MSc HT15. Further Statistical Methods: MCMC

Lecture 5-6: Markov chains; Metropolis Hastings MCMC

Notes and Practicals available at

www.stats.ox.ac.uk/~nicholls\MScMCMC15
Markov chain Monte Carlo Methods

Our aim is to estimate $\mathbb{E}_p(\phi(X))$ for $p(x)$ some pmf (or pdf) defined for $x \in \Omega$.

Up to this point we have based our estimates on iid draws from either $p$ itself, or some proposal distribution with pmf $q$.

In MCMC we simulate a correlated sequence $X_0, X_1, X_2, ...$, which satisfies $X_t \sim p$ (or at least $X_t$ converges to $p$ in distribution) and rely on the usual estimate $\hat{\phi}_n = n^{-1} \sum_{t=0}^{n-1} \phi(X_t)$.

We will suppose $\Omega$, the space of states of $X$, is finite (and therefore discrete).

MCMC methods are applicable to countably infinite and continuous state spaces, and are one of the most versatile classes of Monte Carlo algorithms we have.
Markov chains

Let $\{X_t\}_{t=0}^{\infty}$ be a homogeneous Markov chain of random variables on $\Omega$ with starting distribution $X_0 \sim p^{(0)}$ and transition probability

$$P_{i,j} = \mathbb{P}(X_{t+1} = j | X_t = i).$$

Denote by $P_{i,j}^{(n)}$ the $n$-step transition probabilities

$$P_{i,j}^{(n)} = \mathbb{P}(X_{t+n} = j | X_t = i)$$

and by $p^{(n)}(i) = \mathbb{P}(X_n = i)$.

The transition matrix $P$ is *irreducible* if and only if, for each pair of states $i, j \in \Omega$ there is $n$ such that $P_{i,j}^{(n)} > 0$. The Markov chain is *aperiodic* if $P_{i,j}^{(n)}$ is non zero for all sufficiently large $n$. 
Markov chains

Here is an example of a periodic chain: \( \Omega = \{1, 2, 3, 4\} \), \( p^{(0)} = (1, 0, 0, 0) \), and transition matrix

\[
P = \begin{pmatrix}
0 & 1/2 & 0 & 1/2 \\
1/2 & 0 & 1/2 & 0 \\
0 & 1/2 & 0 & 1/2 \\
1/2 & 0 & 1/2 & 0 \\
\end{pmatrix},
\]

since \( P_{1,1}^{(n)} = 0 \) for \( n \) odd.

**Exercise**: show that if \( P \) is irreducible and \( P_{i,i} > 0 \) for some \( i \in \Omega \) then \( P \) is aperiodic.
The Stationary Distribution and Reversible Markov chains

The pmf $\pi(i), i \in \Omega$, $\sum_{i \in \Omega} \pi(i) = 1$ is a stationary distribution of $P$ if $\pi P = \pi$. If $p^{(0)} = \pi$ then

$$p^{(1)}(j) = \sum_{i \in \Omega} p^{(0)}(i) P_{i,j},$$

so $p^{(1)}(j) = \pi(j)$ also. Iterating, $p^{(t)} = \pi$ for each $t = 1, 2, ...$ in the chain, so the distribution of $X_t \sim p^{(t)}$ doesn’t change with $t$, it is stationary.

In a reversible Markov chain we cannot distinguish the direction of simulation from inspection of a realization of the chain and its reversal, even with knowledge of the transition matrix.

Most MCMC algorithms are based on reversible Markov chains.
Denote by $P'_{i,j} = \mathbb{P}(X_{t-1} = j | X_t = i)$ the transition matrix for the time-reversed chain.

It seems clear that a Markov chain will be reversible if and only if $P = P'$, so that any particular transition occurs with equal probability in forward and reverse directions.

**Theorem.**

(I) If there is a probability mass function $\pi(i), i \in \Omega$ satisfying $\pi(i) \geq 0$, $\sum_{i \in \Omega} \pi(i) = 1$ and

"Detailed balance": $\pi(i)P_{i,j} = \pi(j)P_{j,i}$ for all pairs $i, j \in \Omega$,

then $\pi = \pi P$ so $\pi$ is stationary for $P$.

