8 Feb 2013

eBay-Google Bayesian short course: Problem set 4

1. (First practice with `WinBUGS`.) Write a `WinBUGS` program to use Gibbs sampling to analyze the data in the length-of-stay case study, using the same Gamma prior and Poisson likelihood as in that example. Obtain MCMC approximations both for the posterior distribution of λ given the data vector y and the predictive distribution $p(y_{n+1}|y)$, and compare summaries of these distributions (means, SDs, histograms or density traces) with the theoretical conjugate results we got in the case study. You don't need to worry about MCMC diagnostics in this simple example, because Gibbs sampling when there's only one parameter amounts to IID sampling from the relevant posterior and predictive distributions. Justify your choices of initial values for the Markov chain and length of burn-in period. Use one of the formulas given in class to work out how long you need to monitor the chain to report 3-significant-figure accuracy of the posterior mean estimates for both λ and y_{n+1}, and verify that you do indeed achieve that level of accuracy (at least up to Monte Carlo noise) in your simulation. What length of monitoring run is necessary to report 3-significant-figure accuracy of the posterior **SD** estimate? Explain briefly, and report all relevant calculations (simulation or otherwise).

2. (Second practice with `WinBUGS`.) In problem 3 of Problem Set 2 we used conjugate inference to fit an Exponential sampling model to the wire failure data given in that problem, and you may remember noticing that the biggest data value (21194) seemed a bit large in the Exponential context, which tentatively called the Exponential distribution into question. Recalling that the basic Bayesian idea for improving a model is to *expand* it by embedding it in a richer class of models of which it's a special case, the natural thing to try is to fit a model to this data set in which the sampling distribution is Gamma (we saw in part 2 of the lecture notes that the Exponential is a special case of the $\Gamma(\alpha, \beta)$ family with $\alpha = 1$). Write a `WinBUGS` program to use MCMC to fit the model

$$\begin{aligned} (\alpha, \beta) &\sim p(\alpha, \beta) \\ (y_i|\alpha, \beta) &\overset{\text{IID}}{\sim} \Gamma(\alpha, \beta), \quad i = 1, \dots, n \end{aligned} \tag{1}$$

to the wire failure data. For this problem, by way of prior information (unlike the situation in Problem Set 2) let's use a diffuse prior on α and β. Since they both live on $(0, \infty)$ it's natural to try independent $\Gamma(\epsilon, \epsilon)$ priors for both of them, with (as usual) a small value for ϵ like 0.001; or you could use an initial run with $\Gamma(\epsilon, \epsilon)$ priors to see where the likelihood is appreciable and then use $U(0, c_\alpha)$ and $U(0, c_\beta)$ priors for α and β, where c_α and c_β are chosen to be big enough not to truncate the likelihood but not much larger than that. Summarize the posterior distribution on α and β to an appropriate degree of Monte Carlo accuracy. Does the $\Gamma(\alpha, \beta)$ family appear to provide a better fit to the wire failure data than the Exponential sampling distribution used in Problem Set 2? Explain briefly.

3. (Multinomial data and the Dirichlet distribution as a prior; based on Section 3.5 in Gelman et al.) In late October 1988, CBS News conducted a survey which was equivalent to a simple random sample of $n = 1,447$ American adults to learn about voter preferences

in the Presidential election which was to be held a few weeks later. $y_1 = 727$ of these people supported George Bush (the elder), $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or responded "no opinion." This situation is a lot like the AMI mortality case study in class except that there are three outcome categories (Bush, Dukakis, other) instead of two (died, lived): before any data arrives you would probably agree that your uncertainty about the string of 1,447 individual outcomes from each sampled person (which is summarized by the counts $y = (y_1, y_2, y_3) = (727, 583, 137)$) is exchangeable. This leads by an easy generalization of de Finetti's representation theorem for binary outcomes to the following model for the summary counts:

$$\begin{aligned} (\theta_1, \ldots, \theta_k) &\sim p(\theta_1, \ldots, \theta_k) \\ p(y_1, \ldots, y_k | \theta_1, \ldots, \theta_k) &= c \prod_{j=1}^{k} \theta_j^{y_j}, \end{aligned} \qquad (2)$$

where $0 < \theta_j < 1$ for all $j = 1, \ldots, k$ and $\sum_{j=1}^{k} \theta_j = 1$. The second line of (2) (the sampling distribution of the vector y, which defines the likelihood function) is the *multinomial* distribution, an obvious generalization of the binomial to $k > 2$ categories (in this voting problem $k = 3$). Evidently in this model the conjugate prior for the vector $\theta = (\theta_1, \ldots, \theta_k)$ is of the form

$$p(\theta_1, \ldots, \theta_k | \alpha_1, \ldots, \alpha_k) = c \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}; \qquad (3)$$

this distribution turns out to be well-behaved for any choice of the hyperparameter vector $\alpha = (\alpha_1, \ldots, \alpha_k)$ such that $\alpha_j > 0$ for all $j = 1, \ldots, k$. This is the *Dirichlet*(α) distribution, a generalization of the Beta distribution to more than two categories. With this prior the model becomes

$$\begin{aligned} (\theta_1, \ldots, \theta_k) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \\ (y_1, \ldots, y_k | \theta_1, \ldots, \theta_k) &\sim \text{Multinomial}(n; \theta_1, \ldots, \theta_k) \end{aligned} \qquad (4)$$

