P8160 - Baysian modeling of hurrican trajectories

Jiahao Fan, Tianwei Zhao, Yijin Wang, Youlan Shen, Yujia Li

2023-05-02

Contents

Background and Objectives

As hurricanes affect people, the ability to forecast hurricanes is essential for minimizing the risks in suffered areas. Hereby, a hierarchical Bayesian strategy for modeling North Atlantic hurricane counts since 1950 is illustrated. Model integration would be expected to achieve through a Markov chain Monte Carlo algorithm. Contingent on the availability of values for the covariates, the models can be used to explore the seasonality among track data and make predictive inferences on hurricanes.

The data given has 703 observations and following features are recorded for each hurricanes in the North Atlantic:

- ID: ID of the hurricanes

- Season: In which the hurricane occurred

- Month: In which the hurricane occurred

- Nature: Nature of the hurricane (ET-Extra tropical;DS-Disturbance;NR-Not rated;SS-Sub tropical; TS-Tropical storm)

- time: dates and time of the record

- Latitude and Longitude: The location of a hurricane check point

- Wind.kt: Maximum wind speed (in Knot) at each check point

Methods

Baysian Model

MCMC Algorithm

Given the fact that we cannot directly sample from the joint posterior distribution calculated above. We choose to use Gibbs sampling. Gibbs sampling is a sequential sampler that samples from the conditional distribution of each parameter so that the final collection of samples forms a Markov chain whose stationary distribution is the joint distribution that we are interested in.

In our case, conditional posteriors are more manageable than the joint posterior. Conditional posteriors for each parameter are calculated as the following:

!!! ADD CONDITONAL POSTERIOR

Still, we could not find a regular distribution that has the density function that matches with the conditional posterior of σ^2. Therefore, we incorporated Metropolis Hasting Random Walk into Gibbs sampling. Metropolis Hasting Random Walk finds approximated values of σ^2 by accepting and rejecting proposed value with an acceptance probability that helps maintain the detailed balance condition. In our algorithm, we work with log-scaled posterior values instead of the original-scaled posterior because $\frac{1}{\sigma^N}$ dominates the entire posterior value where N is the total number of observations in the data set.

In our investigation of hurricane wind speed, we used Gibbs sampling and Metropolis Hasting Random Walk as the following:

Algorithm 1 Gibbs Sampling

1: **function** Gibbs(starting_values, V,S,X,D,Y, nitr,ν,H, hurricane_idx)
2:     create posts results lists to collect results for B, μ, A, γ, σ^2
3:     $B^1 \leftarrow$ f(B|start_μ, start_σ^2, start_A, start_γ, Y)
4:     $\mu^1 \leftarrow$ f(μ|B^1, start_σ^2, start_A, start_γ, Y)
5:     $A^1 \leftarrow$ f(A|B^1, μ^1, start_σ^2, start_γ, Y)
6:     $\gamma^1 \leftarrow$ f(γ|B^1, μ^1, start_σ^2, A^1, Y)
7:     σ^2_chain $\leftarrow$ metropolis_hasting_sigma_squared(start_σ^2, iter = 500, a = 0.8,Y,D,B^1,X,γ^1)
8:     $(\sigma^2)^1 \leftarrow$ mean(σ^2_chain[401:500])
9:     **for** i in 2:nitr **do**
10:         $B^i \leftarrow$ f(B|μ^{i-1}, $(\sigma^2)^{i-1}$, A^{i-1}, γ^{i-1}, Y)
11:         $\mu^i \leftarrow$ f(μ|B^i, $(\sigma^2)^{i-1}$, A^{i-1}, γ^{i-1}, Y)
12:         $A^i \leftarrow$ f(A|B^i, μ^i, $(\sigma^2)^{i-1}$, γ^{i-1}, Y)
13:         $\gamma^i \leftarrow$ f(γ|B^i, μ^i, $(\sigma^2)^{i-1}$, A^i, Y)
14:         σ^2_chain $\leftarrow$ metropolis_hasting_sigma_squared($(\sigma^2)^{i-1}$, iter = 500, a = 0.8,Y,D,B^i,X,γ^i)
15:         $(\sigma^2)^i \leftarrow$ mean(σ^2_chain[401:500])
16:     **end for**
17:     **return** posts results lists that collect results for B, μ, A, γ, σ^2
18: **end function**

