Tutorial Lectures on MCMC I

Sujit Sahu [a]

University of Southampton

http://www.maths.soton.ac.uk/staff/sahu/

Utrecht: August 2000.

- Introduction to MCMC, especially for computation in Bayesian Statistics.
- Basic recipes, and a sample of some techniques for getting started.
- No background in MCMC assumed.
- Not for experts!

---

[a] In close association with Gareth Roberts

Markov Chain Monte Carlo (MCMC)

Introduction

Outline:

- Motivation
- Monte Carlo integration
- Markov chains
- MCMC

Bayesian Inference

Data: Y (realisation y)

Parameters, latent variables:

$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$

Likelihood: $L(y|\boldsymbol{\theta})$

Prior: $\pi_0(\boldsymbol{\theta})$

Inference is based on the *joint posterior*

$$\begin{aligned} \pi(\boldsymbol{\theta}|y) &= \frac{L(y|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{\int L(y|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto L(y|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}) \\ i.e. \quad Posterior &\propto Likelihood \times Prior \end{aligned}$$

Example 1

Let $Y_1, \dots, Y_n \overset{i.i.d.}{\sim} N(\theta, 1)$ and $\pi_0(\theta) = \frac{1}{\pi\,(1+\theta^2)}$.

Posterior:

$$\begin{aligned} \pi(\theta|y) &\propto \exp\left\{-\frac{\sum_{i=1}^n (y_i-\theta)^2}{2}\right\} \times \frac{1}{1+\theta^2} \\ &\propto \exp\left\{-\frac{n(\theta-\bar{y})^2}{2}\right\} \times \frac{1}{1+\theta^2}. \end{aligned}$$

Things of interest to Bayesians:

- Posterior Mean = $\mathbb{E}(\theta|y)$.
- Posterior Variance = var$(\theta|y)$.
- Credible interval $\{a(y), b(y)\}$ for θ s.t. $Pr\{a(y) < \theta < b(y)|y\} = 0.95$.

Example 2

Data $Y_1, \dots, Y_n$ are a random sample from $N(\mu, \sigma^2)$. Non-informative prior is:

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2},$$

Joint posterior:

$$\begin{aligned} \pi(\mu, \sigma^2 | y) &\propto \left(\tfrac{1}{\sigma^2}\right)^{n/2+1} \\ &\times \exp\left\{-\tfrac{\sum (y_i - \mu)^2}{2\sigma^2}\right\} \end{aligned}$$

which is not of standard from.

Outline:

- Motivation
- $\boxed{\text{Monte Carlo integration}}$
- Markov chains
- MCMC

General problem: evaluating

$$\mathbb{E}_\pi[h(X)] = \int h(x)\pi(x)dx$$

can be difficult. ($\int |h(x)|\pi(x)dx < \infty$).

However, if we can draw samples

$$X^{(1)}, X^{(2)}, \ldots, X^{(N)} \sim \pi(x)$$

then we can estimate

$$\mathbb{E}_\pi[h(X)] \approx \bar{h}_N = \frac{1}{N}\sum_{t=1}^{N} h\left(X^{(t)}\right).$$

This is *Monte Carlo (MC) integration*

Changed notation:

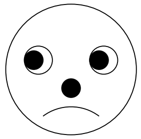
$\boldsymbol{\theta} \equiv x;\ \ \pi(\boldsymbol{\theta}|Y) = \pi(x)$

Consistency

For independent samples, by Law of Large numbers,

$$\begin{aligned} \bar{h}_N &= \frac{1}{N}\sum_{t=1}^{N} h\left(X^{(t)}\right) \\ &\rightarrow \mathbb{E}_\pi[h(X)] \text{ as } N \rightarrow \infty. \end{aligned} \quad (1)$$

But independent sampling from $\pi(x)$ may be difficult.



It turns out that (1) still applies if we generate samples using a Markov chain.

But first, some revision of Markov chains.

