves and probabilities. Now: $p(\mathrm{left}) = 0.5$, and $p(\mathrm{right}) = 0.5$.
- At time $t = 1$ 100% chance of being in starting position.
- For $t = 2$:

$$p(\theta = 3) = 0.5 \cdot \left({}^{P(3)}/_{P(4)}\right)$$
$$p(\theta = 4) = 0.5 \cdot \left(1 - {}^{P(3)}/_{P(4)}\right)$$
$$p(\theta = 5) = 0.5 \cdot (1)$$

- For $t = n$: Iterate...



A Simple Case of the Metropolis Algorithm (IV)

- Uses *proposal distribution*, *acceptance criterion*, and target distribution *ratio*.
- Convergence to $P(\theta)$ (up to multiplicative constant).

Short form acceptance probability for move:

$$p_{\text{move}} = \min\left(\frac{P(\theta_{\text{proposed}})}{P(\theta_{\text{current}})}, 1\right) \tag{7.1}$$

Intuition for convergence:

$$\frac{p(\theta \rightarrow \theta + 1)}{p(\theta + 1 \rightarrow \theta)} = \frac{0.5 \min P(\theta + 1)/P(\theta)}{0.5 \min P(\theta)/P(\theta + 1)} = \frac{P(\theta + 1)}{P(\theta)} \tag{7.2}$$

Favour travelling to $P(\theta + 1)$ more than $P(\theta)$.
(Details: Target distribution is fix point of transition matrix)

The Metropolis Algorithm more generally

Idea: Random walk through parameter space.

- Proposal distribution, e.g. $p(M) \mathrm{for} M \in \mathrm{left}, \mathrm{right}$, or $p(M) \sim \mathcal{N}$.
- *Unnormalised* target distribution. Must be calculable for any θ e.g. $P(\theta) = p(\theta|D)p(\theta)$.
- Acceptance criterion, e.g. ${}^{P(\theta_{t+1})}/_{P\theta_t}$

Propose a move in paramter space. Use *acceptance criterion* to shape distribution of generated samples to match *target distribution*.

Contiuous case (I)

Example: Estimate bias of coin given some data and prior. Coin bias: Continuous chain of tiny islands
Note: $P(\theta)$ is tractable b.c. $p(D|\theta)$ and $p(\theta)$ are tractable.

- Generate jump $\Delta\theta \sim \mathcal{N}(\wr, \sigma)$. $\theta_{\text{pro}} = \theta_{\text{cur}} + \Delta\theta$.
- Calculate accpetance
  $p_{\text{move}} = \min\left(1, \frac{\theta_{\text{pro}}^{z}(1-\theta_{\text{pro}})^{N-z}\theta_{\text{pro}}^{(}a-1)(1-\theta_{\text{pro}})^{(}b-1)/B(a,b)}{\theta_{\text{cur}}^{z}(1-\theta_{\text{cur}})^{N-z}\theta_{\text{cur}}^{(}a-1)(1-\theta_{\text{cur}})^{(}b-1)/B(a,b)}\right)$
- Accept move if $x \sim \mathcal{U}$ is less than p_{move}, else tally θ_{cur} again.

Note:

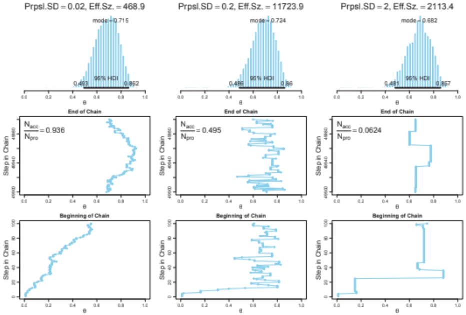
- More data $\rightarrow$ narrower posterior.
- More samples $\rightarrow$ better approximation of posterior.

Contiuous case (II)

Example run

- Left (small steps): High acceptance, inefficient exploration. Low ESS.
- Mid (medium steps): Medium acceptance, most efficient exploration (of the 3). High ESS.
- Right (large steps): Low acceptance, inefficient exploration. Low ESS.

ESS: Effective sample size



Aside: Effective Sample Size

Caveat: I might have misunderstood this!

Samples in MCMC chain can be correlated.

Effective sample size (ESS) estimates the correlation and reports how many *uncorrelated*, i.e. direct, samples of the target distribution the chain corresponds to.