Figure 1: Gibbs Sampling

Algorithm 2 Metropolis Hasting sigma squared

1: **function** METROPOLIS HASTING SIGMA SQUARED(start_sigma_squared, nitr, a, B, X,D,Y, H, hurricane_idx)
2:     Initialize chain
3:     $chain[1] \leftarrow$ start_sigma_squared
4:     **for** i in 2:itr **do**
5:         $curr \leftarrow chain[i-1]$
6:         $prop \leftarrow$ a sample from Unif(curr-a, curr+a)
7:         prop_posterior $\leftarrow$ sigma_squared_log_posterior(prop,...)
8:         curr_posterior $\leftarrow$ sigma_squared_log_posterior(curr,...)
9:         posterior_ratio $\leftarrow$ exp(prop_posterior - curr_posterior)
10:         $\alpha \leftarrow$ min(posterior_ratio, 1)   ▷ acceptance probability
11:         $u \leftarrow runif(1)$
12:         **if** u $< \alpha$ **then**
13:             $chain[i] \leftarrow prop$ ▷ accept proposal
14:         **else**
15:             $chain[i] \leftarrow curr$ ▷ reject proposal
16:         **end if**
17:     **end for**
18:     **return** $chain$
19: **end function**

Figure 2: Metropolis Hasting Random Walk

We decide starting values of each parameters by fitting a linear mixed effects model that is the same as the proposed model in a frequentist way and then extract its corresponding coefficients. Starting values for γ are the fixed effects coefficients in the linear mixed effects model.

Table 1: Starting Values for Gibbs Sampling

Terms	Starting Value
γ	Fixed effect coefficients in the fitted Linear Mixed Effects Model
$\beta_{i_1,\ldots,i_n}$	Random effect coefficients in the fitted Linear Mixed Effects Model
σ^2	Residual Variance in the fitted Linear Mixed Effects Model
β_0	Sum of random effect and fixed effect intercepts in the fitted Linear Mixed Effects Model
μ	Mean of the $\beta_{i_1,\ldots,i_n}$ for each predictor
$A = \Sigma^{-1}$	Mean of the Wishart distribution with $\nu = 5$ and scale matrix with 0.7 on the main diagnoal and 0.2 on off-diagnoal

With the stating values above, we assume that the true σ^2 will be around 27 and tried a few values for window size a. Finally, with $a = 0.8$, σ^2 samples converge pretty quickly and the acceptance rate is around 46% with 500 iterations. Therefore, we set $a = 0.8$ in each iteration of Gibbs Sampling and burn the first 400 samples of σ^2.

Although our algorithm is not the most efficient algorithm, it has two main advantages: 1) It is simple to compute and does not rely on the calculation of gradient of log densities as Hamiltonian Monte Carlo; 2) By incorporating the Random Walk, we have the flexibility to control the range of sampling area when the conditional distribution of parameter of interest is difficult to simulate from and when our initial prior is not informative .

MCMC Results

We ran our algorithm with these starting values for 6000 iterations and decided to burn the first 3000 samples for each parameters for analysis in the rest of this project. We use diagnostic plots to evaluate the quality of these samples. We created trace plots for all 6000 samples for each parameter, and autocorrelation plots and histogram plots for the second half of the samples after burn-in.

Results

Seasonal Analysis

Prediction

Accurately predicting hurricane wind speeds is crucial in hurricane forecasting. To achieve this, we calculated the posterior means of the parameters by averaging the last three thousand values in each MCMC chain, and used these means to predict wind speeds at each time point.

To assess the performance of our Bayesian model, we used the Root Mean Square Error (RMSE) of the predicted wind speeds compared to actual wind speeds for several hurricanes. In the table on the left, we list the 10 hurricanes with the least RMSE, with the first hurricane having only six observations but still achieving a small weighted average error of 1.095. This result indicates that our model can make accurate predictions even for hurricanes with fewer observations.

Hurricane ID	Number of Observation	Prediction RMSE
SIXTEEN.2008	6	1.095
FABIAN.1991	9	1.245
FOUR:UNNAMED.1988	5	1.297
TEN.2005	19	1.420
LORENZO.2001	14	1.471
FIFTEEN:UNNAMED.1988	4	1.857
FABIAN.1997	15	1.933
ANDREW.1986	14	1.947
CINDY.1987	20	1.984
GUSTAV.1996	27	2.000

Figure 3: Prediction Standard Error of Wind Speed in Ascending Order

We further analyzed the model's performance on hurricanes with more than 50 observations and those with fewer than 50 observations, labeled 'more observations' and 'fewer observations', respectively, as shown in the graph on the right. The average RMSE for both categories is around 5, suggesting that our model's accuracy is not significantly affected by the amount of data available. This finding is reassuring and indicates that our model is effective in predicting wind speeds for hurricanes with varying amounts of data.
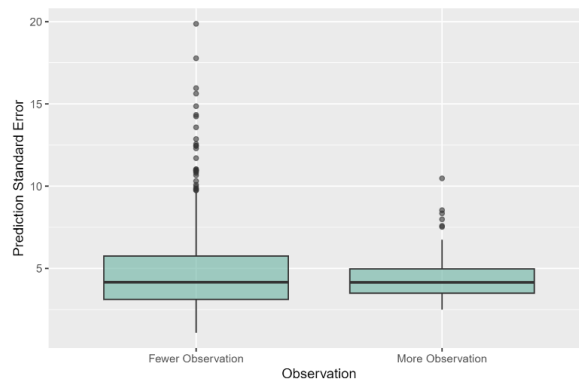

