MCMC Short Course - Examples

George Casella

July 13, 2007

Introduction

1. For each of the following, plot the function in R. Use various values for μ, θ, σ and τ.

   (a) The censored data density arising from observing

   $$Z = X \wedge Y = \min(X, Y),$$

   where

   $$X \sim \mathcal{N}(\theta, \sigma^2) \quad \textbf{and} \quad Y \sim \mathcal{N}(\mu, \tau^2),$$

   which is given by

   $$\left[1 - \Phi\left(\frac{z-\theta}{\sigma}\right)\right] \times \tau^{-1}\varphi\left(\frac{z-\mu}{\tau}\right) + \left[1 - \Phi\left(\frac{z-\mu}{\tau}\right)\right] \sigma^{-1}\varphi\left(\frac{z-\theta}{\sigma}\right)$$

   where φ and Φ are the density and cdf of the normal $\mathcal{N}(0, 1)$ distribution.

   (b) The mixture of two normal distributions,

   $$p\mathcal{N}(\theta, \sigma^2) + (1-p)\mathcal{N}(\mu, \tau^2)\ ,$$

   with density

   $$p\sigma^{-1}\varphi\left(\frac{x-\theta}{\sigma}\right) + (1-p)\ \tau^{-1}\ \varphi\left(\frac{x-\mu}{\tau}\right)]$$

   (c) The likelihood based on observing $(X_1, X_2, X_3) = (0, 5, 9)$ from the Student's t density proportional to

   $$\sigma^{-1}\left(1 + \frac{(x-\theta)^2}{p\sigma^2}\right)^{-(p+1)/2},$$

   with $p = 1$ and $\sigma = 1$ (the standard Cauchy).

   - **Plot.txt**

Random Variable Generation

2. Check the R uniform random number generator:

   (a) Generate 1,000 uniform random variables and make a histogram

(*b*) Generate uniform random variables $(X_1, \ldots, X_n)$ and plot the pairs (X_i, X_{i+1}) to check for autocorrelation.

- **Uniform.txt**

3. (*a*) Generate a binomial(n, p) random variable with $n = 25$ and $p = .2$. Make a histogram and compare it to the binomial mass function, and to the R binomial generator.

   - **binomial.txt**

   (*b*) Generate $5,000$ *logarithmicseries* random variables with mass function

   $$P(X = x) = \frac{-(1-p)^x}{x \log p}, \quad x = 1, 2, \ldots \quad 0 < p < 1.$$

   Make a histogram and plot the mass function.

   - **logarithmic.txt**

4. In each case generate the random variables and compare to the density function

   (*a*) Normal random variables using a Cauchy candidate in Accept/Reject

   (*b*) Gamma$(4.3, 6.2)$ random variables using a Gamma$(4, 7)$.

   (*c*) Truncated normal - Standard normal truncated to $(2, \infty)$

   - **RandomVariables.txt**

Integration

5. For the Bayes estimator

   $$\delta(x) = \frac{\int_{-\infty}^{\infty} \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}{\int_{-\infty}^{\infty} \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}$$

   (*a*) Plot the integrand and use MCI to calculate the integral.

   (*b*) Monitor the convergence with the standard error of the estimate. Obtain three digits of accuracy with probability .95.

   - **BayesEstimator.txt**

6. For a standard normal random variable Z, calculate $P(Z > 2.5)$ using

(*a*) Monte Carlo sums based on indicator functions

(*b*) Importance sampling based on a candidate exponential $f(x) = e^{-(x-2.5)}$, $x > 2.5$.

Note that $P(Z > 2.5) = .0062$.

- **TailProb.txt**

Optimization

7. Use the R functions "optim" and "optimize" to find the maximum of

$$f(x) = [\cos(50x) + \sin(20x)]^2 .$$

Compare to the results of a stochastic exploration.

- **Optimize.txt**

8. For each of the likelihood functions in Exercise 1, find the maximum with "optimize" and stochastic exploration.

9. The follow are genotype data on blood type

Genotype	Probability	Observed	Probability	Frequency
AA	p_A^2	A	$p_A^2 + 2p_Ap_O$	$n_A = 186$
AO	$2p_Ap_O$			
BB	p_B^2	B	$p_B^2 + 2p_Bp_O$	$n_B = 38$
BO	$2p_Bp_O$			
AB	$2p_Ap_B$	AB	p_Ap_B	$n_{AB} = 13$
OO	p_O^2	O	p_O^2	$n_O = 284$

Because of dominance, we can only observe the genotype in the third column, with probabilities given by the fourth column. The interest is in estimating the allele frequencies p_A, p_B, and p_O (which sum to 1).