(II) If in addition $p^{(0)} = \pi$ then $P' = P$ and the chain is reversible with respect to $\pi$. 
Proof of (I): sum both sides of detailed balance equation over \( i \in \Omega \). Now \( \sum_i P_{j,i} = 1 \) so \( \sum_i \pi(i)P_{i,j} = \pi(j) \).

Proof of (II), we have \( \pi \) a stationary distribution of \( P \) so \( \mathbb{P}(X_t = i) = \pi(i) \) for all \( t = 1, 2, \ldots \) along the chain. Then

\[
P_{i,j}' = \mathbb{P}(X_{t-1} = j|X_t = i) = \mathbb{P}(X_t = i|X_{t-1} = j)\frac{\mathbb{P}(X_{t-1} = j)}{\mathbb{P}(X_t = i)} \text{ (Bayes rule)}
\]

\[
= P_{j,i}\pi(j)/\pi(i) \text{ (stationarity)}
\]

\[
= P_{i,j} \text{ (detailed balance)}.
\]
Convergence and the Ergodic Theorem

If the (finite state space) MC is irreducible and aperiodic then the stationary distribution is unique and $p(t) \to \pi$ as $t \to \infty$. We say the chain “targets” $\pi$. If we simulate the MC $X_0, X_1, \ldots X_n$ to large enough $n$ from any start $X_0 = x_0$ then since $X_t \sim p^t$ and $p^t \sim \pi$ at large $t$, 'most' of the samples are 'nearly' distributed according to $\pi$.

We will use $\{X_t\}_{t=0}^{n-1}$ to estimate $\mathbb{E}_\pi(\phi(X))$. The ‘obvious’ estimator is

$$\hat{\phi}_n = \frac{1}{n} \sum_{t=0}^{n-1} \phi(X_t),$$

but the $X_t$ are correlated and only converge in distribution to $\pi$. 
Theorem. If $\{X_t\}_{t=0}^{\infty}$ is an irreducible and aperiodic Markov chain on a finite space of states $\Omega$, and is reversible wrt $\pi$, then as $n \to \infty$

$$P(X_n = i) \to \pi(i) \text{ and } \hat{\phi}_n \xrightarrow{\text{P}} \mathbb{E}_\pi(\phi(X))$$

for any bounded function $\phi : \Omega \to \mathbb{R}$.

We refer to such a chain as ergodic with equilibrium $\pi$.

$\hat{\phi}_n$ is consistent. A more general statement asks for a positive recurrent chain. The conditions are simpler here because we are assuming a finite state space for the Markov chain.

We would really like to have a CLT for $\hat{\phi}_n$ formed from the Markov chain output, so we have confidence intervals $\pm \sqrt{\text{var}(\hat{\phi}_n)}$ as well as the central point estimate $\hat{\phi}_n$ itself. These results hold for all the examples discussed later but are a little beyond us at this point.
Metropolis-Hastings Algorithm

Suppose we need samples from a pmf $p(x), x \in \Omega$. We give an algorithm simulating a Markov chain targeting $p$. It is enough to give a rule simulating $X_{t+1}$ give $X_t$. The algorithm determines the transition probabilities $P(X_{t+1} = y|X_t = x)$ and the transition matrix $P$.

Let $p(x) = \tilde{p}(x)/Z_p$ be the pmf on finite state space $\Omega = \{1, 2, \ldots, m\}$. We will call $p$ the target distribution.

Choose a ‘proposal’ transition matrix $q(y|x)$. We will use the notation $Y \sim q(\cdot|x)$ to mean $\Pr(Y = y|X = x) = q(y|x)$. 
Metropolis Hastings MCMC: the following algorithm simulates a Markov chain. If the chain is irreducible and aperiodic then it is ergodic with equilibrium distribution $p$.

Let $X_t = x$. $X_{t+1}$ is determined in the following way.

[1] Draw $y \sim q(\cdot|x)$ and $u \sim U[0, 1]$.

[2] If

$$u \leq \alpha(y|x) \text{ where } \alpha(y|x) = \min \left\{ 1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)} \right\}$$

set $X_{t+1} = y$, otherwise set $X_{t+1} = x$.