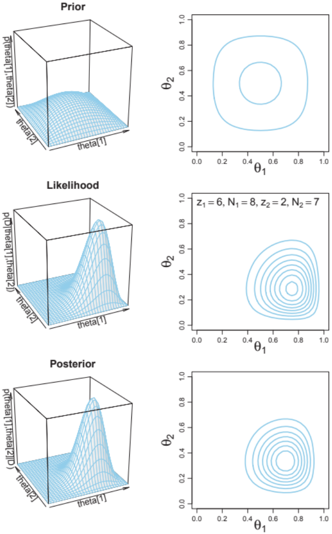
Source: `http://www.nowozin.net/sebastian/blog/effective-sample-size-in-importance-sampling.html`

Toward Gibbs sampling (I)

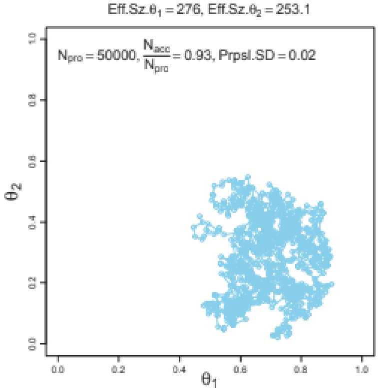
Alternative to Metropolis: Gibbs sampling, usually for more than one parameter.

Suppose two coins, each with bias: $\theta = \{\theta_1, \theta_2\}$. Samples from each coin independent.

$$
\begin{aligned}
p(\theta) =& \beta(\theta|a,b) \\
p(D|\theta) =& \mathrm{Bernoulli}(D|\theta) \\
p(\theta_1,\theta_2|D) =& \frac{\theta_1^{z_1+a_1-1}(1-\theta_1)^{N_1-z_1+b_1-1}}{B(z_1+a_1, N_1-z_1+b_1)} \\
& \frac{\theta_2^{z_2+a_2-1}(1-\theta_2)^{N_2-z_2+b_2-1}}{B(z_2+a_2, N_2-z_2+b_2)}
\end{aligned}
$$



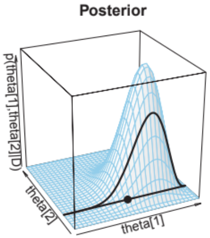
Metropolis estimate of posterior:

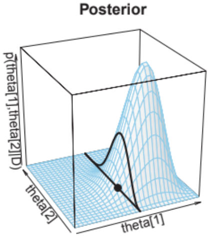


Gibbs sampling (I)

Idea: Walk only in one paramter direction at a time. Cycle paramters to update.

- Requires tractable $p(\theta_i | \{\theta_{i \neq j}\}, D)$
- Cycle paramters $(\theta_1, \theta_2, ..., \theta_1, \theta_2, ..., ...)$ to better cover paramter space.
- Special case of Metropolis: variable proposal distribution





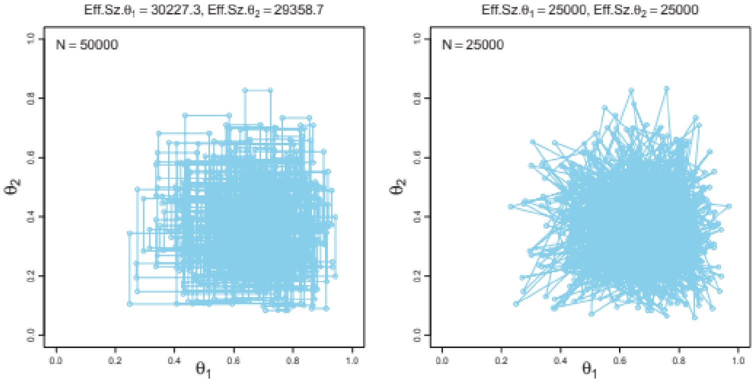
Gibbs sampling (II)

Becase the proposal distribution exatcly mirrors the posterior probability for that paramter, the move is always accepted.

Linger at θ_1, building up approximiation of $P(\theta_1, \theta_2)$ for that value. Gibbs sampling does this lingering only one sample at a time.

