AMS 207 - Homework 2

Red Hawk Data

Mary Silva

Write the Bayes factor, BIC, DIC and Gelfand and Ghost criterion to compare a model where n observations are assumed to be sampled with a poisson distribution with a gamma prior, to a model where the observations are sampled from a binomial distribution, with a fixed, large, number of trials and beta prior for the probability of success.

1. Consider the data on red tailed hawks. Fit the data using the two different models. Notice that there is no hierarchical structure in this case, as opposed to what was assumed in the take home part of Test 1. Ignore the route counts.

We first fit the data using a Poisson distribution with a Gamma prior:

$$\begin{aligned} M_1 : &X_i \sim \text{Poisson}(\lambda); \qquad i = 1, ..., n \\ &\lambda \sim Gamma(\alpha, \beta) \end{aligned}$$

The likelihood for model 1 is given by:

$$f_1(\boldsymbol{x}|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

The prior for model 1 is given by

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp -\lambda\beta$$

Which leads to the following posterior

$$\begin{aligned} \pi_1(\lambda|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\lambda)p(\lambda) && (1) \\ &\propto \frac{\beta^\alpha}{\Gamma(\alpha)\prod x_i!}\exp\Big\{-\lambda(n+\beta)\Big\}\lambda^{\sum x_i+\alpha-1} && (2) \\ & && (3) \end{aligned}$$

Which we recognize as the kernel of a Gamma distribution. Thus,

$$\lambda|\boldsymbol{x} \sim \sim Gamma\left(\alpha^* = \sum x_i + \alpha, \beta^* = n + \beta\right) \qquad (4)$$

For the second model, we fit the data using a binomial distribution with a fixed, large number of trials and a beta prior

$$\begin{aligned} M_2 : &X_i \sim Binom(n, \theta); \qquad i = 1, ..., n \\ &\theta \sim Beta(\alpha, \beta) \end{aligned}$$

The Likelihood:

$$f_2(\boldsymbol{x}|\theta) = \prod_{i=1}^{n}\binom{N}{x_i}\theta^{\sum x_i}(1-\theta)^{nN-\sum x_i}$$

The prior:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Throughout this paper, I use large $N = 10^6$.
The posterior distribution for model 2 up to a constant is:

$$\pi_2(\lambda|\boldsymbol{x}) \propto \left[\prod_{i=1}^{n} \binom{N}{x_i}\right] \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\sum x_i+\alpha-1}(1-\theta)^{nN+\beta-\sum x_i -1} \tag{5}$$

Which we recognize as the kernel for Beta distribution, therefore

$$\theta|\boldsymbol{x} \sim Beta\left(\alpha^* = \sum x_i + \alpha, \beta^* = nN + \beta - \sum x_i\right) \tag{6}$$

2. Perform a prior sensitivity analysis.

   We are going to explore various non-informative and informative priors for both models. For model M_1, as an informative prior, we want to choose α and β such that the prior mean, $E(\lambda) = \frac{\alpha}{\beta}$, is proportional to the maximum likelihood estimator $\hat{\lambda}$ = mean(data) = 376.1. So we will set $\alpha = 376.1$ and $\beta = 1$ as our informative prior and compare to weakly informative priors (Figure 1). We plot the posterior distributions using different α and β values and attaining the α^* and β^* values obtained in equation (4).