We initialise this with $X_0 = x_0, p(x_0) > 0$ and iterate for $t = 1, 2, 3, \ldots n$ to simulate the samples we need.
Example: Simulating a Discrete Distribution

Let \( p(x) = x/Z_p \) with \( Z_p = \sum_{x=1}^{m} x \).

Give a MH MCMC algorithm ergodic for \( p(x), x = 1, 2, \ldots, m \).

Step 1: Choose a proposal distribution \( q(y|x) \). It needs to be easy to simulate and determine a irreversible chain.

A simple distribution that 'will do' is \( Y \sim U\{1, 2, \ldots, m\} \), so

\[
q(y) = 1/m, \quad y = 1, 2, \ldots, m.
\]

This proposal scheme is clearly irreducible (we can get from \( A \) to \( B \) in a single hop).
Step 2: write down the algorithm.

If \( X_t = x \), then \( X_{t+1} \) is determined in the following way.

[1] Simulate \( y \sim U\{1, 2, \ldots, m\} \) and \( u \sim U[0, 1] \).

[2] If

\[
    u \leq \min \left\{ 1, \frac{\tilde{p}(y) q(x|y)}{\tilde{p}(x) q(y|x)} \right\}
\]

\[
    = \min \left\{ 1, \frac{y}{x} \right\}
\]

set \( X_{t+1} = y \), otherwise set \( X_{t+1} = x \).
#MCMC simulate $X_t$ according to $p=[1:m]/\text{sum}(1:m)$.

```
m<-30
n<-10000; X<-rep(NA,n); X[1]<-1
for (t in 1:(n-1)) {
    x<-X[t]
    y<-ceiling(m*runif(1))
    a<-min(1,y/x)
    U<-runif(1)
    if (U<=a) {
        X[t+1]<-y
    } else {
        X[t+1]<-x
    }
}
```
Left: The $x$-axis is step counter $t = 1, 2, 3...200$. The $y$-axis is Markov chain state $X_t$ for $	ilde{p}(x) = x$, $x = 1, ..., m$, $m = 30$.

Right: histogram of $X_1, X_2, ..., X_n$ for $n = 1000$. 
Example: Simulating a Poisson Distribution

We want to simulate

\[ p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \propto \frac{\lambda^x}{x!} \]

Step 1: Define the proposal. We want something simple to simulate with a simple density. We use

\[ q(y|x) = \begin{cases} 
\frac{1}{2} & \text{for } y = x \pm 1 \\
0 & \text{otherwise,}
\end{cases} \]

i.e. toss a coin and add or subtract 1 to \( x \) to obtain \( y \).
Step 2: Write down the algorithm, giving the acceptance probability \( \alpha(y|x) \) in closed form.

If \( X_t = x \), then \( X_{t+1} \) is determined in the following way.

[1] Simulate \( y \sim U\{x - 1, x + 1\} \) and \( u \sim U[0, 1] \).

[2] The acceptance probability is
\[
\alpha(y|x) = \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\}.
\]

If \( y = -1 \) then \( \alpha = 0 \). Otherwise,
\[
\alpha(y|x) = \begin{cases} 
\min \left(1, \frac{\lambda}{x+1}\right) & \text{if } y = x + 1 \\
\min \left(1, \frac{x}{X}\right) & \text{if } y = x - 1
\end{cases}
\]

If \( u \leq \alpha(y|x) \), set \( X_{t+1} = y \), otherwise set \( X_{t+1} = x \).
MCMC for state spaces which are not finite. Does this work for continuous rv? Computers use the “computer measure”. The reals are discretised.

Let $x^*$ be the computer truncation of $x$ and $\delta x = \{y : x^* = x\}$. The length $|\delta x|$ of the cell containing $x$ is not constant. Roughly $|\delta x|/x \simeq 10^{-15}$. $\pi(x)$ is approximated by $[\pi(x)]^*$. 

