P8160 Group Project 3: Bayesian Modeling of Hurricane Trajectories

Anyu Zhu | Haotian Wu | Wenhan Bao | Yiming Li | Qihang Wu

5/6/2022

Objective

In this project, we firstly built a Bayesian model based on the track data of 703 hurricanes in the North Atlantic area since 1950. A Markov Chain Monte Carlo (MCMC) algorithm was designed to generate the distribution of parameters. With the start time and type of each hurricane, the estimated coefficients from the Bayesian model were used to explore the seasonal differences and wind speed changes over years. Finally, we also explored the characteristics of hurricanes associated with the damage and deaths.

Background

Hurricanes are large rotating tropical storms with winds in excess of 119 kilometers per hour (74 mph). They usually form between June 1 and November 30 in the Atlantic Ocean but can develop in other oceans as well. They are known as typhoons in the western Pacific and cyclones in the Indian Ocean[1].

When a hurricane approaches land, tremendous damage can occur to the nearby cities. Therefore, scientists continue to improve their ability to forecast hurricanes. The sooner they can access accurate information about a hurricane's location and intensity, the better the chances to minimize the hurricane's impacts.

Data Descriptions and Preprocessing

The first data `hurrican703.csv` collected the track data of 703 hurricanes in the North Atlantic area since 1950. For all the hurricanes, their locations (longitude & latitude) and maximum wind speed were recorded every 6 hours. The variables include `ID`, `Season`, `Month`, `Nature`, `time`, `Latitude`, `Longitude`, and `Wind.kt`. Main data processing steps for this data are listed as follows:
1) Created 3 new variables including the changes of latitude and longitude, as well as the wind speed between the time t and $t-6$;
2) Removed 9 hurricanes with observations less than 5 to ensure the data partition.
Finally, we have totally **691** hurricanes in the updated dataset.

The second data `hurricanoutcome2.csv` recorded the damages and death caused by 46 hurricanes in the U.S, and some features extracted from the hurricane records. To explore the characteristics related with death and damage, we combined this data with the coefficients results obtained from the first model by the hurricane ID. For this dataset, we converted the start month for each hurricane into the variable `season`, which includes Spring, Summer, Fall, and Winter

Statistical Methods

Likelihood

For each hurricane i and k_i's time points, we have the following Bayesian model:

$$Y_i(t+6) = \beta_{0i} + \beta_{1i}Y_i(t) + \beta_{2i}\Delta_{i1}(t) + \beta_{3i}\Delta_{i2}(t) + \beta_{4i}\Delta_{i3}(t) + \varepsilon_i(t),$$

where $Y_i(t)$ is the wind speed at time t, Δ_{i1}, Δ_{i2}, and Δ_{i3} are the changes of latitude, longitude, and the wind speed between time point t and $t-6$, respectively. $\varepsilon_i(t)$ follows a normal distributions with mean zero and variance σ^2. The above Bayesian model can be simplified as:

$$Y_i(t+6) = x_i(t) + \varepsilon_i(t),$$

where $\beta_i = (\beta_{0i}, \beta_{1i}, ..., \beta_{4i}) \sim N\left(0, \sigma^2\right)$. Based on the property of the multivariate linear regression model, for each hurricane i, we have:

$$Y_i \mid X_i \sim N_{k_i}\left(x_i\beta_i, \sigma^2 I_{k_i}\right),$$

where I_{k_i} is an identity matrix with k_i dimensions.

Thus, we can consider the following distribution of each hurricane i:

$$f\left(y_i \mid \beta_i, \sigma^2\right) = \left[(2\pi)^{k_i} \cdot \det\left(\sigma^2 I_{k_i}\right)\right]^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}\left(y_i - x_i\beta_i\right)^\top \left(\sigma^2 I_{k_i}\right)^{-1}\left(y_i - x_i\beta_i\right)\right\}$$

.