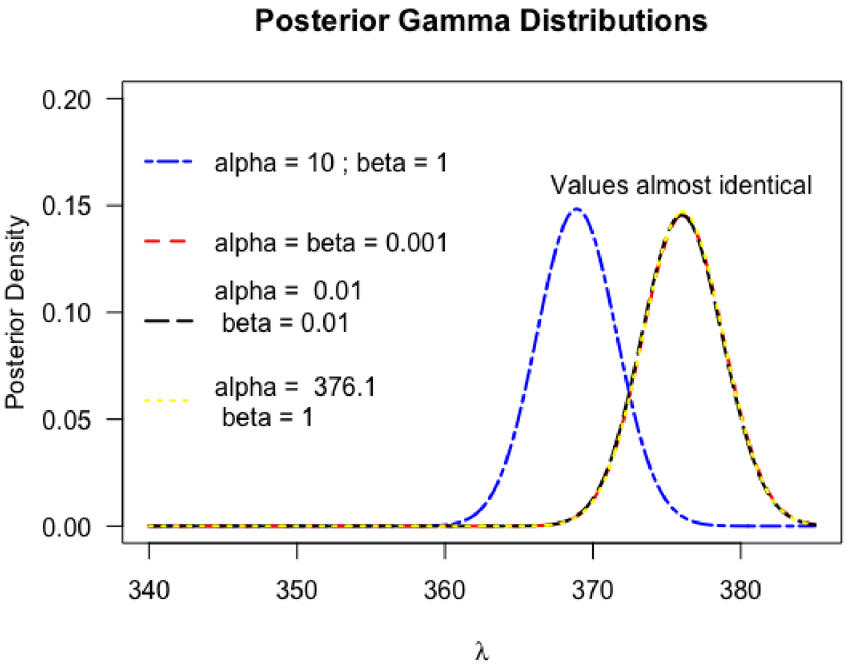

Figure 1: Posterior densities for $\lambda|\boldsymbol{x} \sim Gamma(\alpha^*, \beta^*)$. We see that the posterior distribution for model 1 is not sensative to prior selection when the prior mean is centered around the MLE.

   Similarly, for model M_2, we want to select informative priors α and β for the Beta distribution so that the prior mean is proportional to the MLE. In other words, we want the prior mean to be centered around $376.1/N$ = 3.761e-04 (Figure 2). We note that, based on the plot, prior selection makes very little difference for the second model.

3. With n being the number of red tailed hawk observations, we define the following criterion:

BIC

$$BIC = -2 \times \text{Likelihood} - \log(n)$$

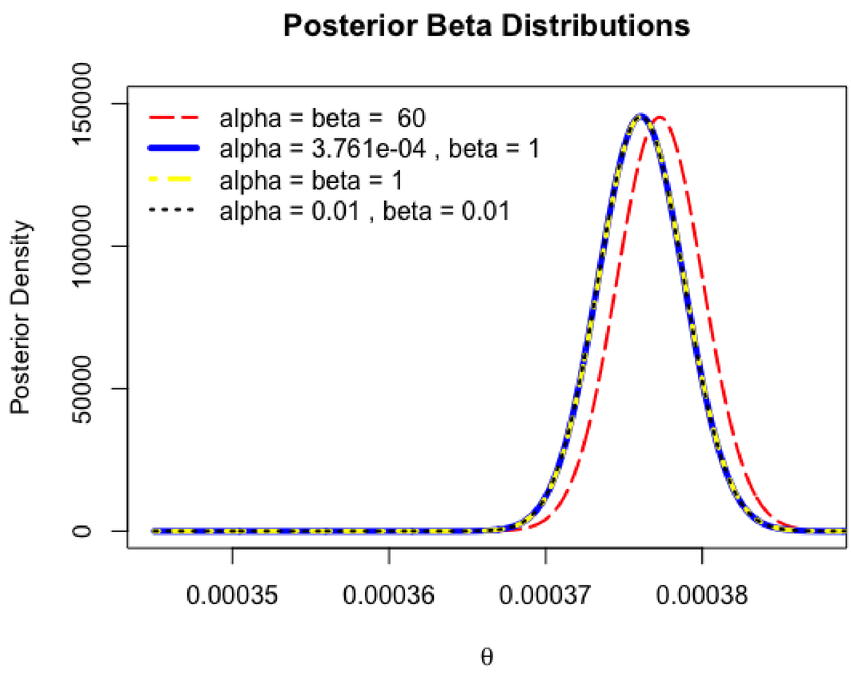

Figure 2: Posterior densities for $\theta|\boldsymbol{x} \sim Beta(\alpha^*, \beta^*)$. We see that the posterior distribution for model 1 is not sensative to prior selection, thus any of the above priors would be justified in using.

DIC

For the Deviance Information Criterion (DIC), the deviance statistic is given by

$$D(\theta) = -2 \log \text{Likelihood} + 2 \log h(x)$$

where $h(x)$ is the standardizing function. Then the DIC is given by

$$DIC = \bar{D} + (\bar{D} - D(\bar{\theta})) = 2\bar{D} - D(\bar{\theta})$$

Where $D(\bar{\theta})$ is the mean of posterior samples.