Gibbs sampling (III)

Gibbs sampling estimate of the posterior:



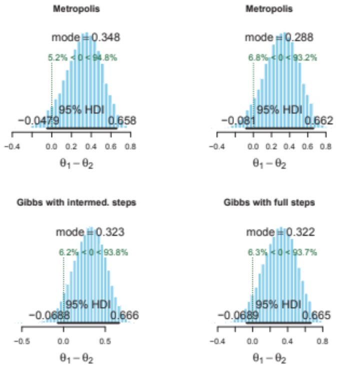
Representing the posterior

Histogram the representative sample to approximate posterior distribution.

Figure compares results from 4 different generated samples.

Computes estimations of mode and HDI, e.g. for comparing if statistically significant difference.

In limit, all dists should look the same.



End!

Representativeness

Problem: Starting point could be in low-prob area. Algorithm could get stuck in part of the distribution.

Test for convergence (is sample representative): Difficult — State of the art: Eyeball it.

Burn-in: Up to several thousand samples. Initial samples can be highly non-representative.

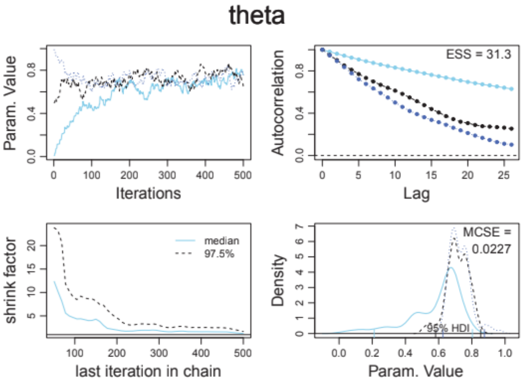


Figure 7.11 — Todo

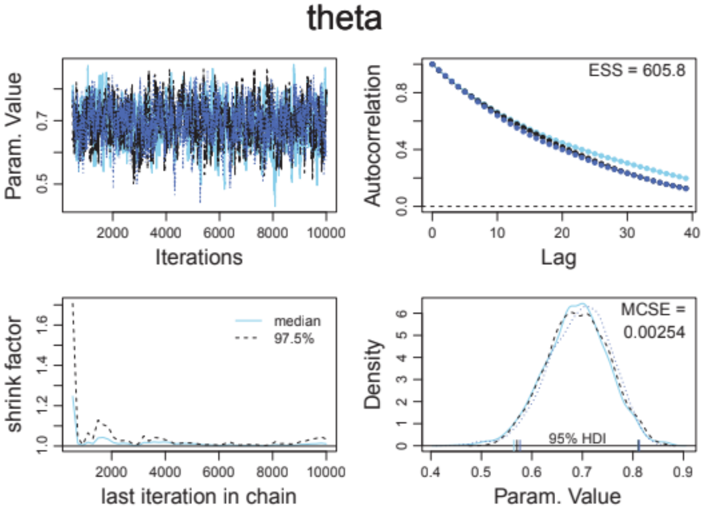


Figure 7.12 — Todo

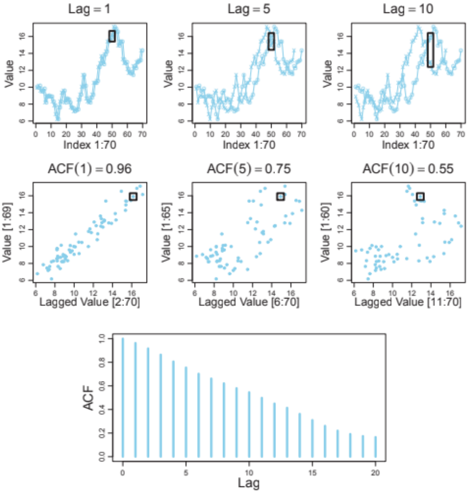


Figure 7.13 — Todo

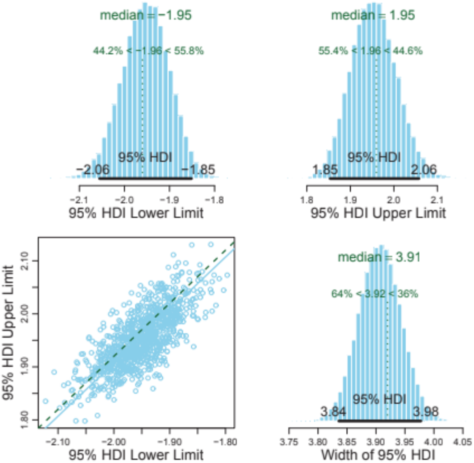


Figure 7.14 — Todo