From above, we derive the following likelihood function:

$$\begin{aligned} f\left(y \mid B, \sigma^2\right) &= \prod_{i=1}^{n} f\left(y_i \mid \beta_i, \sigma^2\right) \\ &= \prod_{i=1}^{n}\left(\left[(2\pi)^{k_i} \cdot \det\left(\sigma^2 I_{ki}\right)\right]^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}\left(y_i - x_i\beta_i\right)^\top \left(\sigma^2 I_{k_i}\right)^{-1}\left(y_i - x_i\beta_i\right)\right\}\right) \end{aligned}$$

Prior distributions

We assume the following non-informative prior distributions:

$$\beta_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{4i}) \sim N_5(\mu, \Sigma),$$

where $B = \left(\beta_1^\top, \beta_2^\top, \ldots, \beta_n^\top\right)^\top$ and n is the number of hurricanes. So,

$$\pi(B|\mu, \Sigma^{-1}) = \prod_{i=1}^{n} f\left(\beta_i\right) \propto \det(\Sigma)^{-n/2} \cdot \exp\left\{-\frac{1}{2}\sum_i\left[\left(\beta_i - \mu\right)^\top (\Sigma)^{-1}\left(\beta_i - \mu\right)\right]\right\}.$$

Also, $\pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$; $\pi(\mu) \propto 1$; $\pi\left(\Sigma^{-1}\right) \propto |\Sigma|^{-(d+1)} \cdot \exp\left(-\frac{1}{2}\Sigma^{-1}\right)$.

Conditional posteriors

The posterior distribution is the product of the likelihood and the prior:

$$g\left(B, \sigma^2, \mu, \Sigma^{-1} \mid y\right) \propto f\left(y \mid B, \sigma^2\right) \cdot \pi(B \mid \mu, \Sigma^{-1}) \cdot \pi\left(\sigma^2\right) \cdot \pi(\mu) \cdot \pi\left(\Sigma^{-1}\right),$$

so we have:

$$\begin{aligned} \pi\left(\sigma^2 \mid \cdot\right) &\propto \prod_{i=1}^{n} \det\left(\sigma^2 I_{k_i}\right)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}\Sigma_i\left[\left(y_i - x_i\beta_i\right)^\top \left(\sigma^2 I_{ki}\right)^{-1}\left(y_i - x_i\beta_i\right)\right]\right\} \cdot \sigma^{-2} \\ &= \left(\sigma^2\right)^{-\frac{1}{2}\Sigma_i k_i} \cdot \exp\left\{-\frac{1}{2\sigma^2}\Sigma_i\left[\left(y_i - x_i\beta_i\right)^\top \left(y_i - x_i\beta_i\right)\right]\right\} \cdot \sigma^{-2} \\ &= \left(\sigma^2\right)^{-1-\frac{1}{2}\Sigma_i k_i} \cdot \exp\left\{-\frac{1}{2\sigma^2}\Sigma_i\Sigma_{ti}\left(y_{i,t} - x_{i,t}\beta_i\right)^2\right\} \end{aligned}$$

Therefore, $\sigma^2 \sim$ Inverse Gamma $\left(\frac{1}{2}\Sigma_i k_i, \frac{1}{2}\Sigma_i\Sigma_{t_i}\left(y_{i,t} - x_{i,t}\beta_i\right)^2\right)$.

$$\begin{aligned}
\pi(\Sigma^{-1} \mid \cdot) &\propto \det(\Sigma)^{-n/2} \cdot \exp\left\{-\frac{1}{2}\Sigma_i \left(\beta_i - \mu\right)^\top \Sigma^{-1} \left(\beta_i - \mu\right)\right\} \cdot \det(\Sigma)^{-(d+1)} \cdot \exp\left\{-\frac{1}{2}\Sigma^{-1}\right\} \\
&= \det(\Sigma)^{-(n/2+d+1)} \cdot \exp\left\{-\frac{1}{2}\left[\Sigma^{-1} + \Sigma_i \left(\beta_i - \mu\right)^\top \Sigma^{-1} \left(\beta_i - \mu\right)\right]\right\} \\
&\propto \det(\Sigma^{-1})^{(n+2d+2)/2} \cdot \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1} \cdot \left(I + \Sigma_i \left(\beta_i - \mu\right)\left(\beta_i - \mu\right)^\top\right]\right\} \\
&\propto \det(\Sigma^{-1})^{(n+3d+3-d-1)/2} \cdot \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1} \cdot \left(I + \Sigma_i \left(\beta_i - \mu\right)\left(\beta_i - \mu\right)^\top\right]\right\}
\end{aligned}$$

