Markov chain Monte Carlo (MCMC)

TMA4300: Computer Intensive Statistical Methods
(Spring 2014)

Andrea Riebler

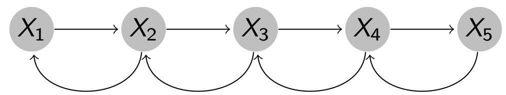
Introduction

Application of ordinary Monte Carlo methods is difficult if the unknown parameter is of high dimension. However, Markov chain Monte Carlo (MCMC) methods will then be a useful alternative.

en.wikipedia.org/wiki/Markov_chain

Andrey Markov (1856 – 1922),
Russian mathematician.

Markov chain:



Given the previous observation X_{i-1}, X_i is independent of the sequence of events that preceded it.

Markov chain Monte Carlo

Idea

Simulate a Markov chain $X_1, \dots, X_i, \dots$, which is designed in a way such that $\mathrm{P}(X_i = x)$ converges to the target distribution, e.g. the posterior distribution.

Properties:

- After convergence, one obtains random samples from the target distribution, which can be used to estimate posterior characteristics.
- Samples will typically be dependent.

Terminology:

- Burn-in: The initial samples may follow a very different distribution. These initial samples are thrown away.
- Thinning: Only keep every k-th iteration to save storage space.

Metropolis-Hastings algorithm

There is great liberty in the actual choice of the underlying random samples of MCMC algorithms.

- Random samples typically depend on the current state X_i of the Markov chain.
- Random samples are generated from some arbitrary proposal distribution with density $Q(x_i|x_{i-1})$, say.

The Metropolis–Hastings algorithm is the most general approach.

History of Metropolis-Hastings

- The algorithm was presented 1953 by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller from the Los Alamos group. It is named after the first author Nicholas Metropolis.
- W. Keith Hastings extended it to the more general case in 1970.
- It was then ignored for a long time.
- Since 1990 it has been used more intensively.

Metropolis-Hastings algorithm

Metropolis-Hastings counts a rejection as an iteration and records the past value as the current value.

```
Init x_0 ~ g(x_0)
for i = 1 to N − 1 do
    u ~ U(0, 1)
    Generate a proposal x* ~ Q(x|x_{i−1})
    if u < min(1, π(x*)/π(x_{i−1}) × Q(x_{i−1}|x*)/Q(x*|x_{i−1})) then
        x_i ← x*
    else
        x_i ← x_{i−1}
    end if
end for
```

$$\textbf{if } u < \underbrace{\min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})} \times \underbrace{\frac{Q(x_{i-1}|x^\star)}{Q(x^\star|x_{i-1})}}_{\text{Proposal ratio}}\right)}_{\text{Acceptance probability } \alpha} \textbf{ then}$$

Remarks on the Metropolis-Hastings algorithm

- Under some regularity conditions it can be shown that the Metropolis-Hasting algorithm converges to the target distribution regardless of the specific choice of $Q(x|x_{i-1})$.
- However, the speed of convergence and the dependence between the successive samples depends strongly on the proposal distribution.
- Since we only need to compute the ratio $\pi(x^\star)/\pi(x_{i-1})$, the proportionality constant is irrelevant.
- Similarly, we only care about $Q(.)$ up to a constant.
- Often it is advantageous to calculate the acceptance probability on log-scale, which makes the computations more stable.

For more comments and details see: Chib, S. and Greenberg, E. (1995), *Understanding the Metropolis-Hastings algorithm, The American Statistician, 49: 327–335*

Special cases of the Metropolis-Hastings algorithm

Metropolis algorithm

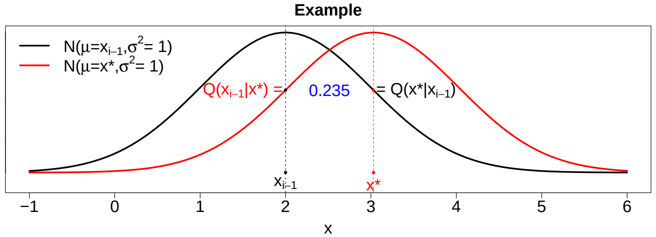
The proposal density is symmetric around the current values, that means

$$Q(x_{i-1}|x^\star) = Q(x^\star|x_{i-1}).$$

Hence,

$$\alpha = \min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})} \times \frac{Q(x_{i-1}|x^\star)}{Q(x^\star|x_{i-1})}\right) = \min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})}\right)$$

A particular case is the random walk proposal, defined as the current value x_{i-1} plus a random variate of a 0-centered symmetric distribution.



