Bayensian Learning

Homework 2: Bayesian Prediction from a Parametric Model

Emilie Krutnes Engen

Master in Big Data Analytics
Carlos III University of Madrid

Universidad
Carlos III de Madrid

Bayesian Prediction of Hurricane Activity

In this exercise we consider the hurricane activity on the coast of the Atlantic Ocean as a stochastic process where the number of hurricanes h follows a Poisson distribution. The Poisson distribution is given by the following equation.

$$f(h|\lambda) = e^{-\lambda}\frac{\lambda^h}{h!}, \quad h = 0, 1, 2, \ldots, \lambda > 0 \tag{1}$$

The objective is to estimate the occurrence rate λ. The standard estimate for λ is the maximum likelihood estimate (MLE) (Albert, 2009). However, when the number of occurrences is close to zero, this estimate show poor performance. Thus, it is desirable to deploy a Bayesian estimate that uses prior knowledge about the occurrence rate, λ to make statements about future hurricane frequency. Because the process is stochastic the predictions are given in terms of probabilities.

The Dataset

The data is retrieved from NHC (2017) nhc and is a collection of tropical cyclone data, contain information on the location, time, maximum winds, central pressure of tropical and subtropical cyclones.

The National Hurricane Center (NHC) conducts a post-storm analysis of tropical cyclones in the Atlantic basin and the North Pacific Ocean to determine the official assessment of the cyclone's history. The analysis performed in this study is solely based on the hurricane data from the Atlantic Ocean between 1851 and 2014.

Initial Data Cleaning

Before performing the analysis, some initial cleaning is conducted to obtain the input data needed for this analysis. To describe the hurricane activity as a Poisson process, the number of hurricanes are counted. To obtain the annual number of hurricanes a subset of the data with *Status* HU (tropical cyclone of hurricane intensity) and distinct *ID* is selected. The year of occurrence is then retrieved from the variable *Date* and finally the number of hurricanes for each year is counted. The result is a total of 890 hurricane occurrences over a time period of 163 years.

The Prior Distribution

Because we want to use prior information to estimate future hurricane frequency, we need to make a decision about the prior distribution. A convenient choice for a prior distribution is a member of the gamma density $\Gamma(\alpha, \beta)$ of the form

$$p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0 \tag{2}$$

where β is the time interval and α is the number of hurricanes that occurred during this interval (Albert, 2009). A convenient source of prior information is historical hurricane records. By assuming that the records from the first 50 years serve as a good indicator of the hurricane frequency, we observe the number of hurricanes h_j in the time interval β ($j = 1, \ldots, \beta$). Since h_j follows a Poisson distribution, then the updated distribution for λ is given by

$$p(\lambda) \propto \lambda^{\sum_{j=1}^{\beta} h_j - 1} exp(-\beta\lambda) \tag{3}$$

Hence we have a gamma prior, $\Gamma(\alpha, \beta)$, for λ, where $\alpha = \sum_{j=1}^{\beta} h_j$. To estimate the prior parameters we use bootstrapping. Bootstrapping is a resampling method that provides an estimate of the variation for the given statistic. By applying the *bootstrap*() function we obtain a bootstrap sample of the mean. Although we cannot say with certainty the true hurricane rate for this period, we can obtain a confidence interval for λ given by the records for the first 50 years. From the 90 % confidence interval constructed, we are 90 % confident that the annual number of hurricanes are within [4.680,5.681]. Based on the confidence interval, we further want to find the best estimate for the number of hurricanes and the length of the time interval. This is done by apply the optimization function *optim*() in R where the objective function is defined as the absolute value of the difference between the gamma and target quantiles. From this we obtain estimates for the gamma parameters, $\alpha = 288.48$ an $\beta = 55.80$.

The Posterior Distribution

We now have the information required to obtain the posterior distribution for λ, the prior parameters $\alpha = 288.48$ an $\beta = 55.80$ and the likelihood statistics based on the data from the reliable time period 1900-2014. The likelihood function is given by the following equation.

$$\ell(\lambda|h) \propto exp(-h\lambda)\lambda^{\sum_{i=1900}^{2014} h_i} (4)$$

Of importance here is the fact that if the observed number of hurricanes h for a given time interval t is Poisson and λ is assigned the gamma prior, $\Gamma(\alpha, \beta)$, then the posterior distribution is also on the gamma form with parameters $\alpha + h$ an $\beta + t$, $\Gamma(\alpha + h, \beta + t)$. Thus the gamma density is the conjugate prior for the Poisson rate λ. The total number of hurricanes from this period and the length of the period is $h = 631$ an $t = 113$. Thus we have that the posterior parameters are $h^* = h + \alpha = 919.48$ and $t^* = t + \beta = 168.80$. By applying Bayes Theorem

we have that the posterior distribution, given the prior and and the Poisson sampling density is given by

$$f(\lambda|h^*, t^*) \propto \lambda^{\alpha+h-1} exp(-(\beta + t)\lambda) \tag{5}$$

Notice that the Bayesian model includes assumptions of the sampling density and the prior density. We investigate the validity of the proposed model by inspecting the predictive density. If the observed data α is consistent with the predictive density, $(\lambda|h^*, t^*)$ the model is appropriate. However, if α is in the tail of the predictive density, it suggest that the Bayesian model is either inappropriate, or that the prior- or sampling density is poorly defined. The prior, likelihood, and posterior are all in the gamma family. We therefore compute the values of the prior, likelihood, and posterior by applying the *dgamma*() function in R.

