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

Problem Set 10

Problem 1 (categorical data and the Dirichlet distribution). Consider again the data on the number of children of men in their 30s from Problem 3 on Problem Set 2. These data could be considered as categorical data, as each sample Y lies in the discrete set $\{1, \ldots, 8\}$ (8 here actually denotes "8 or more" children). Let $\boldsymbol{\theta}_A = (\theta_{A,1}, \ldots, \theta_{A,8})$ be the proportion in each of the eight categories from the population of men with bachelor's degrees.

(a) Write in compact form the conditional probability given $\boldsymbol{\theta}_A$ of observing a particular sequence $\{y_{A,1}, \ldots, y_{A,8}\}$ for a random sample from the A population. This should be the so-called multinomial model. Explain how it generalizes the binomial model.

(b) Identify the sufficient statistic. Show that the Dirichlet family of distributions, with densities of the form

$$p(\boldsymbol{\theta}|\mathbf{a}) = \frac{\Gamma(a_1 + \ldots + a_K)}{\Gamma(a_1)\cdots\Gamma(a_K)}\theta_1^{a_1-1}\cdots\theta_K^{a_K-1}$$

are a conjugate class of prior distributions for this sampling model.

(c) Let $X_1, \ldots, X_K$ be independent with $X_j \sim \Gamma(a_j, 1)$ for $j = 1, \ldots, K$. Then the ratios

$$Z_1 = \frac{X_1}{X_1 + \cdots X_K}, \ldots, Z_K = \frac{X_K}{X_1 + \cdots + X_K}$$

follow a Dirichlet distribution with parameter $\mathbf{a} = (a_1, \ldots, a_K)$. Show this using the following: If X has density $f(x)$, and $Z = h(X)$ is a one-to-one transformation with inverse $X = h^{-1}(Z)$, then the density of Z is

$$g(z) = f(h^{-1}(z))\left|\frac{\partial h^{-1}(z)}{\partial z}\right|.$$

The function `rdir()` thus samples from the Dirichlet distribution:

```
rdir <- function(nsamp=1, a) # a is a vector
{
Z <- matrix(rgamma(length(a) * nsamp, a, 1), nsamp, length(a), byrow=T)
Z / apply(Z, 1, sum)
}
```

(d) Using the function in (c), generate 5000 or more samples of $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ from their posterior distributions. Also, obtain samples from the respective posterior predictive distributions.

(e) Obtain comparable samples from the posterior predictive distributions under the model in Problem 3 on Problem Set 2. Compare the results here to the results in (d).

Solutions will be discussed in class on November 7.