Figure 4: Prediction Standard Error of Wind Speed by Number of Observation

Additional plots are presented below to further investigate the model's performance. The graph displaying the average prediction points by month shows that our model performs well across different months, except for January, which has limited data points. When we evaluate our predictions by year or by nature, we see consistent performance across different years and types of hurricanes.
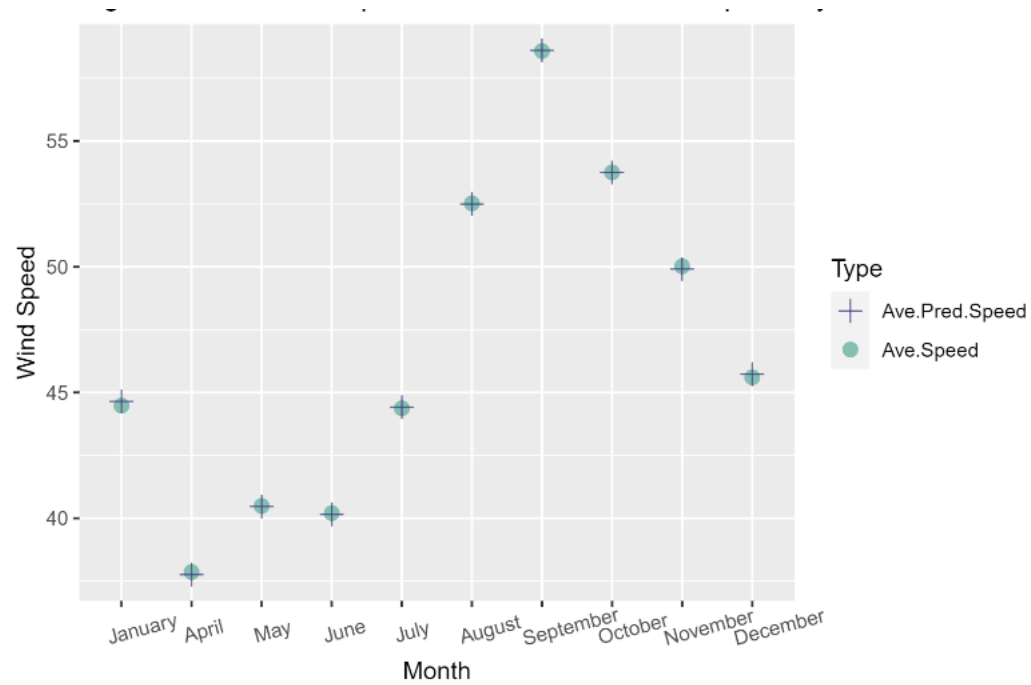
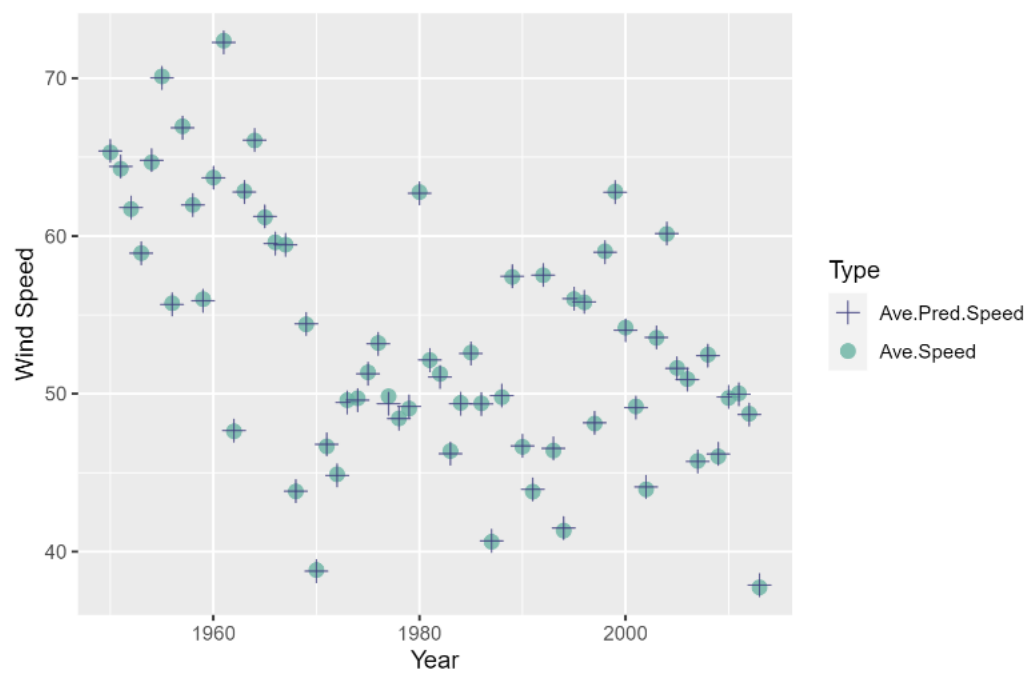Figure 5: Actual Wind Speed vs.Predicted Wind Speed by Month