Thus $\Sigma^{-1} \sim \text{Wishart}\left(n + 3d + 3, [I + \Sigma_i \left(\beta_i - \mu\right)\left(\beta_i - \mu\right)^\top\right]^{-1}$), that is:

$$\Sigma \sim \text{Inverse Wishart}\left(n + 3d + 3, I + \Sigma_i \left(\beta_i - \mu\right)\left(\beta_i - \mu\right)^\top\right)$$

$$\begin{aligned}
\pi(\mu \mid \cdot) &\propto \exp\left\{-\frac{1}{2}\Sigma_i \left[\left(\beta_i - \mu\right)^\top \Sigma^{-1} \left(\beta_i - \mu\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\Sigma_i \left(\beta_i^\top \Sigma^{-1} \beta_i + \mu^\top \Sigma^{-1} \mu - 2\beta_i^\top \Sigma^{-1} \mu\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\Sigma_i \beta_i^\top \Sigma^{-1} \beta_i + \mu^\top n\Sigma^{-1} \mu - 2\Sigma_i \beta_i^\top \Sigma^{-1} \mu\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mu^\top n\Sigma^{-1} \mu - 2\Sigma_i \beta_i^\top \Sigma^{-1} \mu + \Sigma_i \beta_i^\top \Sigma^{-1} \beta_i\right)\right\} \\
&= \exp\left\{-\frac{1}{2}(\mu^\top \underbrace{n\Sigma^{-1}}_{M} \mu - 2\mu^\top \underbrace{\Sigma_i \Sigma^{-1} \beta_i}_{N} + \Sigma_i \beta_i^\top \Sigma^{-1} \beta_i)\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\left(\mu - M^{-1}N\right)^\top M \left(\mu - M^{-1}N\right)\right]\right\}.
\end{aligned}$$

Therefore, $\mu \sim MVN\left(M^{-1}N, M^{-1}\right)$.

$$\begin{aligned}
&\pi(B \mid \cdot) \propto \exp\left\{-\frac{1}{2}\Sigma_i \left[\left(y_i - x_i\beta_i\right)^\top \left(\sigma^2 I_{ki}\right)^{-1} \left(y_i - x_i\beta_i\right)\right]\right\} \cdot \exp\left\{-\frac{1}{2}\Sigma_i \left[\left(\beta_i - \mu\right)^\top (\Sigma)^{-1} \left(\beta_i - \mu\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\Sigma_i \left[\left(y_i - x_i\beta_i\right)^\top \left(\sigma^2 I_{ki}\right)^{-1} \left(y_i - x_i\beta_i\right) + \left(\beta_i - \mu\right)^\top \Sigma^{-1} \left(\beta_i - \mu\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\Sigma_i [y_i^\top \sigma^{-2} I_{ki} y_i + \beta_i^\top x_i^\top \sigma^{-2} I_{k_i} x_i \beta_i - 2y_i^\top \sigma^{-2} I_{ki} x_i \beta_i + \beta_i^\top \Sigma^{-1} \beta_i + \mu^\top \Sigma^{-1} \mu - 2\mu^\top \Sigma^{-1} \beta_i]\right\} \\
&= \exp\left\{-\frac{1}{2}\Sigma_i [y_i^\top \sigma^{-2} I_{ki} y_i + \mu^\top \Sigma^{-1} \mu + \beta_i^\top \left(\Sigma^{-1} + x_i^\top \sigma^{-2} I_{ki} x_i\right) \beta_i - 2\left(y_i^\top \sigma^{-2} I_{ki} x_i + \mu^\top \Sigma^{-1}\right) \beta_i]\right\}
\end{aligned}$$