Outline:

- Motivation
- Monte Carlo integration
- Markov chains
- MCMC

A Markov chain is generated by sampling

$$X^{(t+1)} \sim p(x|x^{(t)}), t = 1, 2, \dots .$$

p is the transition kernel.

So $X^{(t+1)}$ depends only on $X^{(t)}$, not on $X^{(0)}, X^{(1)}, \dots, X^{(t-1)}$.

$$p(X^{(t+1)}|x^{(t)}, x^{(t-1)}, \dots) = p(X^{(t+1)}|x^{(t)})$$

For example:

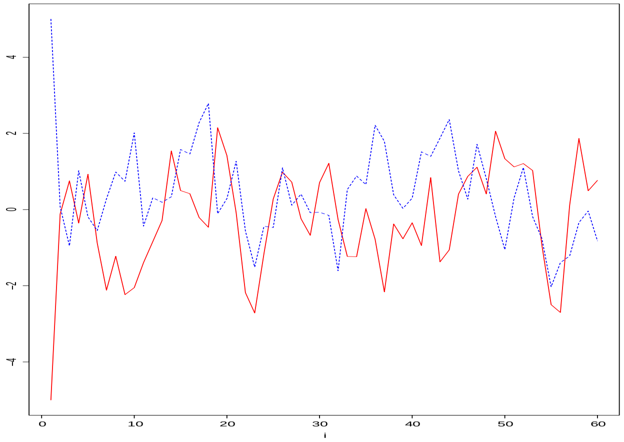
$$X^{(t+1)}|x^{(t)} \sim N(0.5\, x^{(t)}, 1.0).$$

This is called a first order *auto-regressive process* with lag-1 auto-correlation 0.5

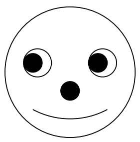
Simulation of the chain:

$$X^{(t+1)}|x^{(t)} \sim N(0.5\,x^{(t)}, 1.0).$$

Two different starting points are used.



After about 5–7 iterations the chains seemed to have forgotten their starting positions.



Stationarity

As $t \to \infty$, the Markov chain converges to its *stationary* distribution.

in distribution

or invariant

In the above example, this is

$$X^{(t)}|x^{(0)} \sim N(0.0, 1.33), \text{ as } t \to \infty$$

which does not depend on $x^{(0)}$.

Does this happen for all Markov chains?

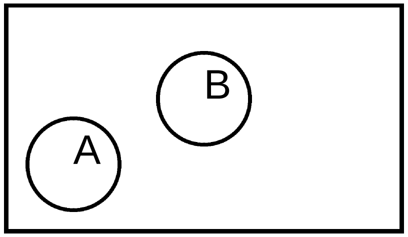
Irreducibility

Assuming a stationary distribution exists, it is unique if the chain is *irreducible*.

Irreducible means any set of states can be reached from any other state in a finite number of moves.

An example of a reducible Markov chain:

Suppose $p(x|y) = 0$ for $x \in A$ and $y \in B$ and vice versa.



Aperiodicity

A Markov chain taking only finite number of values is *aperiodic* if greatest common divisor of return times to any particular state i say, is 1.

- Think of recording the number of steps taken to return to the state 1. The g.c.d. of those numbers should be 1.
- If the g.c.d. is bigger than 1, 2 say, then the chain will return in cycles of 2, 4, 6, ... number of steps. This is not allowed for aperiodicity.
- Definition can be extended to general state space case.

Ergodicity

Assume the Markov chain:

- has the stationary distribution $\pi(x)$
- is aperiodic and irreducible.

then we have an *ergodic theorem*:

$$\begin{aligned} \bar{h}_N &= \frac{1}{N}\sum_{t=1}^{N} h\left(X^{(t)}\right) \\ &\to \mathbb{E}_\pi[h(X)] \quad \text{as } N \to \infty. \end{aligned}$$

$\bar{h}_N$ is called an ergodic average.