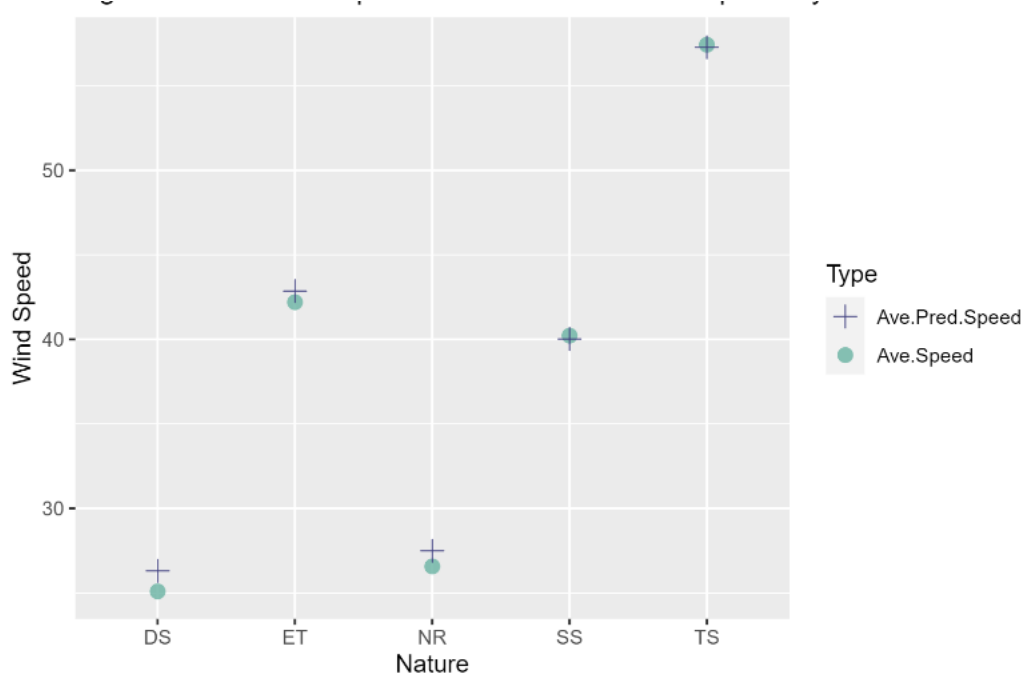Figure 6: Actual Wind Speed vs.Predicted Wind Speed by Year

Figure 7: Actual Wind Speed vs.Predicted Wind Speed by Nature

In conclusion, our analysis using RMSE and additional graphs demonstrates that our Bayesian model is capable of making accurate predictions for hurricane wind speeds. These results suggest that our model is effective in predicting wind speeds for hurricanes with varying amounts of data, and performs consistently across different months, years, and types of hurricanes. These findings have significant implications for the field of hurricane prediction and can inform future research in this area.

Discussion

In summary, our application of the MCMC algorithm to the hurricane Bayesian model has yielded valuable insights into the prediction of hurricane wind speeds. Our analysis revealed significant seasonal differences in wind speeds and did not find any evidence to support the claim of increasing wind speeds over the years. These findings emphasize the importance of accounting for seasonal differences when modeling hurricane wind speeds. Although our final model did not achieve perfect convergence for every parameter, we were still able to produce reliable predictions. The use of MCMC has allowed us to overcome the challenges posed by the complicated distribution and high-dimensional space of hurricane wind speed data, demonstrating the usefulness of this algorithm in Bayesian analysis.

MCMC is a powerful tool in Bayesian statistics that is particularly useful for handling complicated distributions in high-dimensional spaces. One of its main advantages is its ability to explore the parameter space efficiently and estimate the posterior distribution of model parameters. This is particularly helpful when direct sampling is practically impossible in high-dimensional spaces.

Despite its many benefits, MCMC has some drawbacks that need to be addressed. One major issue is the assumption of prior distribution, which can introduce bias and affect the estimation of the posterior distribution. Additionally, the method cannot guarantee convergence due to the correlation of samples, which can lead to errors in the estimation. To address these issues, the method can be improved by changing the sampler, increasing the number of samples, or trying different parameterizations of the model. These solutions can lead to more accurate and efficient estimation of the posterior distribution.

Group Contributions

Appendices

Figures and Tables