We can define the following terms:

$$\begin{aligned}
R &= y_i^\top \sigma^{-2} I_{k_i} y_i + \mu^\top \Sigma^{-1} \mu \\
V &= \Sigma^{-1} + x_i^\top \sigma^{-2} I_{ki} x_i \\
M &= \sigma^{-2} x_i^\top y_i + \Sigma^{-1} \mu
\end{aligned}$$

Thus, $\pi(B \mid \cdot) \propto \left(\beta_i - V^{-1}M\right)^\top V \left(\beta_i - V^{-1}M\right) \sim MVN\left(V^{-1}M, V^{-1}\right)$

Gibbs Sampling

Since directly generating the above jointly density is rather complicated, we can implement the gibbs sampler to sample each variable in turn. Based on the previous conditional distributions, we updated the four parameters including B, σ^2, $\sum$, and μ in sequence. Finally, we selected 10,000 iterations. Note that we ignored the first 8,000 iterations to ensure the stationary distribution of the Markov chain was reached.

Regression Models

To explore seasonal and annual difference in the wind speed, we took the β coefficients from the Bayesian model as the response and built the following linear regression model:

$$\beta_j \sim Season + Year + Nature, j = 0, ..., 4,$$

where season and nature are categorical, and year is continuous variable.

To predict the hurricane-induced damage, we incorporated the coefficients from the Bayesian model and the new predictors in the second data, selected some featured variables by lasso, and finally fit a regression model. The variables β_0, Year, max speed, total affected population, and the affected population reside in the United States were selected in the model.

$$Damage \sim \beta_0 + Year + Maxspeed + Total.Pop + Percent.USA$$

We built a poisson regression model to evaluate characteristics related to deaths.

$$Death \sim \beta_4 + MonthSummer + Maxspeed + Maxpressure + Meanpressure + Hours + Total.Pop + Percent.Poor$$

