Chase Joyner

882 Homework 1

September 7, 2016

## Problem:

In this problem, we will consider developing a Bayesian model for Poisson data; i.e., our observed data will consist of $Y_1, ..., Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Recall, a random variable $Y$ is said to follow a Poisson distribution with mean parameter $\lambda$ if its pmf is given by

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} I\big(y \in \{0, 1, 2, ...\}\big).$$

Note, the Poisson model is often used to analyze count data.

(a) For the Poisson model, identify the conjugate prior. This should be a general class of priors.

**Solution:** The conjugate prior is a Gamma distribution. To see this, impose the prior $\lambda \sim \text{Gamma}(a, b)$. Then, we have

$$\begin{aligned} p(\lambda|\mathbf{y}) &\propto p(\mathbf{y}|\lambda) \cdot \pi(\lambda) \\ &\propto \prod_{i=1}^{n} e^{-\lambda}\lambda^{y_i} \cdot \lambda^{a-1}e^{-b\lambda} \\ &= e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i} \cdot \lambda^{a-1}e^{-b\lambda} \\ &= \lambda^{n\overline{y}+a-1}e^{-(b+n)\lambda}. \end{aligned}$$

Therefore, the posterior for $\lambda$ is $\text{Gamma}(a + n\overline{y}, b + n)$ and hence the conjugate prior under the Poisson model is the Gamma distribution.

(b) Under the conjugate prior, derive the posterior distribution of $\lambda|y$. This should be a general expression based on the choice of the hyper-parameters specified in your prior.

**Solution:** Shown above.

(c) Find the posterior mean and variance of $\lambda|y$. These should be general expressions based on the choice of the hyper-parameters specified in your prior.

**Solution:** Since the posterior distribution is Gamma, we see that

$$E[\lambda|\mathbf{y}] = \frac{a + n\overline{y}}{b + n} \quad \text{and} \quad V\big(\lambda|\mathbf{y}\big) = \frac{a + n\overline{y}}{(b + n)^2}.$$

(d) Obtain the MLE of $\lambda$. Develop and discuss a relationship that exists between the MLE and posterior mean identified in (c).

**Solution:** The log-likelihood function is

$$\ell(\lambda|\mathbf{y}) = \log p(\mathbf{y}|\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$
$$= \log \frac{e^{-n\lambda}\lambda^{n\overline{y}}}{\prod_{i=1}^{n} y_i!} = -n\lambda + n\overline{y}\log\lambda - \sum_{i=1}^{n} \log y_i!.$$

Taking the derivative wrt to $\lambda$, we have

$$-n + \frac{n\overline{y}}{\lambda} \overset{\text{set}}{=} 0$$

which yields that $\hat{\lambda}_{MLE} = \overline{y}$. The relationship between the MLE of $\lambda$ and the posterior mean for $\lambda$ is a weighted average. Specifically, the MLE for $\lambda$ is $\overline{y}$, and the posterior mean can be thought of as a weighted average of $\overline{y}$ and the prior mean.

(e) Write two separate R programs which can be used to find both a $(1-\alpha)100\%$ equal-tailed credible interval and a $(1-\alpha)100\%$ HPD credible interval for the Poisson model. These programs should take as arguments the following inputs: the observed data, prior hyperparameters, and significance level.

**Solution:** See appendix for code. Can also email code file.

(f) Find a data set which could be appropriately analyzed using the Poisson model. This data set should be of interest to you, and you should discuss, briefly, why the aforementioned model is appropriate; i.e., consider independence, identically distributed, etc. You will also need to provide the source of the data.

**Solution:** The dataset that I chose to analyze includes the number of major (category 3 and above) hurricanes in the United States per decade, starting in 1850 and ending in 2000, i.e. years 1851-1860, 1861-1870, ..., 1991-2000. The source for this dataset is: *http://www.nhc.noaa.gov/pastdec.shtml.* The dataset can be found in the table below:

| Decade | 1851-1860 | 1861-1870 | 1871-1880 | 1881-1890 | 1891-1900 | 1901-1910 | 1911-1920 | 1921-1930 | 1931-1940 | 1941-1950 | 1951-1960 | 1961-1970 | 1971-1980 | 1981-1990 | 1991-2000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hurricanes | 6 | 1 | 7 | 5 | 8 | 4 | 7 | 5 | 8 | 10 | 8 | 6 | 4 | 5 | 5 |

It is reasonable to assume a Poisson model since we have count data, i.e. integer values in the set $\{0, 1, 2, ...\}$. Additionally, the independent and identically distributed assumptions are reasonable since the amount of hurricanes in one decade does not really provide information on the amount of hurricanes that will occur in the next decade and there is no reason to assume more hurricanes will occur in one decade over another.

(g) Analyze the data set you have selected in (e). Provide posterior point estimates of $\lambda$, credible intervals, etc. Your analysis should be accompanied by an appropriate discussion of your findings.

**Solution:** Using hyper-parameters of $a = 0$ and $b = 0$, the posterior mean for $\lambda$ is

$$E[\lambda|\mathbf{y}] = \frac{a + n\overline{y}}{b + n} = \frac{0 + 15 \cdot \frac{89}{15}}{0 + 15} = \frac{89}{15} = 5.933.$$

Also, we have the equal-tailed credible interval to be

$$[4.765, 7.228]$$

and the HPD credible interval to be

$$[4.724, 7.181].$$

Therefore, we conclude there is a 95% probability that the true value of $\lambda$ falls between 4.724 and 7.181. A reasonable statement based on these findings is that the number of hurricanes in a decade is Poisson distributed with parameter $\lambda = 5.933$. However, using this one point estimate does not include the uncertainty for $\lambda$; i.e. it does not account for other possible outcomes but instead only on the data that we saw.

# APPENDIX

```
######################################################################
######################################################################
######
######  Two functions:  one to create (1-alpha)100% equal-tailed
######                  credible interval and a (1-alpha)100%
######                  HPD credible interval for lambda.
######
######################################################################
######################################################################

##### (1-alpha)100% equal-tailed credible interval #####

ETCI = function(y, a = 0, b = 0, alpha){
        n = length(y)
        return(qgamma(c(0.025, 0.975), a + n * mean(y), b + n))
}

ETCI(y, alpha = 0.05)

##### (1-alpha)100% HPD interval #####

HPD.h = function(y, a = 0, b = 0, h = 0.1){
        n = length(y)
        apost = a + n * mean(y)
        bpost = b + n
        mode = (apost - 1) / bpost
        dmode = dgamma(mode, apost, bpost)

        ## divide by dmode below to get on scale of 0 to 1 ##
        lint = uniroot(f = function(x){dgamma(x, apost, bpost) / dmode - h},
                       lower = 0, upper = mode)$root
        uint = uniroot(f = function(x){dgamma(x, apost, bpost) / dmode - h},
                       lower = mode, upper = 10000)$root

        ## calculate coverage, should be around (1-alpha)100% ##
        coverage = pgamma(uint, apost, bpost) - pgamma(lint, apost, bpost)

        return(c(lint, uint, coverage, h))
}
```

```
HPD = function(h, y, alpha){
        cov = HPD.h(y, h = h)[3]
        res = (cov - (1 - alpha))^2
        return(res)
}

h.final = optimize(HPD, c(0,1), y = y, alpha = 0.05)$minimum
HPD.h(y, h = h.final)
```