SC7/SM6 Bayes Methods HT20

Lecturer: Geoff Nicholls

Lecture 4: Markov Chain Monte Carlo Methods II

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/
MCMC for state spaces which are not finite*

Does this work for continuous rv? Computers use the “computer measure”. The reals are discretised.

Let $\tilde{\pi}(x) = c\tilde{\pi}(x), x \in R$. Let $x^* \in R^*$ be the computer truncation of $x$ to the interval $\delta x = \{y \in R : y^* = x^*\}$ with $R^*$ the set of values which can be represented on the computer.

The length $|\delta x|$ depends on $x$. Roughly $|\delta x|/x \simeq 10^{-15}$.

Let $[\tilde{\pi}(x)]^*$ denote the computer approximation to $\tilde{\pi}(x)$ and let $\pi^*(x) = b[\tilde{\pi}(x)]^*$ be the normalised computer PMF on $x \in R^*$.

*Material in this slide and next not examinable.
Our MCMC targets $\pi^*(x)$. If $\Pr_{\pi}(X \in \delta x) \simeq \pi^*(x) | \delta x|$ then

$$\frac{[\tilde{\pi}(y)]^* [q(x | y)]^*}{[\tilde{\pi}(x)]^* [q(y | x)]^*} \simeq \frac{\pi^*(y) | \delta y | [q(x | y)]^* | \delta x|}{\pi^*(x) | \delta x | [q(y | x)]^* | \delta y|} \frac{\Pr_{\pi}(Y' \in \delta y) \Pr_q(X' \in \delta x | Y = y)}{\Pr_{\pi}(X' \in \delta x) \Pr_q(Y' \in \delta y | X = x)}$$

so the variation in the length $|\delta x|$ is not itself a problem.

If we apply MCMC to densities, we simulate the approximate distribution. Our discussion of Markov chains on finite spaces is at least relevant as we only ever target $\pi^*(x), x \in \mathbb{R}^*$. 

†A coupling argument can be used to show we get the same output as an exact chain that uses the same random bits to drive the process.
MH example: an equal mixture of bivariate normals

\[ \pi(\theta) = (2\pi)^{-1} \left( 0.5e^{-(\theta - \mu_1)\Sigma_1^{-1}(\theta - \mu_1)/2} + 0.5e^{-(\theta - \mu_2)\Sigma_2^{-1}(\theta - \mu_2)/2} \right) \]

with \( \theta = (\theta_1, \theta_2) \). Use \( \mu_1 = (1, 1)^T \), \( \mu_2 = (4, 4)^T \) and \( \Sigma_1 = \Sigma_2 = I_2 \) for this illustration.

Step 1. For a proposal distribution \( q \) we want something simple to sample. A simple choice that "will do" is

\[ \theta'_i \sim U(\theta_i - a, \theta_i + a), \]

with \( a > 0 \) a fixed constant. We jump uniformly in a box of side \( 2a \), easy to sample, and

\[ q(\theta'|\theta) = q(\theta|\theta') = 1/4a^2. \]
Step 2. The algorithm is, given $\theta^{(n)} = (\theta_1, \theta_2)$,

[1] for $i = 1, 2$ simulate $\theta_i' \sim U(\theta_i - a, \theta_i + a)$

[2] with probability

$$\alpha(\theta' | \theta) = \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \right\}$$

set $\theta^{(n+1)} = \theta'$ otherwise set $\theta^{(n+1)} = \theta$.

Step 3. This algorithm is ergodic for any $a > 0$ but we will see that the choice of $a$ makes a difference to efficiency.
#MCMC simulate $X_t$ according to a mixture of normals

```r
f<-function(x,mu1,mu2,S1i,S2i,p1=0.5) {
    #mixture of normals, density up to constant factor
    c1<-exp(-t(x-mu1)%*%S1i%*%(x-mu1))
    c2<-exp(-t(x-mu2)%*%S2i%*%(x-mu2))
    return(p1*c1+(1-p1)*c2)
}
```

```r
a=3; n=2000
mu1=c(1,1); mu2=c(4,4); S=diag(2); S1i=S2i=solve(S);
X=matrix(NA,2,n); X[,1]=x=mu1
for (t in 1:(n-1)) {
    y<-x+(2*runif(2)-1)*a
    MHR<-f(y,mu1,mu2,S1i,S2i)/f(x,mu1,mu2,S1i,S2i)
    if (runif(1)<MHR) x<-y
    X[,t+1]<-x
}
```

Exercise: try this, experimenting with different values of $a$. 