Also for such chains with

$$\sigma_h^2 = \text{var}_\pi[h(X)] < \infty$$

- the central limit theorem holds and
- convergence occurs geometrically.

Numerical standard errors (nse)

The nse of $\bar{h}_N$ is $\sqrt{\text{var}_\pi(\bar{h}_N)}$, and for large N

$$\text{nse}\left(\bar{h}_N\right) \approx \sqrt{\frac{\sigma_h^2}{N}\left\{1 + 2\sum_{l=1}^{N-1}\rho_l(h)\right\}}$$

where $\rho_l(h)$ is the lag-l auto-correlation in $\left\{h(X^{(t)})\right\}$.

- In general no simpler expression exist for the nse.
- See Geyer (1992), Besag and Green (1993) for ideas and further references.

- If $\left\{h(X^{(t)})\right\}$ can be approximated as a first order auto-regressive process then

  $$\text{nse}\left(\bar{h}_N\right) \approx \sqrt{\frac{\sigma_h^2}{N}\frac{1+\rho}{1-\rho}},$$

  where ρ is the lag-1 auto-correlation of $\left\{h(X^{(t)})\right\}$.
- The first factor is the usual term under independent sampling.
- The second term is usually > 1.
- And thus is the penalty to be paid because a Markov chain has been used.

Moreover,

- the nse may not be finite in general.
- it is finite if the chain converges geometrically
- If the nse is finite, then we can make it as small as we like by increasing N.
- the ‘obvious’ estimator of nse is not consistent.

See later.

Markov chains – summary

- A Markov chain may have a stationary distribution.
- The stationary distribution is unique if the chain is irreducible.
- We can estimate nse’s if the chain is also geometrically convergent.

Where does this all get us?

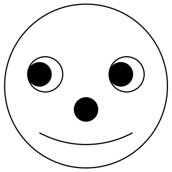
Outline:

- Motivation
- Monte Carlo integration
- Markov chains
- MCMC

How do we construct a Markov chain whose stationary distribution is our target distribution, $\pi(x)$?

Metropolis *et al* (1953) showed how.

The method was generalized by Hastings (1970).



This is called

Markov chain Monte Carlo (MCMC).

Metropolis-Hastings algorithm

At each iteration t

Step 1 Sample $y \sim q\left(y|x^{(t)}\right)$.

“candidate” point “Proposal” distribution

Step 2 With probability

$$\alpha(x^{(t)}, y) = \min\left\{1, \frac{\pi(y)q\left(x^{(t)}|y\right)}{\pi\left(x^{(t)}\right)q\left(y|x^{(t)}\right)}\right\}$$

set

$$x^{(t+1)} = y \quad \text{(acceptance)},$$

else set

$$x^{(t+1)} = x^{(t)} \quad \text{(rejection)}.$$

Note that:

- The normalising constant in $\pi(x)$ is not required to run the algorithm. It cancels in the ratio.
- If $q(y|x) = \pi(y)$, then we obtain independent samples.
- Usually q is chosen so that $q(y|x)$ is easy to sample from.
- Theoretically, any density $q(\cdot|x)$ having the same support should work. However, some q's are better than others. See later.
- The induced Markov chains have the desirable properties under mild conditions on $\pi(x)$.

Implementing MCMC

- Flavours of Metropolis-Hastings
- Gibbs Sampler
- Number of Chains
- Burn-in and run length
- Numerical standard errors

The Metropolis algorithm

Proposal is symmetric:

$$q(x|y) \equiv q(y|x)$$

– as proposed by Metropolis *et al.* (1953).