Gelfand and Ghosh

For the Gelfand and Ghosh criterion, denote the observed data as $\boldsymbol{x}$ and z as the new predicted value given x. Therefore we have the following posterior predictive distribution for

$$\begin{aligned} M_1 : p_1(z|\boldsymbol{x}) &= \int_0^\infty f_1(z|\lambda)\pi_1(\lambda|\boldsymbol{x})d\lambda \\ &= \frac{\Gamma(z + \sum x_i + \alpha)}{\Gamma(\sum x_i + \alpha)\Gamma(z+1)} \left(\frac{n+\beta}{n+\beta+1}\right)^{\sum x_i + \alpha} \left(\frac{1}{n+\beta+1}\right)^z \end{aligned}$$

Which is a negative binomial distribution with mean

$$\frac{\sum x_i + \alpha}{n+\beta}$$

and variance

$$\frac{\sum x_i + \alpha}{(n+\beta)^2}(n+\beta+1)$$

The posterior predictive distribution for model 2 is

$$\begin{aligned} M_2 : p_2(z|\boldsymbol{x}) &= \int_0^1 f_2(z|\theta)\pi_2(\theta|x)d\theta \\ &= \binom{N}{z} \frac{Beta(z+\alpha_0^*, n-z+\beta_0^*)}{Beta(\alpha_0^*, \beta_0^*)} \end{aligned}$$

where

$$\alpha_0^* = \sum x_i + \alpha$$

and

$$\beta_0^* = nN + \beta - \sum x_i$$

We recognize this as a Beta-Binomial distribution with mean

$$\frac{N\alpha_0^*}{\alpha_0^* + \beta_0^*}$$

and variance

$$\frac{N\alpha_0^*\beta_0^*(\alpha_0^* + \beta_0^* + n)}{(\alpha_0^* + \beta_0^*)^2(\alpha_0^* + \beta_0^* + 1)}$$

(Sorry, there's a lot of alphas and betas going on around here, its getting ugly; but Wikipedia, am I right?) Let $\mu_l = E(z_l|\boldsymbol{x})$ and $\sigma_l^2 = Var(z_l|\boldsymbol{x})$. Then the Gelfand and Ghosh criterion is given by

$$D = G + P,$$

where

$$G = \sum_l (\mu_l - x_l)^2$$

and

$$P = \sum_l \sigma_l^2$$

D seeks to reward goodness of fit penalizing complexity. So the smaller the D the better.

Bayes Factor

Next we look at the Bayes factor, which is simply a ratio of the marginals for model 1 and model 2.

Here we compute the marginals:

$$\begin{aligned} m_1(x) &= \int f_1(x|\lambda)p_1(\lambda)d\lambda & (7) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)\prod(x_i!)} \int e^{-\lambda(n+\beta)}\lambda^{\sum x_i + \alpha - 1} d\lambda & (8) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)\prod(x_i!)} \frac{\Gamma(\sum x_i + \alpha)}{(n+\beta)^{\sum x_i + \alpha}} & (9) \end{aligned}$$

$$\begin{aligned} m_2(x) &= \int f_2(x|\lambda)p_2(\lambda)d\lambda & (10) \\ &= \left[\prod_1^n \binom{N}{x_i}\right] \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{\sum x_i \alpha - 1}(1-\theta)^{nN+\beta-\sum x_i - 1} & (11) \\ &= \left[\prod_1^n \binom{N}{x_i}\right] \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i + \alpha)\Gamma(nN + \beta - \sum x_i)}{\Gamma(\alpha + nN + \beta)} & (12) \end{aligned}$$

To calculate the Bayes factor, we use our previously defined α and β parameters from part 2, and compute the ratio of the log marginals and then exponentiate the value.

$$\text{Bayes Factor} = \exp\{\log m_1 - \log m_2\} = 0.82$$

Our Bayes factor is close to 1, therefore there is no real evidence in favor of either model over the other.

Criterion Summary/Conclusions

We provide a table specifying our R calculations with using the informative priors previously specified.

	BIC	DIC	Gelfand and Ghosh	Bayes Factor
M1	2116.62	2122.52	549018.3	0.82
M2	2124.98	2123.03	549019.5	
Difference	8.35	0.51	1.17	

From the table, we see that the BIC, DIC, and G&G criterion are slightly smaller for model 1 (the gamma posterior model). This implies that model 1 is the "better" model. But in this case since the difference is slight, so we conclude that for model 2 (the beta posterior model) when we use a fixed, large, number of trials for the binomial prior as we have done, both model 1 and model 2 are essentially equivalent in terms of fitting the data, which is to be expected.