SC7/SM6 Bayes Methods HT18

Lecturer: Geoff Nicholls

Lecture 2: Monte Carlo Methods

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/

and via the MSc weblearn pages.
Why Monte Carlo?

Many of the important quantities in Bayesian inference are expectations over the posterior distribution: posterior probabilities, marginal likelihoods, many point estimates and interval estimates - all given in terms of typically intractable integrals.

Suppose $\Theta \in \Omega$ is a parameter vector with posterior density $\Theta \sim \pi(\theta|y)$ given data $Y = y$, suppose $f : \Omega \rightarrow \mathbb{R}$ is some function and we want to evaluate the posterior expectation

$$E_{\Theta|y}(f(\Theta)|Y = y) = \int_{\Omega} f(\theta)\pi(\theta|y)d\theta.$$ 

If we have $X_t \in \Omega, t = 1, 2, \ldots, T$ with $X_t \sim \pi(\cdot|y)$ then

$$\hat{f}_T = T^{-1} \sum_{i=1}^{T} f(X_t)$$

is an unbiased estimator for $E_{\Theta|y}(f(\Theta)|Y = y)$. 
Markov chain Monte Carlo Methods

MCMC is a family of algorithms for simulating $X_0, X_1, X_2, \ldots$ so that $X_t \sim p$ (or at least $X_t$ converges to $p$ in distribution) for a user-defined probability distribution $p$.

MCMC methods are one of the most versatile classes of Monte Carlo algorithms we have, and are in routine use across statistics.

I will set out theory for the case that $\Omega$, the space of states of $X$, is finite (and therefore discrete) because it is simpler. However, it also captures many of the essential issues. When we work on a computer we approximate any continuous quantities like $\theta$, $L(\theta; y)$, $\pi(\theta)$ and $\pi(\theta|y)$ using finite precision arithmetic so we are really working with finite $\Omega$ anyway.
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

\[
P_{i,j} = \mathbb{P}(X_{t+1} = j \mid X_t = i).
\]

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

\[
P_{i,j}^{(n)} = \mathbb{P}(X_{t+n} = j \mid X_t = i)
\]

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

The transition matrix \( P \) is \textit{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 \textit{aperiodic} if \( P_{i,j}^{(n)} \) is non zero for all sufficiently large \( n \).
The Stationary Distribution and Detailed Balance

In discussing Markov chains we will work with a generic “target” distribution \( p(i), i \in \Omega \). This is the distribution we will try to sample. When we come to apply the MCMC methods to Bayesian inference, the target distribution will be the posterior \( p(\theta) = \pi(\theta | y) \).

The probability mass function (PMF) \( p(i), i \in \Omega, \sum_{i \in \Omega} p(i) = 1 \) is a stationary distribution of \( P \) if \( pP = p \). If \( p^{(0)} = p \) then

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

so \( p^{(1)}(j) = p(j) \) also. Iterating, \( p^{(t)} = p \) for each \( t = 1, 2, \ldots \) in the chain, so the distribution of \( X_t \sim p^{(t)} \) doesn’t change with \( t \), it is stationary.
We want $X_t \xrightarrow{D} p$. The convergence theorem for finite irreducible Markov chains tells us that if $\Omega$ is finite, $pP = p$, and $P$ is irreducible and aperiodic, then indeed $X_t \xrightarrow{D} p$. To show this works for given $P$ we need to check $pP = p$, but that is hard, as we have to sum over all $\Omega$ to evaluate $pP$. However...

**Detailed Balance.**
If there is a probability mass function $p(i), i \in \Omega$ satisfying $p(i) \geq 0$, $\sum_{i \in \Omega} p(i) = 1$ and

“Detailed balance”: $p(i)P_{i,j} = p(j)P_{j,i}$ for all pairs $i, j \in \Omega$, then $p = pP$ so $p$ is stationary for $P$.

Detailed balance is sufficient for stationarity, and much easier to check. A Markov chain satisfying DB is “reversible”. 
Exercise: prove that if \( p(i)P_{i,j} = p(j)P_{j,i} \) for all \( i, j \in \Omega \) then \( p = pP \). Give the corresponding expressions for the continuous case where \( \Omega = \mathbb{R}^n \) say, expressing stationarity as an integral equation satisfied by \( p \), and write DB in terms of probability densities.
Convergence and the Ergodic Theorem

We choose some “start state” \( X_0 \sim p^{(0)} \) to initialise the Markov chain. If the chain converges to the target distribution \( p \), then \( X_t \sim p^{(t)} \) with \( p^{(t)} \approx p \) at large \( t \), so when we look at our Markov chain samples \( X_0, X_1, ..., X_T \), “most” of the samples are “nearly” distributed according to \( p \). Is this good enough?

Let \( \hat{f}_T \) be the estimator for \( E(f(X)), X \sim p \) we defined above.

**Theorem.** If \( \{X_t\}_{t=0}^{\infty} \) is an irreducible and aperiodic Markov chain on a finite space of states \( \Omega \) satisfying detailed balance with respect to the probability distribution \( p \), then as \( T \to \infty \)

\[
P(X_T = i) \to p(i) \quad \text{and} \quad \hat{f}_T \xrightarrow{a.s.} E(f(X))
\]

for any bounded function \( f : \Omega \to R \). [For proof see eg Norris *Markov Chains*, CUP, (1997)]
We refer to such a chain as ergodic with target $\pi$.