Special case: Random-walk Metropolis

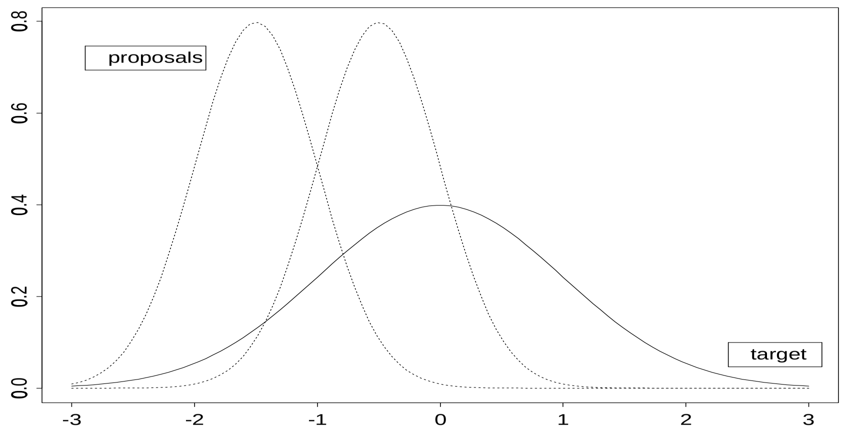
$$q(x|y) \equiv q(|y-x|).$$

In this case:

$$\alpha(x^{(t)}, y) = \min\left\{1, \frac{\pi(y)}{\pi\left(x^{(t)}\right)}\right\}$$

Example:

$$\begin{aligned} \pi(x) &\propto \exp\left\{-\frac{x^2}{2}\right\} \\ q(y|x) &\propto \exp\left\{-\frac{(y-x)^2}{2(0.5)^2}\right\} \end{aligned}$$



Proposal depends on where you are.

The Independence Sampler

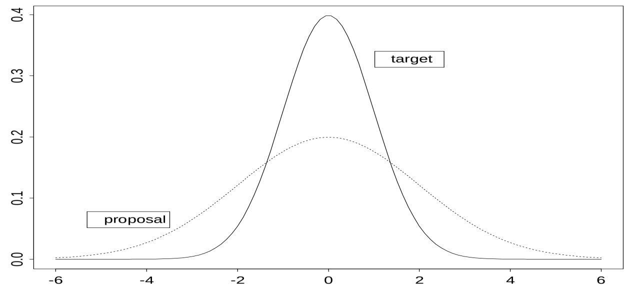
Proposal does not depend on x :

$$q(y|x) \equiv q(y)$$

So $\alpha(x, y)$ has a simpler form.

Beware: Independence samplers are either very good or very bad.

Tails of $q(y)$ must be heavier than tails of $\pi(x)$ for geometric convergence.



Return to the Normal-Cauchy example.

Example 1: Let

$Y_1, \ldots, Y_n \sim i.i.d. N(\theta, 1)$ and

$\pi_0(\theta) = \frac{1}{\pi\,(1+\theta^2)}$.

Posterior:

$$\pi(\theta|y) \propto \exp\left\{-\frac{n(\theta - \bar{y})^2}{2}\right\} \times \frac{1}{1+\theta^2}.$$

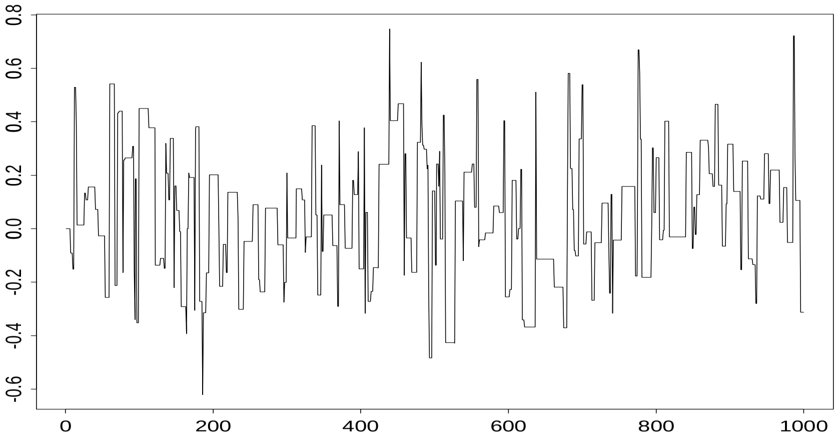
Suppose $n = 20$, $\bar{y} = 0.0675$. With the x notation we have

$$\pi(x) \propto \exp\left\{-\frac{n(x - 0.0675)^2}{2}\right\} \times \frac{1}{(1+x^2)}.$$

Example continued...

