STK 4021 - Applied Bayesian Analysis and Numerical Methods
Thordis L. Thorarinsdottir
Fall 2014

Problem Set 5

Problem 1 (sampling from a truncated normal). Let

$$p(z) \propto \frac{1}{\sigma}\varphi\Big(\frac{z-\mu}{\sigma}\Big)\mathbb{1}\{z \in (a,b)\},$$

where φ is the density of the standard normal distribution. Prove that the following sampling method generates a sample from the distribution with density p.

1. sample $u \sim \mathcal{U}\big([\Phi((a-\mu)/\sigma), \Phi((b-\mu)/\sigma)]\big)$
2. set $z = \mu + \sigma\Phi^{-1}(u)$

Here, Φ is the cdf of the standard normal distribution.

Problem 2 (disease rates). The number of occurences of a rare, nongenetic birth effect in a five-year period for six neighboring counties is $\mathbf{y} = (1, 3, 2, 12, 1, 1)$. The counties have populations of $\mathbf{x} = (33, 14, 27, 90, 12, 17)$, given in thousands. The second county has higher rates of toxic chemicals (PCBs) present in the soil samples, and it is of interest to know if this town has a high disease rate as well. We will use the following hierarchial model to analyze these data:

- $Y_i|\theta_i, x_i \sim \text{Po}(\theta_i x_i)$;
- $\theta_1, \dots, \theta_6|a, b \sim \Gamma(a, b)$;
- $a \sim \Gamma(1, 1)$; $b \sim \Gamma(10, 1)$.

(a) Describe in words what the various components of the hierarchical model represent in terms of the observed and expected disease rates.

(b) Identify the form of the conditional distribution of $p(\theta_1, \dots, \theta_6|a, b, \mathbf{x}, \mathcal{D})$, and from this identify the full conditional distribution of the rate for each county $p(\theta_i|\theta_{-i}, a, b, \mathbf{x}, \mathcal{D})$.

(c) Write out the ratio of the posterior densities comparing a set of proposal values (a^*, b^*, θ) to values (a, b, θ). Note that the value of θ, the vector of county-specific rates, is unchanged.

(d) Construct a Metropolis-Hastings algorithm which generates samples of (a, b, θ) from the posterior. Do this by iterating the following steps:

1. Given the current value (a, b, θ), generate a proposal (a^*, b^*, θ) by sampling a^* and b^* from a symmetric proposal distribution centered around a and b, but making sure all proposals are positive. Accept the proposal with the appropriate probability.

2. Sample new values of θ_j's from their full conditional distributions.

Perform diagnostic tests on your chain and modify if necessary.

(e) Make posterior inference on the infection rates using the samples from the Markov chain. In particular,

(i) Compute marginal posterior distributions of $\theta_1, \ldots, \theta_6$ and compare them to $y_1/x_1, \ldots, y_6/x_6$.

(ii) Examine the posterior distribution of a/b, and compare it to the corresponding prior distribution as well as to the average of y_i/x_i across the six counties.

(iii) Plot samples of θ_2 versus θ_j for $j \neq 2$, and draw a 45 degree line on the plot as well. Also, estimate $\mathbb{P}(\theta_2 > \theta_j | \mathbf{x}, \mathcal{D})$ for each $j \neq 2$ and $\mathbb{P}(\theta_2 = \max\{\theta_1, \ldots, \theta_6\} | \mathbf{x}, \mathcal{D})$. Interpret the results of these calculations, and compare them to the conclusions one might obtain if they just examined y_j/x_j for each county j.

Solutions will be discussed in class on October 3.