DTC Statistics module 2010

9:30-10:30 Gibbs Sampler, theory and examples

10:30-12 First Gibbs sampler problem

12-12:30 Feedback

14:00-5:00 Second Gibbs Sampler problem

17:00-17:30 Feedback

Geoff Nicholls `nicholls@stats.ox.ac.uk` 27/01/2010

Gibbs sampling...

1. is a special case of Metropolis Hastings MCMC.
2. is used to sample complex realistic posterior distributions for Bayesian inference.
3. was the most commonly used form of MCMC until recently
4. is popular because it is simple and intuitively natural.

Why do MCMC?

Let $p(x)$ be some posterior pdf of interest. Suppose we want to estimate $\mathsf{E}(g(X))$ for some function $g(x)$.

Draw J samples $X^{(j)} \sim p(\cdot)$, $j = 1, 2, 3, \ldots J$. Let

$$g_J = \frac{1}{J} \sum_{j=1}^{J} g(X^{(j)}).$$

Use the ergodic property of the samples -

$$g_J \rightarrow \mathsf{E}(g(X))$$

in probability (or better). In fact a stronger property 'usually' holds

$$g_J \rightarrow \mathsf{N}(\mathsf{E}(X)), \sigma_g^2 \tau_g / J).$$

Here σ_g^2 is the variance of $g(X)$ if $X \sim p(\cdot)$, and $\tau_g \geq 1$ with $\tau_g = 1$ if the samples $X^{(t)}, j = 1, 2, ..., J$ are jointly independent.

Metropolis Hastings MCMC

How to simulate $X^{(j)} \sim p(\cdot), j = 1, 2, 3, ...J$ for a given posterior pdf?

Chose a starting value $X^{(0)} = x^{(0)}$ satisfying $p(x) > 0$. Now suppose $X^{(j)} = x$. The value of $X^{(j+1)}$ is given as follows:

1. simulate a candidate $x' \sim q(x'|x)$
2. calculate
$$\alpha(x'|x) = \min\left\{1,\ \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right\}$$
3. draw a uniform random number $U \sim U(0, 1)$. If $U < \alpha(x'|x)$ accept the candidate (set $X^{(j+1)} = x'$) and otherwise reject the candidate (set $X^{(j+1)} = x$).

Example: suppose $X \sim N(\mu, \Sigma)$, $\Sigma_{ii} = \text{var}(X_i) = \sigma^2$, for both $i = 1, 2$, and $\Sigma_{1,2} = \text{cov}(X_1, X_2) = \rho\sigma_1\sigma_2$. In this case

$$p(x_1, x_2) \propto \exp\left(-\frac{(x_1-\mu_1)^2 - 2\rho(x_1-\mu_1)(x_2-\mu_2) + (x_2-\mu_2)^2}{2\sigma^2(1-\rho^2)}\right)$$

For example, we will take $\sigma^2 = 3$ and $\rho = -2/3$.

Construct a MH MCMC algorithm to sample this distribution.

For the proposal step, fix a constant $a > 0$ and make random jumps of size up to a along the two axes:

$$q(x'|x) = \frac{1}{4a^2} \quad \text{for} \quad x_1 - a \leq x'_1 \leq x_1 + a, x_2 - a \leq x'_2 \leq x_2 + a$$

ie the uniform density in a box of side $2a$ centred at x with sides aligned to the axes.

Check $q(x'|x) = q(x|x')$: the probability density to propose y given x is equal to the probability density to propose x given y (if y is in reach of x then x is in reach of y so both densities equal $1/4a^2$).

Here is a Metropolis Hastings Markov Chain Monte Carlo algorithm simulating $X^{(j)} \sim N(\mu, \Sigma), j = 1, 2, ..., J$.

```
%%
%MH MCMC for bivariate normal example
clear;
mu=[5;5];
ssq=3;
rho=-2/3;

%%
J=2000;
a=5;
x=[0;0];
X=zeros(2,J);

%%
for j=1:J
  xp=x+a*(2*rand(2,1)-1);
  alpha=min(1,p(xp,mu,ssq,rho)/p(x,mu,ssq,rho));
  if (rand<alpha), x=xp; end

  X(:,j)=x;
end

mean(X')'
cov(X')
```

The Gibbs Sampler

Where the parameter is multidimensional, there is a particularly simple, and efficient, scheme for MCMC.

Suppose the parameter (x above) has k components: $x = (x_1, \ldots x_k)$.

The Gibbs sampler updates the components of the parameter sequentially (either in a fixed or a random order). When updating the ith component, treat all the other components as if they were known, and choose a new value for x_i from the conditional distribution of this component given the other components.