(see the associated R-file for plotting commands)
Convergence and mixing

We want to estimate $E_p(f(X))$ using our MCMC samples $X_0, X_1, X_2, \ldots, X_n$ targeting $p(x)$ and calculate the estimate $\overline{f}_n = n^{-1}\sum_t f(X_t)$. The ergodic theorem tells us this estimate converges in probability to $E_p(f(X))$.

How large should we take $n$? There are two issues, bias and variance, respectively “convergence” and “mixing”.

First, we dont start the chain in equilibrium. Samples from the first part of the chain are biased by initialization. We drop the first part of the MCMC run (called “burn-in”) to reduce the
initialization bias. We know $p(t) \to p$ as $t \to \infty$ so choose a cut-off $T$ such that $p(t) \simeq p$ for $t \geq T$ is a good approximation. Take $n \gg T$ so that retained samples are representative of $p$.

If $n \gg T$ then the bias in $\bar{f}_n$ due to burn-in will be slight. If you need to drop a lot of states from the start of the chain to reduce this bias, you may not have run the chain long enough anyway.

Second, suppose $p^{(0)}(x) = p(x)$, so we start the chain in equilibrium. The variance, $\text{var}(\bar{f}_n)$, of $\bar{f}_n$ will get smaller as $n$ increases. We should choose $n$ large enough to ensure $\text{var}(\bar{f}_n)$ is small enough so that $\bar{f}_n$ has useful precision.

The following figures show output from two MCMC runs for the normal mixture with jump size $a = 2, 4$. We see
MCMC variance in equilibrium*

\(X_0, X_1, X_2, \ldots\) are correlated so generally \(\text{var}(\bar{f}_n) \neq \frac{\text{var}(f(X))}{n}\).

Suppose the chain has reached equilibrium. Correlation at lag \(s\)

\[\rho_s(f) = \frac{\text{cov}(f(X_i), f(X_{i+s}))}{\text{var}(f(X_i))}\]

(so \(\rho_0 = 1\)). Let \(\sigma^2 = \text{var}(f(X_i)), X_i \sim p\). This doesnt depend on \(i\) because the chain is stationary.

Express \(\text{var}(\bar{f}_n)\) in terms of \(\rho_s(f)\). This gives insight and leads to an estimator for \(\text{var}(\bar{f}_n)\), since we can estimate \(\rho_s(f)\).

\[
\text{var}(f) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}(f(X_i), f(X_j)) \\
= \sigma^2 n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho|i-j| \\
= \sigma^2 n^{-1} \left[ 1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \right] \\
\simeq \sigma^2 n^{-1} \left[ 1 + 2 \sum_{s=1}^{n-1} \rho_s \right] \\
= \sigma^2 \tau_f / n,
\]

if \( \rho_s \to 0 \) sufficiently rapidly at large \( s \). Here \( \tau_f \) is the Integrated Autocorrelation Time (IACT) and \( \text{ESS} = n/\tau_f \) is the Effective Sample Size - the number of independent samples giving the same precision for \( \overline{f} \) as the \( n \) correlated samples we have.
We can estimate $\gamma_s = \text{cov}(f(X_i), f(X_{i+s}))$ using

$$\hat{\gamma}_s = \frac{1}{n} \sum_{i=1}^{n-s} (f(X_i) - \hat{f})(f(X_{i+s}) - \hat{f}),$$

and $\gamma_0 = \text{var}(f(X_i))$ (as usual) from the sample output, and compute $\hat{\rho}_s = \hat{\gamma}_s / \hat{\gamma}_0$.

We get an estimate of $\tau_f$,

$$\hat{\tau}_f = 1 + 2 \sum_{s=1}^{M} \hat{\rho}_s,$$

with $M$ a cut-off on the sum. $\hat{\rho}_s$ goes to zero with $s$ and is dominated by noise at large $s$; terms at large $s$ just add noise. We truncate the sum over $s$, at $s = M$. We get $\hat{\tau}_f$ consistent if $M$ is chosen according to suitable criteria.