(see Appendix A in Gelman et al. for the normalizing constants). As with the Beta distribution, the α_j can clearly be seen in this model to represent prior sample sizes; in the voting example, choosing a particular $(\alpha_1, \alpha_2, \alpha_3)$ is equivalent to assuming that the prior is equivalent to a data set with α_1 preferences for Bush, α_2 for Dukakis, and α_3 for other. To create a diffuse prior, which would be a natural choice in the absence of any earlier sampling data (and even with earlier data it's not clear that voter opinion is sufficiently stable over time to make simple use of any previous polling results), we evidently want the α_j to be small; an easy choice that avoids complications with improper priors is to take $\alpha = (1, \ldots, 1)$, a multivariate generalization of the uniform distribution. The main scientific interest in this problem focuses on $\gamma = (\theta_1 - \theta_2)$, the margin by which Bush is leading Dukakis.

(a) Write out the likelihood function for the vector θ in the Multinomial sampling model above, and compute the maximum likelihood estimates of the θ_i and of γ. You can either do this by (i) expressing the log likelihood as a function of θ_1, θ_2, and θ_3 and performing a constrained maximization of it using Lagrange multipliers, or (ii) substituting $\theta_3 = (1 - \theta_1 - \theta_2)$ and $y_3 = (n - y_1 - y_2)$ into the log likelihood and carrying out an unconstrained maximization in the usual way (by setting first partial derivatives to 0 and solving). Do the MLEs have reasonable intuitive forms? Explain briefly. (extra credit) On the web or in a statistics text, read about how Fisher information generalizes when the parameter θ of interest is a vector and use this to compute approximate large-sample standard errors for the MLEs of the θ_i and of γ.

(b) Use `BUGS` or `WinBUGS` with the diffuse prior mentioned above to simulate m draws from the marginal posterior distributions for the θ_i and for γ, where m is large enough to yield results that seem accurate enough to you given the context of the problem (briefly justify your choice of m). How do the posterior means of the θ_i compare with the MLEs? Explain briefly. Report the posterior mean and SD of γ, and compare your estimated posterior density with the plot below, which is taken from Gelman et al. Use your MCMC output to estimate $p(\gamma > 0|y)$, the chance that Bush would win the election if it were held shortly after the data were gathered and the "other" (non-Bush, non-Dukakis) voters behaved appropriately (briefly explain what has to be assumed about these other voters so that $p(\gamma > 0|y)$ $\boxed{\text{is}}$ the chance that Bush would win the election), and attach a Monte Carlo standard error to your estimate of $p(\gamma > 0|y)$. Describe your MCMC sampling strategy (mainly your starting values and the length b of your burnin run; you've already justified your choice of m) and briefly explain why you believe that this strategy has accurately extracted the posterior distribution of interest.

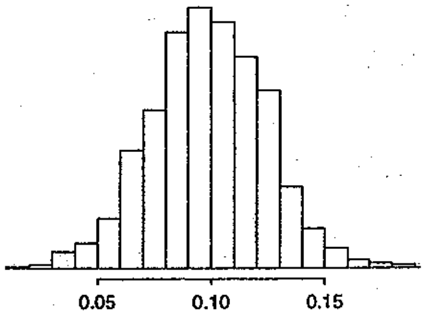

Figure 3.3. ***Histogram of values of $(\theta_1 - \theta_2)$ for 1000 simulations from the posterior distribution for the election polling example.***

(c) (extra credit) Use `Maple` or some equivalent environment (or paper and pen, if you're brave) to see if you can derive a closed-form expression for $p(\gamma|y)$, and compare your mathematical result with your simulation-based findings in (a), using the actual data in this example.

4. (extra credit) Write your own Metropolis-Hastings sampler to analyze the data in the length-of-stay case study, using the same Gamma prior and Poisson likelihood as in that example; using your MH sampler, complete as many of the steps in problem 1 of this assignment as you have time and patience for, and compare the results you obtained in problem 1 with Gibbs sampling. In choosing a proposal distribution for your MH sampler there are two main ways to go: you can either (i) transform λ to the log scale so that it lives on the entire real line and use (something like) a Gaussian proposal distribution for

$\eta = \log(\lambda)$ (in this case you'll be using the simpler Metropolis form for the acceptance probability), or (ii) pick a proposal distribution for λ that simulates from the positive part of the real line (a natural choice would be the family of Gamma distributions; in this case you'll be using the more complicated MH form for the acceptance probability). In either (i) or (ii) you'll find that some measure of scale for the proposal distribution acts like a tuning constant that can be adjusted to achieve optimal MH Monte Carlo efficiency. If you have time it would be good to make a small study of how the MCSE of the posterior mean for λ or η depends on this tuning constant, so that you can find the optimal scaling of the proposal distribution.