(*a*) Under a multinomial model, verify that the observed data likelihood is proportional to

$$(p_A^2 + 2p_Ap_O)^{n_A}(p_B^2 + 2p_Bp_O)^{n_B}(p_Ap_B)^{n_{AB}}(p_O^2)^{n_O}$$

(*b*) With missing data Z_A and Z_B, verify the complete data likelihood

$$(p_A^2)^{Z_A}(2p_Ap_O)^{n_A - Z_A}(p_B^2)^{Z_B}(2p_Bp_O)^{n_B - Z_B}(p_Ap_B)^{n_{AB}}(p_O^2)^{n_O}$$

(c) Verify that the missing data distribution is

$$Z_A \sim \text{binomial}\left(n_A, \frac{p_A^2}{p_A^2 + 2p_A p_O}\right) \text{ and } Z_B \sim \text{binomial}\left(n_B, \frac{p_B^2}{p_B^2 + 2p_B p_O}\right),$$

and write an EM algorithm to estimate p_A, p_B, and p_O

- **EMBlood.txt**

Metropolis-Hastings

10. Calculate the mean of a Gamma$(4.3, 6.2)$ random variables using

    (a) Accept-Reject with a Gamma$(4, 7)$ candidate.

    (b) Metropolis-Hastings with a Gamma$(4, 7)$ candidate.

    (c) Metropolis-Hastings with a Gamma$(5, 6)$ candidate.

    In each case monitor the convergence.

- **GammaAR.txt**

11. The Institute for Child Health Policy (ICHP) at the University of Florida studies the effects of health policy decisions on children's health. A small portion of one of their studies follows.

    The overall health of a child (metq) is rated on a 1-3 scale, with 3 being the worst. Each child is in an HMO (variable np, 1=nonprofit, -1=for profit). The dependent variable of interest (y_{ij})is the use of an emergency room (erodds, 1=used emergency room, 0=did not). The question of interest is whether the status of the HMO affects the emergency room choice.

    (a) An appropriate model is logistic regression

    $$\text{logit}(p_{ij}) = a + bx_i + cz_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i.$$

    where x_i is the HMO type, z_{ij} is the health status of the child, and p_{ij} is the probability of using an emergency room. Verify that the likelihood function is

    $$\prod_{i=1}^{k} \prod_{j=1}^{n_i} \left(\frac{\exp(a + bx_i + cz_{ij})}{1 + \exp(a + bx_i + cz_{ij})}\right)^{y_{ij}} \left(\frac{1}{1 + \exp(a + bx_i + cz_{ij})}\right)^{1-y_{ij}}.$$

    (Here we are only distinguishing between for-profit and non-profit, so $k = 2$.)

(*b*) Run a standard GLM on these data (at LogisticData.txt) and get the estimated mean and variance of a, b, and c.

(*c*) Use normal candidate densities with mean and variance at the estimates in a Metropolis-Hastings algorithm that samples from the likelihood. Get histograms of the parameter values.

- **MHLogistic.txt**

Gibbs Sampling

12. Referring to Exercise 9, estimate p_A, p_B and p_O using a Gibbs sampler. Make a histogram of the samples.

- **GibbsBlood.txt**

13. A subset of the clinical mastitis data is

$$0, 0, 1, 1, 2, 2, 2, 2, 2, 2, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6,$$
$$6, 8, 8, 8, 9, 9, 9, 10, 10, 12, 12, 13, 13, 13, 13, 18,$$
$$18, 19, 19, 19, 19, 20, 20, 22, 22, 22, 23, 25$$

For each herd we model the mean number of occurrences as a Poisson mean λ_i, with a hierarchy to account for overdispersion. A model is

$$\begin{aligned} X_i &\sim \text{Poisson}(\lambda_i), \\ \lambda_i &\sim \text{Gamma}a(\alpha, \beta_i), \\ \beta_i &\sim \text{Gamma}a(a, b), \end{aligned}$$

where α, a, and b are specified. The posterior density of λ_i, $\pi(\lambda_i|\mathbf{x}, \alpha)$, can now be obtained from the Gibbs sampler

$$\begin{aligned} \lambda_i &\sim \pi(\lambda_i|\mathbf{x}, \alpha, \beta_i) = \text{Gamma}a(x_i + \alpha, 1 + \beta_i), \\ \beta_i &\sim \pi(\beta_i|\mathbf{x}, \alpha, a, b, \lambda_i) = \text{Gamma}a(\alpha + a, \lambda_i + b)\,. \end{aligned}$$

(*a*) For $\alpha = .1, a = b = 1$ run the Gibbs sampler. Make histograms and monitor the convergence of λ_5, λ_{15}, and β_{15}

(*b*) Investigate the sensitivity of your answer to the specification of α, a and b.

- **Mastitis.txt**