MCMC convergence

There is no simple generic sufficient condition we can test for convergence. Here some checks we should run to detect poor mixing and identify a burn-in and run length.

[1] Make multiple runs from different start states and check marginal distributions agree.

[2] Plot the autocorrelation function. Check that it falls off to vary around zero. Calculate the ESS and check it is reasonably large.

[3] Plot MCMC traces of the variables and key functions. The chain should be stationary after burn-in.
Here is an example of some of the plots I would use for convergence checking on the normal mixture MCMC sampler. In this example the convergence is quite poor: the three histograms differ by more than the precision I would typically be aiming for.

See associated R-file for further examples.
Mixing updates for multivariate targets

If $\theta = (\theta_1, \ldots, \theta_p)$ and we target $\pi(\theta)$, $\theta \in \Omega$, we commonly update one variable at a time. Small changes to the state will have a higher acceptance probability.

If we have $p$ updates setting $\theta'_{-i} = \theta_{-i}$ and $\theta'_i \sim q_i(\cdot|\theta)$, $i = 1, \ldots, p$ and select $q_i$ with probability $\xi_i$ in MH-MCMC then the overall transition density for the update is

$$K(\theta, \theta') = \sum_i \xi_i K_i(\theta, \theta'),$$

where $K_i(\theta, \theta') = q_i(\theta'_i|\theta)\alpha(\theta'|\theta)$, the MH-MCMC update kernel if we choose $q_i$, is reversible wrt $\pi(\theta)$. A mixture of reversible updates is reversible, so $K$ is reversible wrt $\pi(\theta)$. 
The Gibbs sampler

*Random scan Gibbs* takes $\xi_i = 1/p$ and $q_i(\theta_i|\theta) = \pi(\theta_i|\theta_{-i})$.

$$
\pi(\theta_i|\theta_{-i}) = \frac{\pi(\theta)}{\pi(\theta_{-i})}
$$

is the conditional density. MH-MCMC reduces to

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

1. Simulate $i \sim U\{1, \ldots, p\}$ and $\theta_i' \sim \pi(\cdot|\theta_{-i})$. Set $\theta_{-i}' = \theta_{-i}$.

2. Set $X_{t+1} = \theta'$.

The acceptance probability is one! We still check irreducibility.

*Exercise: write down $\alpha(\theta'|\theta)$ for this case and show it equals one.*
The *sequential-scan Gibbs sampler* visits each variable in turn from $1 \ldots p$. This is also stationary.

Example: Bivariate density $\pi(\theta_1, \theta_2)$ and $X_t = (\theta_1, \theta_2)$.

Algorithm:

1. Simulate $\theta'_1 \sim \pi(\theta'_1 | \theta_2)$ then $\theta'_2 \sim \pi(\theta'_2 | \theta'_1)$.

2. Set $X_{t+1} = (\theta'_1, \theta'_2)$.

Claim: if $X_t \sim \pi$ then after these two steps we have a new correlated sample $X_{t+1} \sim \pi$, so the process is stationary.
The distribution of \( X_{t+1} = (\theta'_1, \theta'_2) \) is

\[
p(\theta'_1, \theta'_2) = \int \pi(\theta_1, \theta_2) \pi(\theta'_1 | \theta_2) \pi(\theta'_2 | \theta'_1) d\theta_1 d\theta_2
\]

\[
= \int \pi(\theta_1, \theta_2) \frac{\pi(\theta'_1, \theta_2)}{\pi(\theta_2)} \frac{\pi(\theta'_1, \theta'_2)}{\pi(\theta'_1)} d\theta_1 d\theta_2
\]

\[
= \int \pi(\theta_1 | \theta_2) \pi(\theta_2 | \theta'_1) \pi(\theta'_1, \theta'_2) d\theta_1 d\theta_2
\]

\[
= \pi(\theta'_1, \theta'_2) \quad \text{as} \quad \pi(\theta_1 | \theta_2) \quad \text{and} \quad \pi(\theta_2 | \theta'_1) \quad \text{are normalised.}
\]

If \( \theta^{(0)} = (\theta_1, \theta_2) \), then \( \theta^{(1)} = (\theta'_1, \theta'_2) \) and we can iterate to simulate \( \theta^{(2)}, \theta^{(3)} \ldots \). The same proof works for \( p > 2 \) variables.
Data Augmentation

Some important early applications of the Gibbs sampler arise for missing data. This is also called “data augmentation”.