Let $q(y|x) = \frac{1}{\pi\,(1+y^2)}$.

Running the independence sampler gives:



	True.mean	M.mean	nse	lag-1.cor
X	0.0620	0.0612	0.006	0.172

- Flavours of Metropolis-Hastings
- Gibbs Sampler
- Number of Chains
- Burn-in and run length
- Numerical standard errors

Gibbs sampling

Suppose that $x = (x_1, x_2, \dots, x_k)$ is $k (\geq 2)$ dimensional.

Gibbs sampler uses what are called the full (or complete) conditional distributions:

$$\pi(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k)$$

$$= \frac{\pi(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_k)}{\int \pi(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_k) dx_j}.$$

Note that the conditional

$$\pi(x_j | x_1, \dots, x_{j-1}, x_{j+1}, ..., x_k)$$

is proportional to the joint. Often this helps in finding it.

Gibbs sampling

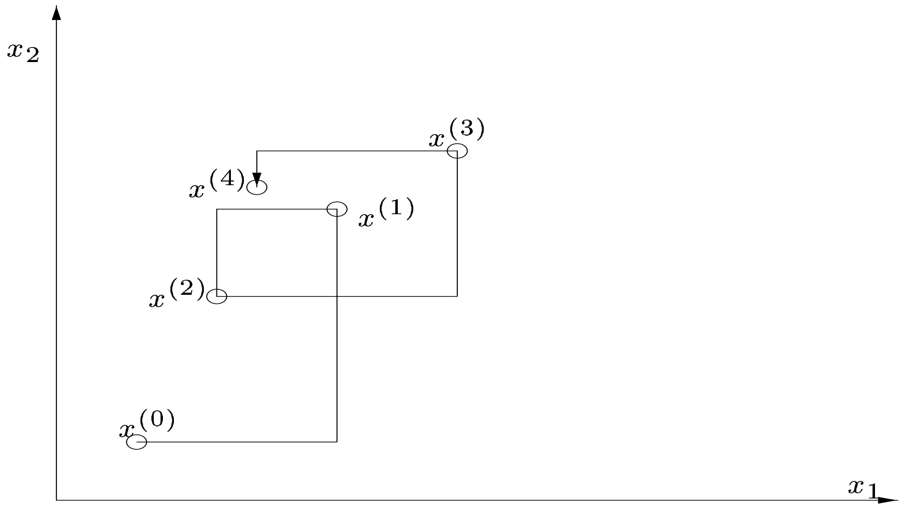
Sample or update in turn:

$$\begin{array}{lcl}
X_1^{(t+1)} & \sim & \pi(x_1 | x_2^{(t)}, x_3^{(t)}, \cdots, x_k^{(t)}) \\
X_2^{(t+1)} & \sim & \pi(x_2 | x_1^{(t+1)}, x_3^{(t)}, \cdots, x_k^{(t)}) \\
X_3^{(t+1)} & \sim & \pi(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \cdots) \\
\vdots & \vdots & \vdots \\
X_k^{(t+1)} & \sim & \pi(x_k | x_1^{(t+1)}, x_2^{(t+1)}, \cdots, x_{k-1}^{(t+1)})
\end{array}$$

Always use the most recent values.

Thus in two dimensions $(k = 2)$, the sample path of the Gibbs sampler will look something like:

ts



Example 2.

Let $Y_i \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ and $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$.