A more general statement asks for a positive or Harris 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 $\hat{f}_n$ formed from the Markov chain output, so we have confidence intervals $\pm \sqrt{\text{var}(\hat{f}_n)}$ as well as the central point estimate $\hat{f}_n$ itself. CLT’s hold for all the examples in this course. [See eg Part C Advanced Simulation]
Metropolis-Hastings Algorithm

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

The basic idea here is to simulate a random walk $X_0, X_1, X_2, ...$ in $\Omega$ by accepting or rejecting proposals from a simple irreducible transition matrix $Q_{i,j}, i, j \in \Omega$ which we get to choose. If $p$ was stationary for $Q$ we could just simulate the chain with transition matrix $Q$, as the chain would target $p$. However we don’t know how to choose such $Q$. The trick is to “correct” proposals drawn from $Q$ to get a new effective transition matrix $P$ which satisfies DB for $p$. 
Metropolis Hastings MCMC: the algorithm simulates a Markov chain. Let $q(j|i) = Q_{i,j}$ be a proposal distribution satisfying $q(j|i) > 0 \iff q(i|j) > 0$. If the chain is irreducible and aperiodic then it is ergodic with target distribution $p$.

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

[1] Draw $j \sim q(\cdot|i)$ and $u \sim U[0,1]$.
[2] If $u \leq \alpha(j|i)$ where $\alpha(j|i) = \min \left\{ 1, \frac{p(j)q(i|j)}{p(i)q(j|i)} \right\}$ then set $X_{t+1} = j$, otherwise set $X_{t+1} = i$.

We initialise this with some $X_0 = i_0, p(i_0) > 0$ and iterate for $t = 1, 2, 3, \ldots T$ to simulate the samples we need.
Example: Simulating the hypergeometric distribution

The hypergeometric distribution \( \text{HyperGeom}(k; K, N, n) \) with parameters \( K = 10, N = 20, n = 10 \) gives the probability for \( k \) successes in \( n \) draws from a population of size \( N \) containing \( K \) successes. If \( p(k) = \text{HyperGeom}(k; K, N, n) \) then

\[
p(k) = \binom{K}{k} \binom{N - K}{n - k} / \binom{N}{n}
\]

Give a MH MCMC algorithm ergodic for \( p(k) \),

\[
\max\{0, n + K - N\} \leq k \leq \min\{n, K\}
\]

(i.e \( k = 0, 1, 2, \ldots, 10 \) here).
Step 1: Choose a proposal distribution $q(j|i)$. It needs to be easy to simulate and determine an irreducible chain. A simple distribution that 'will do' is

$$q(j|i) = \begin{cases} 
1/2 & \text{for } j = i \pm 1 \\
0 & \text{otherwise,}
\end{cases}$$

i.e. toss a coin and add or subtract 1 to $i$ to obtain $j$. This is irreducible (we can get from any state $A$ to any other state $B$ by adding or subtracting 1’s).

Notice that we can “walk out” of the state space $B^-,\ldots,B^+$ given above. If for eg $i = B^+$ and we propose $j = i + 1$ then we have proposed a state $j$ with zero probability in the target distribution. One transparent way to deal with this is give these states probability zero in the target, setting $p(j) = 0$ for all $j \notin \Omega$. 

Step 2: write down the algorithm.

If $X_t = i$, then $X_{t+1}$ is determined in the following way.

[1] Simulate $j \sim U \{i - 1, i + 1\}$ and $u \sim U[0, 1]$.
[2] If $B^- \leq j \leq B^+$ and

$$
u \leq \min \left\{ 1, \frac{p(j)q(i|j)}{p(i)q(j|i)} \right\}$$

$$= \min \left\{ 1, \frac{(Kj)(N-K)}{(j)(N-j)} \right\}$$

then set $X_{t+1} = j$, otherwise (ie if either condition fails) set $X_{t+1} = i$.

The point here is that if $j < B^-$ or $j > B^+$ then $p(j) = 0$ so $\alpha = 0$ and we are bound to reject.
#MCMC simulate $X_t \sim \text{HyperGeom}(K=10, N=20, n=10)$.

```
K<-10; N<-20; n<-10; Bm<-min(0,n+K-N); Bp<-min(n,K);
T<-1000; X<-rep(NA,T);
X[1]<-Bm #start state at lower bound, here X[1]=0
for (t in 1:(T-1)) {
    i<-X[t]
    j<-sample(c(i-1,i+1),1)
    if (j<Bm | j>Bp) {
        X[t+1]<-i #must reject if outside SP
    } else {
        a<-min(1,(choose(K,j)*choose(N-K,n-j)) / (choose(K,i)*choose(N-K,n-i)))
        U<-runif(1)
        if (U<=a) {X[t+1]<-j} else {X[t+1]<-i}
    }
}
```
Left: $x$-axis is step counter $t = 1, 2, 3 \ldots 200$. The $y$-axis is Markov chain state $X_t$ for

$$ p(k) = \text{HyperGeom}(k; K = 10, N = 20, n = 10), \; k = 0, \ldots, 10. $$

Right: histogram of $X_1, X_2, \ldots, X_n$ for $T = 1000$. 

![MCMC step diagram](image)
Ergodicity proof for Metropolis Hastings

We assume the chain is irreducible and aperiodic - this must be checked separately for each MH MCMC algorithm. Under this assumption, it is sufficient (for ergodicity) 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,

$$P_{i,j} \ p(i) = P_{j,i} \ p(j),$$

since that implies $p = pP$.

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

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

\[ p(i)P_{i,j} = p(i)q(j|i)\alpha(j|i) \]

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

\[ = \min \{ p(i)q(j|i), p(j)q(i|j) \} \]

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

\[ = p(j)q(i|j)\alpha(i|j) \]

\[ = p(j)P_{j,i} \]

and we are done.