DA is convenient when the likelihood on the full data is much simpler than the likelihood on the observed data.

Suppose the observation process is

\[ z \sim p(z|\theta), \quad y \sim p(y|z, \theta) \]

and we observe \( y \). The posterior \( \pi(\theta|y) \) is awkward as the likelihood function is an integral,

\[ \pi(\theta|y) \propto \pi(\theta) \int p(y|z, \theta)p(z|\theta)dz \]

In data augmentation we work with the joint posterior density \( p(\theta, z|y) \), thinking of the missing data as another parameter. The posterior is

\[ p(\theta, z|y) \propto p(y|z, \theta)p(z|\theta)p(\theta) \]
A Gibbs sampler for Probit regression - DA simplifies sampling.

Probit regression is a GLM where we have, for the $i$th observation $y_i, i = 1, \ldots, n$, covariates $x_i = (x_{i,1}, \ldots, x_{i,p})$, parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, linear predictor $\eta_i = \sum_j \theta_j x_{i,j}$, observation model

$$y_i \sim Bernoulli(\Phi(\eta_i))$$

and inverse link function $E(Y_i) = \Phi(\eta_i)$ the cdf of $N(0, 1)$.

If the prior for $\theta$ is $\pi(\theta)$ then the posterior for $\theta|y$ is

$$\pi(\theta|y) \propto \pi(\theta) \prod_i \Phi(\eta_i)^{y_i}(1 - \Phi(\eta_i))^{1-y_i}$$

with $\eta_i = \eta_i(\theta), i = 1, \ldots, n$. We cant calculate conditionals $\pi(\theta_j|\theta_{-j}, y)$ as $\theta$ appears inside $\Phi$ so we cant Gibbs sample.
There is another way to represent this model. If
\[ z_i \sim N(\eta_i, 1), i = 1, \ldots, n \]
and we set \( y_i = 1 \) if \( z_i > 0 \) and \( y_i = 0 \) if \( z_i \leq 0 \) then*
\[
\Pr(y_i = 1|\theta) = \int_{z_i > 0} N(z_i; \eta_i, 1)dz_i = \Phi(\eta_i).
\]

The joint posterior augmented with \( z \) is
\[
\pi(\theta, z|y) \propto \pi(z|\theta)\pi(\theta) \prod_i \mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}.
\]
The marginal for \( \theta \) is just \( \pi(\theta|y) \), since
\[
\pi(\theta|y) \propto \pi(\theta) \prod_i \int \pi(z_i|\theta)\mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}dz_i \propto \pi(\theta)p(y|\theta)
\]
*\( z = \eta + W \) with \( W \sim N(0, 1) \) so \( \Pr(z > 0) = \Pr(W > -\eta) = \Phi(\eta) \).
Consider Bayesian inference with a normal prior \( \theta \sim N(0, \Sigma) \).

\[
\pi(\theta, z|y) \propto \pi(z|\theta)\pi(\theta) \prod_i \mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}.
\]

\( \pi(\theta|y, z) \propto \pi(z|\theta)\pi(\theta) \) is jointly normal\(^\dagger\) (conjugate prior) and

\[
\pi(z_i|\theta) \propto N(z_i; \eta_i, 1) \mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}.
\]

Here is the Gibbs sampler. Suppose \( X_t = (\theta, z) \) and \( \eta = \eta(\theta) \).

(1). For \( i = 1, \ldots, n \), simulate \( z'_i \sim N(\eta_i, 1) \mathbb{I}_{y_i=\mathbb{I}_{z'_i>0}} \).

(2). Simulate \( \theta' \sim \pi(\theta|z') \) (multivariate normal, no \( y \) given \( z' \)).

(3) Set \( X_{t+1} = (\theta', z') \).

\(^\dagger\)See PS1 and R code example.