The Hastings ratio we compute is

$$
\frac{[\tilde{p}(y)]^*[q(x|y)]^*}{[\tilde{p}(x)]^*[q(y|x)]^*} = \frac{[p(y)]^*|\delta y|[q(x|y)]^*|\delta x|}{[p(x)]^*|\delta x|[q(y|x)]^*|\delta y|} \simeq \frac{Pr(Y \in \delta y) Pr(X \in \delta x|Y = y)}{Pr(X \in \delta x) Pr(Y \in \delta y|X = x)}
$$

since $Pr(X \in \delta x) \simeq [p(x)]^*|\delta x|$ etc. If we apply this to densities, we simulate the approximate distribution. Our discussion of Markov chains on finite spaces is relevant.
MCMC for the Normal distribution

Suppose want to simulate the standard normal distribution $X \sim N(0, 1)$. The target density is

$$\tilde{p}(x) \propto \exp(-x^2/2).$$

Step 1: Choose the proposal distribution. We need something simple and irreducible. Fix a constant $a > 0$ and choose a new point uniformly at random in a window of length $2a$ centred at $x$. The proposal density is

$$q(y|x) = \frac{1}{2a} \mathbb{1}_{x-a < y < x+a}$$

Notice that $q(y|x) = q(x|y)$. 
Step 2: give the MCMC algorithm. If $X_t = x$ then $X_{t+1}$ is determined in the following way:

[1] Simulate $Y \sim U(x - a, x + a)$ and $U \sim U(0, 1)$.

[2] If $U \leq \alpha(y|x)$ set $X_{t+1} = y$ and otherwise set $X_{t+1} = x$.

Here

$$
\alpha(y|x) = \min \left( 1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)} \right)
= \min \left( 1, \exp(-y^2/2 + x^2/2) \right).
$$
#MCMC simulate $X_t \sim N(0,1)$

a=3; n=2000

X=numeric(n); X[1]=0;

for (t in 1:(n-1)) {
    x<-X[t]
    y<-x+(2*runif(1)-1)*a
    if (runif(1)<exp((x^2-y^2)/2)) {
        X[t+1]<-y
    } else {
        X[t+1]<-x
    }
}
(see the associated R-file for plotting commands)
Claim: The following algorithm simulates a Markov chain. If the chain is irreducible and aperiodic then it is ergodic with equilibrium distribution $p$.

Let $X_t = x$. $X_{t+1}$ is determined in the following way.

[1] Draw $y \sim q(\cdot|x)$ and $u \sim U[0, 1]$.

[2] If

$$u \leq \alpha(y|x) \quad \text{where} \quad \alpha(y|x) = \min \left\{ 1, \frac{\hat{p}(y)q(x|y)}{\hat{p}(x)q(y|x)} \right\}$$

set $X_{t+1} = y$, otherwise set $X_{t+1} = x$.

We initialise this with $X_0 = x_0, p(x_0) > 0$ and iterate for $t = 1, 2, 3, \ldots, n$ to simulate the samples $X_0, X_1, X_2, \ldots, X_n$ we need.
Proof: By the ergodic theorem it is sufficient to show that the Markov chain determined by the random MCMC update has $p$ as a stationary distribution.

We will compute the transition matrix $P$ and show it satisfies detailed balance,

$$P_{x,y} p(x) = P_{y,x} p(y),$$

since that implies $p = pP$.

We don’t need to calculate $P_{x,y}$ when $x = y$ as DB is clear. Suppose $y \neq x$. If $X_t = x$, then the probability $P_{x,y}$ to move to $X_{t+1} = y$ at the next step is the probability to propose $y$ at step 1 times the probability to accept it at step 2, so

$$P_{x,y} = P(X_{t+1} = y|X_t = x) = q(y|x)\alpha(y|x).$$
Now check DB:

\[ p(x)P_{x,y} = p(x)q(y|x)\alpha(y|x) \]

\[ = p(x)q(y|x) \min \left\{ 1, \frac{p(y)q(x|y)}{p(x)q(y|x)} \right\} \]

\[ = \min \{ p(x)q(y|x), p(y)q(x|y) \} \]

\[ = p(y)q(x|y) \min \left\{ \frac{p(x)q(y|x)}{p(y)q(x|y)}, 1 \right\} \]

\[ = p(y)q(x|y)\alpha(x|y) \]

\[ = p(y)P_{y,x} \]

and we are done.