In this setup, at a step of the MH MCMC algorithm that updates component i, the new state is

$$x' = (x_1, x_2, ...x_{i-1}, x'_i, x_{i+1}, ..., x_k)$$

with

$$q(x'|x) = p(x'_i|x_{-i}),$$

and since

$$\begin{aligned} p(x'_i|x_{-i}) &= p(x')/p(x_{-i}) \\ &= p(x')/\sum_{x_i} p(x) \end{aligned}$$

the acceptance probability is

$$\begin{aligned} \alpha(x'|x) &= \min\left\{\frac{p(x')q(x|x')}{p(x)q(x'|x)}\right\} \\ &= \min\left\{1,\ \frac{p(x')p(x_i|x_{-i})}{p(x)p(x'_i|x_{-i})}\right\} \end{aligned}$$

$$
\begin{aligned}
&= \min\left\{1,\ \frac{p(x')p(x)/p(x_{-i})}{p(x)p(x')/p(x_{-i})}\right\} \\
&= 1.
\end{aligned}
$$

so we always accept. Since Gibbs sampling is MH MCMC and MH works (tbc!), Gibbs sampling works (tbc).

Example: Bivariate Normal

For our example $(X_1, X_2) \sim N(\mu, \Sigma)$, the condition distributions are

$$X_1|X_2 = x_2 \sim N(\mu_1 + \rho(x_2 - \mu_2), (1 - \rho^2)\sigma^2)$$

and

$$X_2|X_1 = x_1 \sim N(\mu_2 + \rho(x_1 - \mu_1), (1 - \rho^2)\sigma^2)$$

(you can check this by completing the square for x_1 in the original density).

```
%%
%Gibbs sampler for bivariate normal example
clear;
mu=[5;5];
...
%%
J=1000;
...
%%
for j=1:J
  x(1)=mu(1)+rho*(x(2)-mu(2))+randn*sqrt((1-rho^2)*ssq);
  x(2)=mu(2)+rho*(x(1)-mu(1))+randn*sqrt((1-rho^2)*ssq);
  X(:,j)=x;
end
...
```

Example: Mixture Models

Consider the following simple model for the sexes of pairs of twins.

Independently, each twin pair is identical with probability p, and non-identical with probability $1 - p$.

A pair of identical twins will be both boys with probability q, and both girls with probability $(1 - q)$.

The sexes of each of a pair of non-identical twins are independent, with an individual being a boy with probability q and a girl otherwise.

Suppose we have data on a random sample of n pairs of twins of which u are both boys, v are both girls, and w are of opposite sexes. How do we estimate p and q?

The key is to think of this as a missing data problem. If we knew for each twin pair whether they were identical or not, the problem would be easy.

So introduce missing data $I_1, I_2, \ldots I_n$ where $I_i = 1$ if the ith pair of twins is identical and 0 otherwise.

For a Bayesian approach, adopt a Beta(a, b) prior for p, and an independent Beta(c, d) prior for q.

The "parameter" of interest is $(p, q, I_1, I_2, \ldots, I_n)$.

We have effectively augmented the parameter space by including the missing data. For a Bayesian, there is no conceptual difference between missing data and parameters – they are all just quantities we happen not to know.

This problem (and lots with a similar structure) is tailor-made for using a Gibbs sampler. Update the parameters sequentially as follows, in each case treating all the other parameters as fixed.

1. When we come to update I_i, then if that particular twin pair have different sexes, we must have $I_i = 0$

If both individuals in the pair are male, by Bayes theorem,

$$P(I_i = 1) = pq/(pq + (1-p)q^2),$$

while if both are female,

$$P(I_i = 1) = p(1-q)/(p(1-q) + (1-p)(1-q)^2).$$

2. When we come to update p, since we assume we know the values of the I's it is as if we know how many of the twin pairs are identical. Call this number (given by $\sum I_i$) x.

Then by the earlier example, the posterior distribution for p will be Beta$(a + x, b + n - x)$, and we sample a new value of p from this distribution.

3. When we come to update q, since we assume the I's are known, it is as if we have $x + 2(n - x)$ independent Bernoulli trials (one trial for each identical pair of twins, and two for each non-identical pair).

Now (arbitrarily!) define "male" as a success. For an identical twin pair, count 1 success if they are both males. For a non-identical twin pair, the number of successes is just the number of boys. If we write s for the number of successes, a formula for s is

$$s = \sum_i (\text{number of boys in pair } i)(2 - I_i)/2.$$

Again, by the earlier example, the conditional (aka posterior) distribution of q in this setting will be Beta$(c + s, d + x + 2(n - x) - s)$, and we sample a new value of q from this.

We then just iteratively update each of $p, q, I_1, I_2, \ldots, I_n$.

The Markov chain given by the Gibbs sampler will converge to the joint posterior distribution of $p, q, I_1, I_2, \ldots, I_n$.

If we only wanted to learn about p and q, we could just ignore the values of the I's in the chain.

But for example by focussing on the values of I_k for some k we learn about the marginal posterior probability that the kth twin pair is identical.

(Note that from a frequentist perspective, with this missing data formulation, the mle of p and q can easily be found by the EM algorithm.)