We had:

$$\begin{aligned} \pi(\mu, \sigma^2|y) &\propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} \\ &\times \exp\left\{-\frac{\sum(y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Let $\tau = 1/\sigma^2$. Easy to derive:

$$\begin{aligned} \pi(\mu|\sigma^2, y) &= N(\bar{y}, \sigma^2/n) \\ \pi(\tau|\mu, y) &= \Gamma\left(\frac{n}{2}, \frac{1}{2}\sum\left(y_i - \mu\right)^2\right) \end{aligned}$$

Sampling from full conditionals

We must be able to sample from

$$\pi(x_j|x_1,\dots,x_{j-1},x_{j+1},\dots,x_k)$$

to do Gibbs sampling.

In real problems, full conditionals often have complex algebraic forms, but are usually (nearly) log-concave.

For (nearly) log-concave univariate densities, use adaptive rejection sampling (Gilks and Wild, 1992) and follow-ups.

They have codes (`C` and `Fortran`) available from

www.mrc-bsu.cam.ac.uk

- Flavours of Metropolis-Hastings
- Gibbs Sampler
- $\boxed{\text{Number of Chains}}$
- Burn-in and run length
- Numerical standard errors

How many parallel chains of MCMC should be run ?

Experiment yourself.

- Several long runs (Gelman and Rubin, 1992)
  - gives indication of convergence
  - A sense of statistical security.
- one very long run (Geyer, 1992)
  - reaches parts other schemes cannot reach.

- Flavours of Metropolis-Hastings
- Gibbs Sampler
- Number of Chains
- $\boxed{\text{Burn-in and run length}}$
- Numerical standard errors

Early iterations $x^{(1)}, \ldots, x^{(M)}$ reflect starting value $x^{(0)}$.

These iterations are called *burn-in*.

After the burn-in, we say the chain has ‘converged’.

Omit the burn-in from ergodic averages:

$$\bar{h}_{MN} = \frac{1}{N-M} \sum_{t=M+1}^{N} h\left(X^{(t)}\right).$$

Methods for determining M are called *convergence diagnostics*.

Convergence Diagnostics

Must do:

- Plot the time series for each quantity of interest.
- Plot the auto-correlation functions.

If not satisfied, try some other diagnostics. See for example:

Gelman and Rubin (1992), Robert (1998), Cowles and Carlin (1996) Brooks and Roberts (1998).

But realise that you *cannot* prove that you have converged using any of those.

- Flavours of Metropolis-Hastings
- Gibbs Sampler
- Number of Chains
- Burn-in and run length
- $\boxed{\text{Numerical standard errors}}$

Suppose we decide to run the chain until

$$\text{nse}\left(\bar{h}_{MN}\right)$$

is sufficiently small.

For a given run length N, how can we estimate the nse, taking account of auto-correlations in

$$h\left(X^{(M+1)}\right), \ldots, h\left(X^{(N)}\right)$$

In the method of *batching*, the problem of auto-correlation is overcome by

- dividing the sequence

$$x^{(M+1)}, \ldots, x^{(N)}$$

  into k equal-length batches,
- calculating the mean b_j for each *batch* j,
- checking that the

$$b_1, \ldots, b_k$$

  are approximately uncorrelated.

Then we can estimate

$$\widehat{\text{nse}}\,(\bar{x}_{MN}) = \sqrt{\frac{1}{k(k-1)} \sum (b_i - \bar{b})^2}.$$

Notes:

- Use at least 20 batches.
- Estimate lag-1 autocorrelation of the sequence $\{b_i\}$.
- If the auto-correlation is high, a longer run should be used, giving larger batches.

Again return to Example 2.

Let $S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$. It is easy to find analytically:

$$E(\mu|y) = \bar{y} \text{ and } E(\sigma^2|y) = \frac{S_y^2}{n-3}.$$

Take $N = 2000, M = N/4$.

	T.mean	G.mean	nse	lag-1.c
μ	5.0675	5.0624	0.0046	0.104
σ^2	0.6306	0.6367	0.0062	0.097

When we come back after the break...

- Study Convergence
- Learn Graphical Models
- See **BUGS** illustrations.
- Do Bayesian Model Choice
- Perform Reversible Jump
- Adapt MCMC Methods