Results

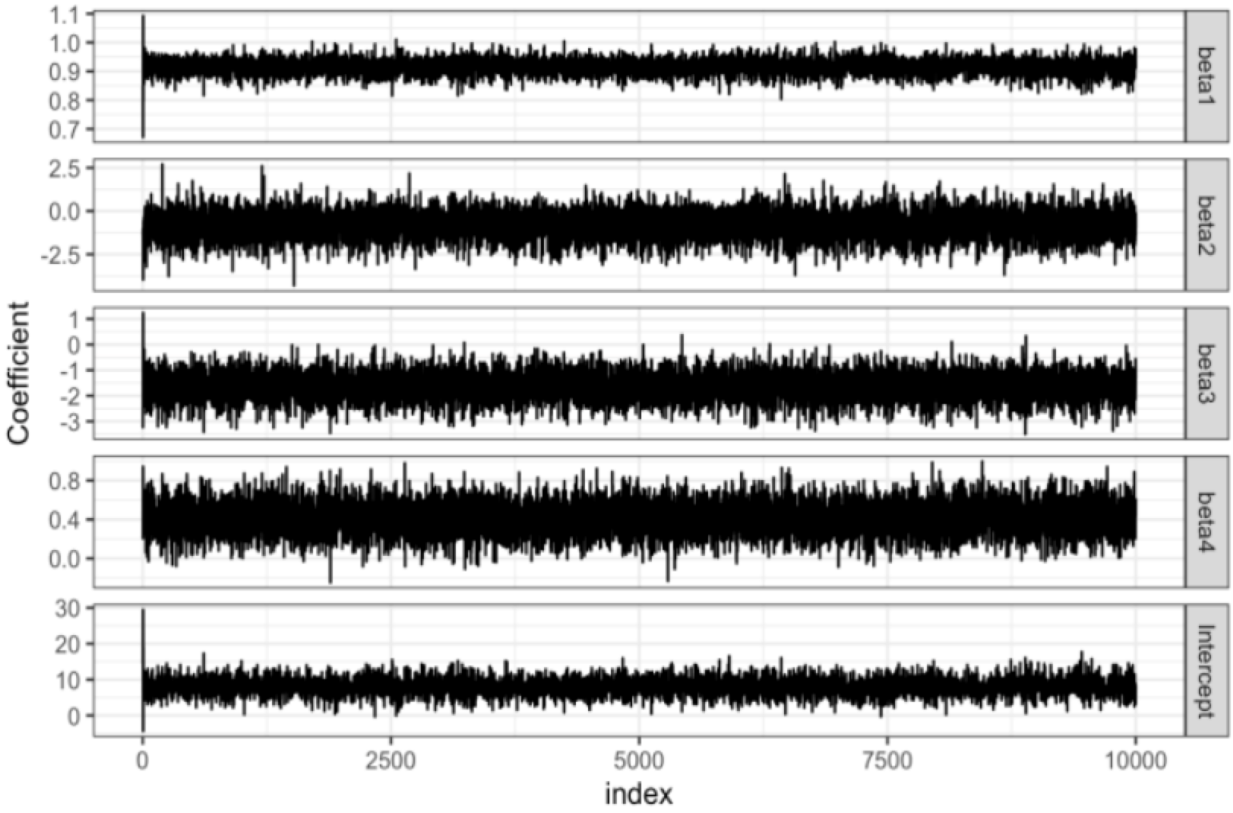
MCMC-Gibbs Sampling

The initial values for μ, Σ^{-1} and σ^2 we set are (50,0,0,0,0), 0.5 and `diag(0.5, 0.5, 0.5)` respectively. In the case of Gibbs results, we set 10000 iterations and record the $\beta_{i,j}^n$ in each iteration with a linear model (from the least square method) to demonstrate how similar the coefficients are for each hurricane's train data and μ^n.

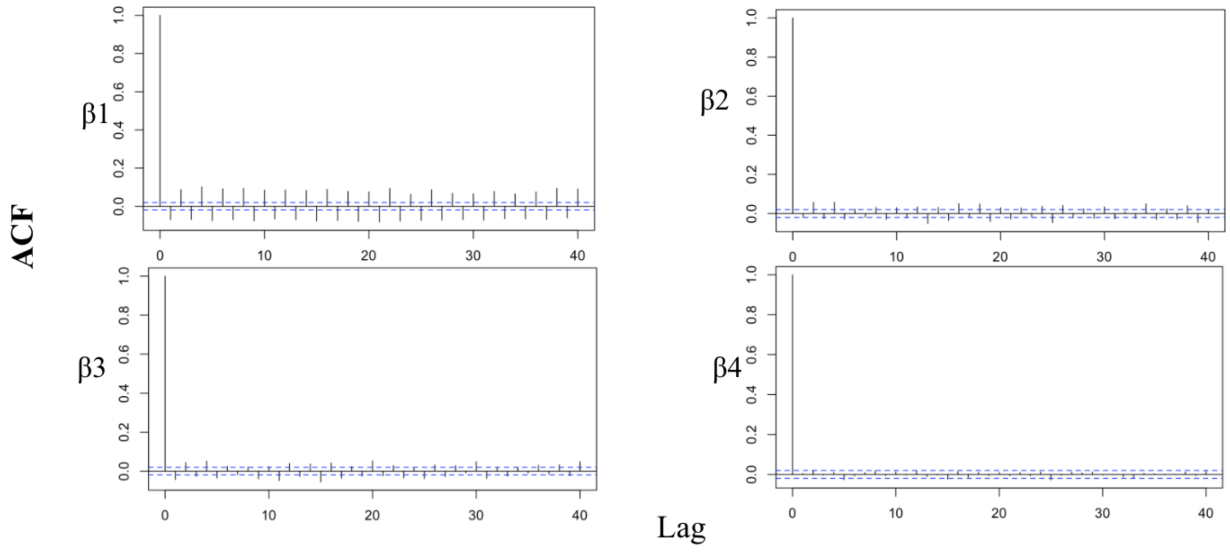
To minimize the variance generated randomly by Gibbs sampling and avoid the non convergent situation at the beginning of the iteration, we only take the coefficients value from 8000 to 10000 times where we take the mean value for each coefficients $\beta_{i,j}$ as the final results. To evaluate how well the estimated Bayesian model tracks the individual hurricanes, we divided the observations within each hurricane ID by 8:2 randomly to create a training and testing set. Then we calculated RMSE for each different hurricane.