Special cases of the Metropolis-Hastings algorithm

Independence proposal

- The proposal distribution does not depend on the current value $\boldsymbol{\theta}^{(i)}$

$$Q(x|x_{i-1}) = Q(x).$$

- The sampler is closer to rejection sampler. However, here if we reject, then we retain the sample.

Acceptance probability always equals unity

This is the case if

$$Q(x^\star|x_{i-1}) = \pi(x^\star),$$

which seems of limited value, as direct sampling from the target density was assumed to be unavailable.
However, let $\boldsymbol{x} = (x^1, \ldots, x^p)$ and consider updating a specific component x^j of $\boldsymbol{x}$ by a sample from $\pi(x^j|\boldsymbol{x}^{-j})$, where $\boldsymbol{x}^{-j}$ denotes $\boldsymbol{x}$ without x^j. Here,

$$\pi(x^j|\boldsymbol{x}^{-j}) \propto \pi(\boldsymbol{x}), \quad j = 1, \ldots, p$$

so that

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{x}^\star)}{\pi(\boldsymbol{x}_{i-1})} \times \frac{\pi(x^j_{i-1}|\boldsymbol{x}^{-j,\star})}{\pi(x^{j,\star}|\boldsymbol{x}^{-j}_{i-1})}\right) = 1$$

Gibbs sampling

Iterative update of all components of $\boldsymbol{x}$. Let $\boldsymbol{x} = (x^1, \ldots, x^p)$. To obtain samples $\boldsymbol{x}_i$ from the joint target distribution $\mathrm{p}(\boldsymbol{x})$ do the following:

- Initialize $\boldsymbol{x}_0$ and let $i = 0$.
- Repeatedly:

$$\begin{aligned}
&\text{Sample} \quad x^1_{i+1} \sim \pi(x^1|\boldsymbol{x}^{-1}) \\
&\text{Sample} \quad x^2_{i+1} \sim \pi(x^2|\boldsymbol{x}^{-2}) \\
&\text{Sample} \quad x^3_{i+1} \sim \pi(x^3|\boldsymbol{x}^{-3}) \\
&\vdots \\
&\text{Sample} \quad x^p_{i+1} \sim \pi(x^p|\boldsymbol{x}^{-p}) \\
&\text{Set i= i+1}
\end{aligned}$$

It is possible to do this block-wise, i.e. sample blocks of the x^j together.

Remarks on Gibbs sampling

- High dimensional updates of $\boldsymbol{x}$ can be boiled down to scalar updates.
- Visiting schedule: Various approaches exist (and can be justified) to ordering the variables in the sampling loop. One approach is random sweeps: variables are chosen at random to resample.
- Gibbs sampling assumes that it is easy to sample from the full-conditional distribution. This is sometimes not so easy. Alternatively, a Metropolis-Hastings proposal can be used for the j-th component, i.e. Metropolis-within-Gibbs.
- Care must be taken when improper prior are used, which may lead to an improper posterior distribution. Impropriety implies that there does not exist a joint density to which the full-conditional distributions correspond.

  Hobert, J. P. and Casella, G. (1996), JASA, 91: 1461–1473.

Efficiency of the Metropolis-Hastings algorithm

The efficiency and performance of the Metropolis-Hastings algorithm depends crucially on the relative frequency of acceptance.

An acceptance rate of one is not always good. Consider the random walk proposal:

- Too large acceptance rate $\Rightarrow$ Slow exploration of the target density.
- Too small acceptance rate $\Rightarrow$ Large moves are proposed, but rarely accepted.

Tuning the acceptance rate:

- For random walk proposals, acceptance rates between 30% and 50% are typically recommended. They can be achieved by changing the variance of the proposal distribution.
- For independence proposals a high acceptance rate is desired, which means that the proposal density is close to the target density.