STAT40380/STAT40390/STAT40850 Bayesian Analysis

Dr Niamh Russell

School of Mathematics and Statistics
University College Dublin

niamh.russell@ucd.ie

March 2016

More on the rejection algorithm

- Last time, we saw the *rejection algorithm*: a method for sampling from the posterior distribution given only $q(\theta|\mathbf{x}) \propto p(\theta|\mathbf{x})$
- The rejection algorithm works by proposing a set of posterior samples from a distribution $h()$, the accepting these with probability according to the ration $q(\theta|\mathbf{x})/Mh(\theta)$ where M is some upper bound for $p(\theta|\mathbf{x})/h(\theta)$
- When we say 'accept with probability' we mean:
  1. Sample a uniform $U(0,1)$ variable
  2. If U is less than the acceptance probability, accept the proposed value of θ, otherwise reject it
- Methods which involve *sampling* from the posterior distribution rather than trying to calculate it directly, are known as *Monte Carlo* methods

Comments on rejection sampling

- The rejection algorithm works well in situations where θ is bounded, eg $(0, 1)$ in last lecture's example
- It is also self-monitoring: if we are getting a low acceptance rate, we can adjust $h(0)$ to produce better samples
- Howver, it can be very hard to get a good acceptance rate when θ is unbounded or worse, long tailed
- Similarly, as the dimension of our problem increases, we find ourselves accepting fewer and fewer values
- Further methods we will study allow us to accept more values in higher dimensions
- These techniques are known as *Markov chain Monte Carlo* or MCMC

Introduction to Monte Carlo

Suppose $X \sim Ga(3,1)$, $Y \sim exp(4)$, $Z \sim \chi^2_4$.
What is the distribution of $Q = \frac{X+Y}{Z+5}$

- Solution 1: Use algebraic techniques on transformations of random variables
- Solution 2: Simulate a large number of X, Y and Z values, and simply compute Q for each value. We can then calculate the mean of Q, the pdf of Q or any other quantity we require

- Technique first invented in Los Alamos during mid-1940s, not widely available for general use until first personal computers of 1980s
- Before Monte Carlo methods existed, practical Bayesian statistics was nearly impossible

Monte Carlo integration

- We can use Monte Carlo to integrate because:

$$\mathbb{E}(x) = \int xp(x)dx \simeq \frac{1}{n}\sum_{i=1}^{n} x_i$$

  where x_i are samples from a chosen probability distribution

- Furthermore, we can create any function of the expected values, via

$$\mathbb{E}(f(x)) = \int f(x)p(x)dx \simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

- Thus, given examples from a distribution, we can compute means, variances, skewness, etc, etc
- The method readily extends to higher-dimensional cases

Markov chain Monte Carlo

- Markov chain Monte Carlo (or MCMC) is a very general method for drawing posterior samples based on values of θ from approximate distributions, and subsequently correcting them so as to draw from the true posterior
- The values are drawn *sequentially*, such that the distribution of the current draw depends only on the previous value drawn, hence a *Markov chain*
- We will label our MCMC simulations are $\theta^0, \theta^1, \theta^2, \ldots$ and will draw a new θ^t from a transition distribution $T_t(\theta^t|\theta^{t-1})$, ie depends only on the previous draw
- (Note that the transition probability might change with t)
- The key to MCMC is to create a Markov chain whose statioary distribution is that of the posterior $p(\theta|\mathbf{x})$. Thus if we run enough samples, we can eventually obtain samples from the posterior

The Gibbs sampler

- The *Gibbs sampler* is used in multivariate problems and works by sampling from each of the parameters in turn (much like the conditional maximisation method we saw in Lecture 13)
- Write $q(\boldsymbol{\theta}|\mathbf{x}) = q(\theta_1, \ldots, \theta_K|\mathbf{x})$
- Steps:
  1. Set inital values of $\theta_1^0, \theta_2^0, \ldots, \theta_K^0$
  2. Loop through iterations $t = 1, 2, \ldots$
     (a) Loop through parameters $k = 1, \ldots, K$
     (b) Sample θ_k^t from $q(\theta_k|\boldsymbol{\theta}_{-k}^{t-1})$ where $\boldsymbol{\theta}_{-k}^t = (\theta_1^t, \ldots, \theta_{k-1}^t, \theta_{k+1}^t, \ldots, \theta_K^t)$

Exercise 1: Gibbs Sampling

Example

Consider a single observation $\mathbf{y} = (y_1, y_2)$ from a bivariate normal likelihood with known covariance matrix and unknown mean $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$.

With a flat prior distribution on $\boldsymbol{\theta}$ the posterior distribution can be writtern as:

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} | \mathbf{y} \sim N\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$$

Outline the steps to perform Gibbs sampling from the posterior distribution when $\mathbf{y} = (0, 0)$ Use the starting values $\theta_1^0, \theta_2^0 = (2.5, 2.5)$ to create two more posterior samples

Exercise 2: Gibbs Sampling

Example

Observations x come from a normal distribution such that $x_i \sim N(\theta, \phi)$ with prior distributions:

$$\theta \sim N(\theta_0, \phi_0) \text{ and } \phi \sim IG(\alpha, \beta)$$

Use Gibbs sampling to outline the steps to update θ and ϕ when: $n = 12, \bar{x} = 119, S = 13045, \alpha = 6, \beta = 1000, \theta_0 = 110$ and $\phi_0 = 20$

Gibbs sampling notes

- Both of the above samples are from conjugate distributions so we can check that our samples are coming from the true posterior. Obviously this is not always the case!
- The distributions $q(\theta_k|\theta_{-k}^{t-1})$ are often known as the *full* or *complete conditionals*
- If the complete conditionals are standard probability distributions (and thus easy to sample from) Gibbs sampling is very efficient
- If the complete conditionals are not standard probability distributions, we may use another algorithm (such as the rejection algorithm)