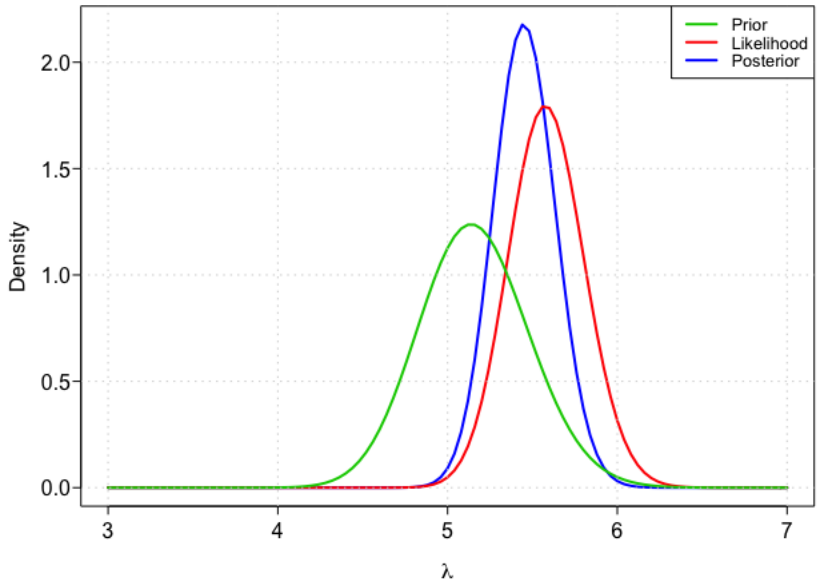

Figure 1: Plot of the likelihood, prior and posterior

From the plot in Figure 1 we observe that posterior resembles the likelihood but is shifted towards the prior. The curve is also narrower, implying a lower variance. We have that the posterior is a weighted average of the prior and the likelihood, where greater the precision imply a higher weight on the prior. The prior density estimate is relatively broad implying low precision. This might be due to the assumption of a Poisson sampling distribution.

Predictive distribution

The information about λ is given by the parameters h^* and t^* and is of gamma density. We now want to use this information to predict the future hurricane activity. Given this data, we have that the posterior predictive density is a negative binomial distribution. Thus, the predictive density for observing $\hat{h}$ hurricanes during the next $\hat{t}$ years is given by

$$p\Big(\hat{h}|h^*, \frac{t^*}{\hat{t}+t^*}\Big) = \frac{\Gamma(\hat{h}+h^*)}{\hat{h}!\Gamma(h^*)}\Big[\frac{t^*}{\hat{t}+t^*}\Big]^{h^*}\Big[\frac{\hat{t}}{\hat{t}+t^*}\Big]^{\hat{h}} \tag{6}$$

where the mean and variance are $\hat{t}\frac{h^*}{t^*}$ and $\hat{t}\frac{h^*}{\hat{t}+t^*}$, respectively.

When you are interested in the probability of a hurricane for the next year, $\hat{t}$ is small compared to t^* and hence it has little influence. However, if you want to predict the hurricane activity for the next 10-30 years, the parameter is of higher importance. We first calculate the predictive posterior probabilities for $h = 1, 2, \ldots, 15$. The probabilities are presented in Table 1.

H	$p(h = H)$
0	0.0043
1	0.0236
2	0.0640
3	0.1159
4	0.1576
5	0.1716
6	0.1559
7	0.1216
8	0.0830
9	0.0505
10	0.0276
11	0.0138
12	0.0063
13	0.0027
14	0.0010
15	0.0004

Table 1: Probability of H number of hurricanes in the next year

In Figure 2 the predictive probabilities are plotted with respect to the number of hurricanes. From the plot we see that the posterior predictive densities estimate a higher occurrence of hurricanes than given by the likelihood. The spread is wide, indicating that there is a lot of uncertainty in the prediction.

We then plot the predictive posterior density and distribution function with respect to the number of hurricanes over the next ten years. The plot is presented in Figure 3.