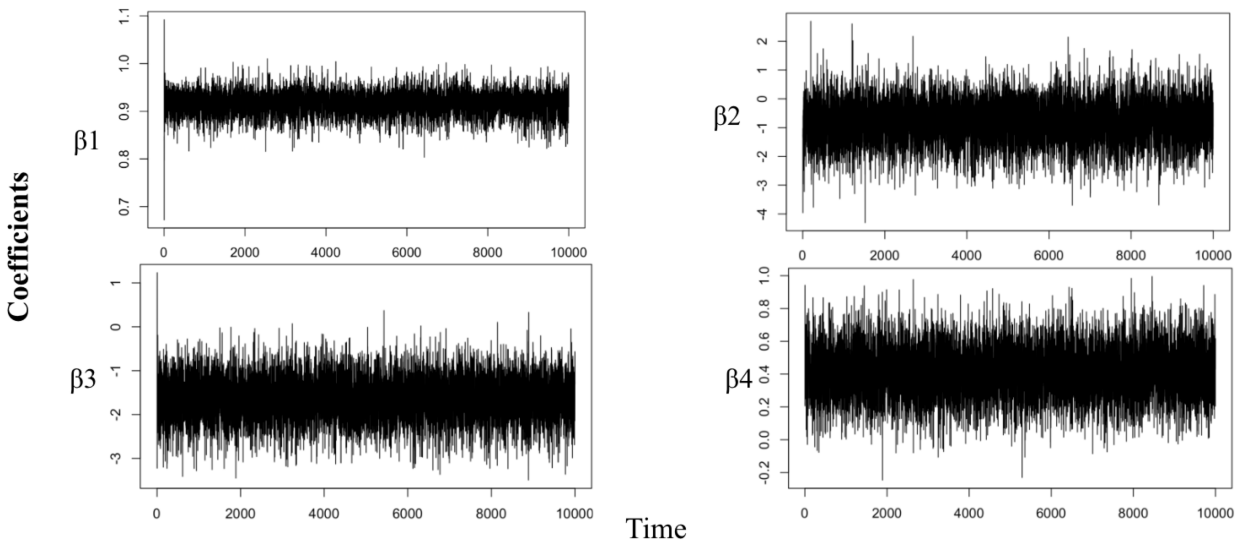
For our final results obtained from Gibbs sampling, you could jump to our github(Column2: RMSE for the test dataset, Column3-7: Mean coefficients $\beta_{i,j}$ for each hurricane by Gibbs sampling, Column8-12: coefficients $\beta_{i,j}$ for each hurricane by lm() function, Column13-17: Mean μ for B)

In order to further elucidate the results from our Gibbs methods, we take `ABLE.1950` as an example. Firstly, we trace the Gibbs step for only extracting the $\beta_{1,j}^n$ from B^n by plotting the value.



Then to look deeply into how the coefficients change with the increasing of iteration times, we also plot the autocorrelation plot(ACF) and time series plot (TS) where we could tell that the correlation for coefficient values between zero and blue line are not statistically significant.





For the case of `ABLE.1950`, the RMSE for the remaining 20% test dataset is 4.25 and the comparison of coefficients $\beta_{1,j}$ is in the table below.

Table 1: Compare the Coefficients from Gibbs Sampling and lm function

model	Intercept	Beta1	Beta2	Beta3	Beta4
Gibbs sampling	7.601	0.919	-0.788	-1.644	0.411
Weibull	8.833	0.912	-1.394	-2.013	2.983

Task 5

Task 6

Discussion and Limitations

Contributions

Reference

[1] https://www.whoi.edu/know-your-ocean/ocean-topics/ocean-human-lives/natural-disasters/hurricanes/?gclid=Cj0KCQjwsdiTBhD5ARIsAIpW8CIO0bvwlo16Pisxe_WV7owA4KdbMfpAhMxqaz8Ir3fsQyfiHPa2MAaAlmZEALw_wcB

Appendix

For codes please click *here*