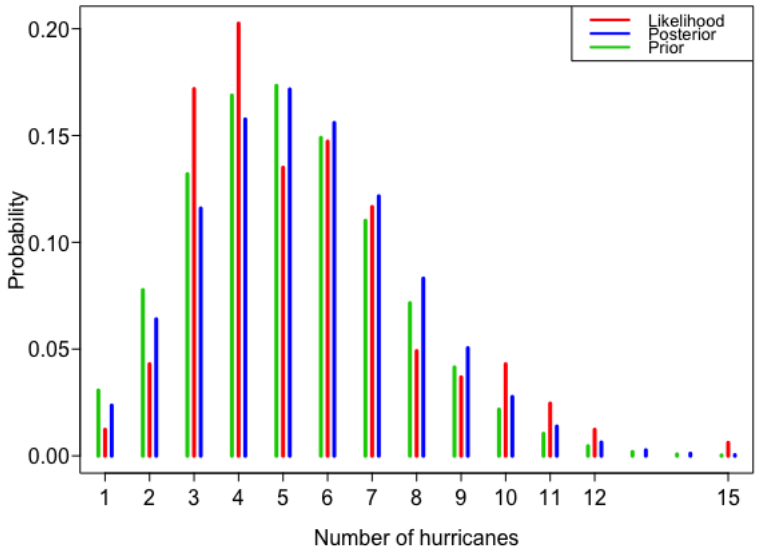

Figure 2: Plot of the likelihood, prior and posterior densities

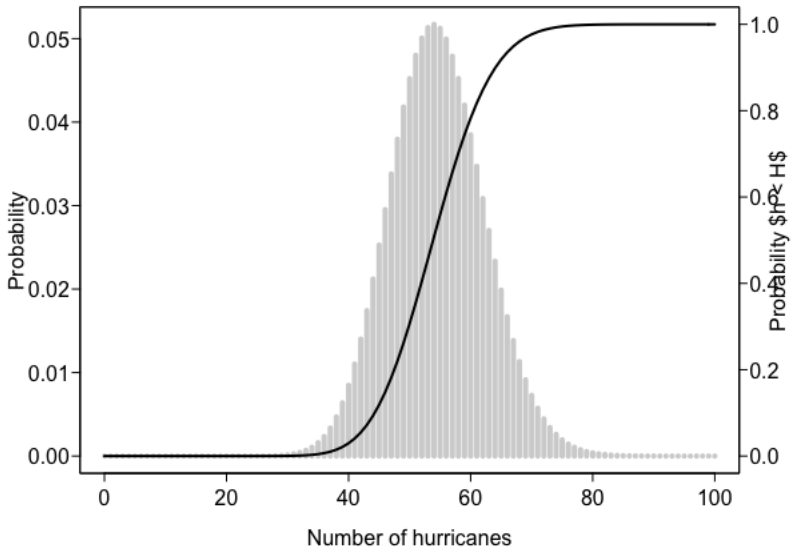

Figure 3: Plot of the predictive posterior density and distribution function with respect to the number of hurricanes over the next ten years

The vertical bars represent the probabilities of H hurricanes in the next ten years with a scale given on the left vertical axis, while the solid line represent the probability that the number of hurricanes will be less than or equal to H with a scale given on the right axis.

We proceed by looking at probability that the number of hurricanes will exceed H for a 10, 20 and 30 years period. From Figure 4 we see that the expected number of hurricanes over the next 30 years is about 170.

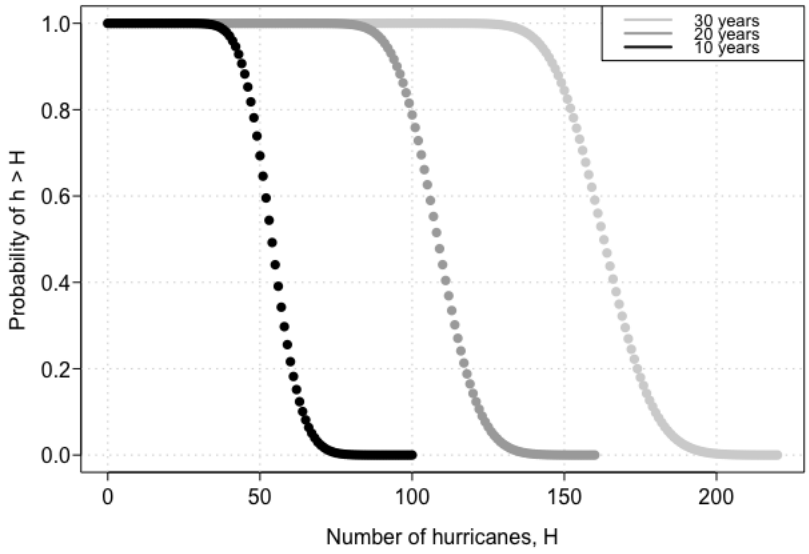

Figure 4: Plot of the predictive posterior probability $p(h > H)$ to a period of 10, 20 and 30 years

The analysis conduction is a good basis for incorporating all previous information about hurricane activity in order to predict future hurricane occurrences. The approach could be further developed by discounting the information over the time period, giving a higher weight to more recent data, than records from earlier decades.

References

Albert, J. (2009). *Bayesian computation with R*, Springer Science & Business Media.

NHC, N. H. C. (2017). Hurricanes and typhoons, 1851-2014. `https://www.kaggle.com/noaa/hurricane-database`.