CMPT 318: Second Group Assignment

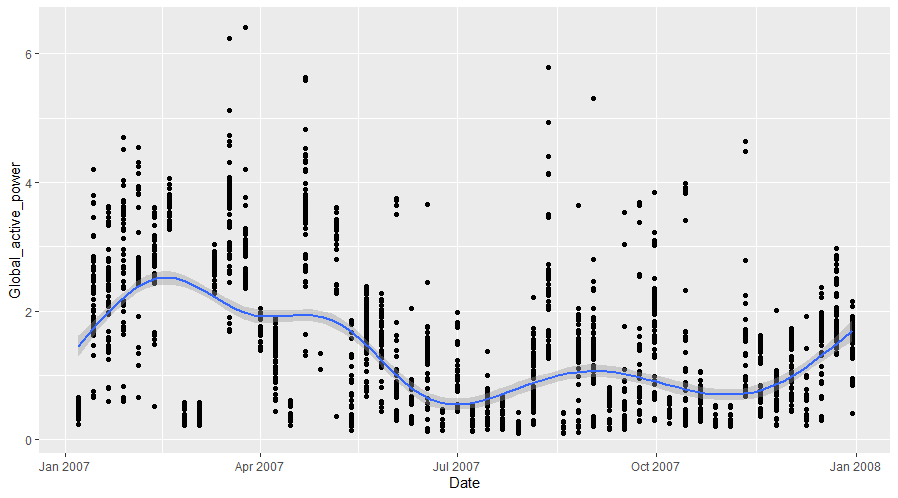
**Group 7:**Wei Yao  
Neil Mukesh Shah  
Karan Sachdeva  
Harry Preet Singh  
Praneer Shrestha

# Task 1

**Time interval chosen:** 09:00 to 10:00

Our team decided to choose this time window because there was consistent fluctuation in the data. We believed that the constant ups and downs as seen in the plot of our graph below, would be a cleaner choice to build the Hidden Markov Models (HMM).

*Figure 1  
Graph plotting the 52 Sundays in the year vs. the Global\_active\_power*



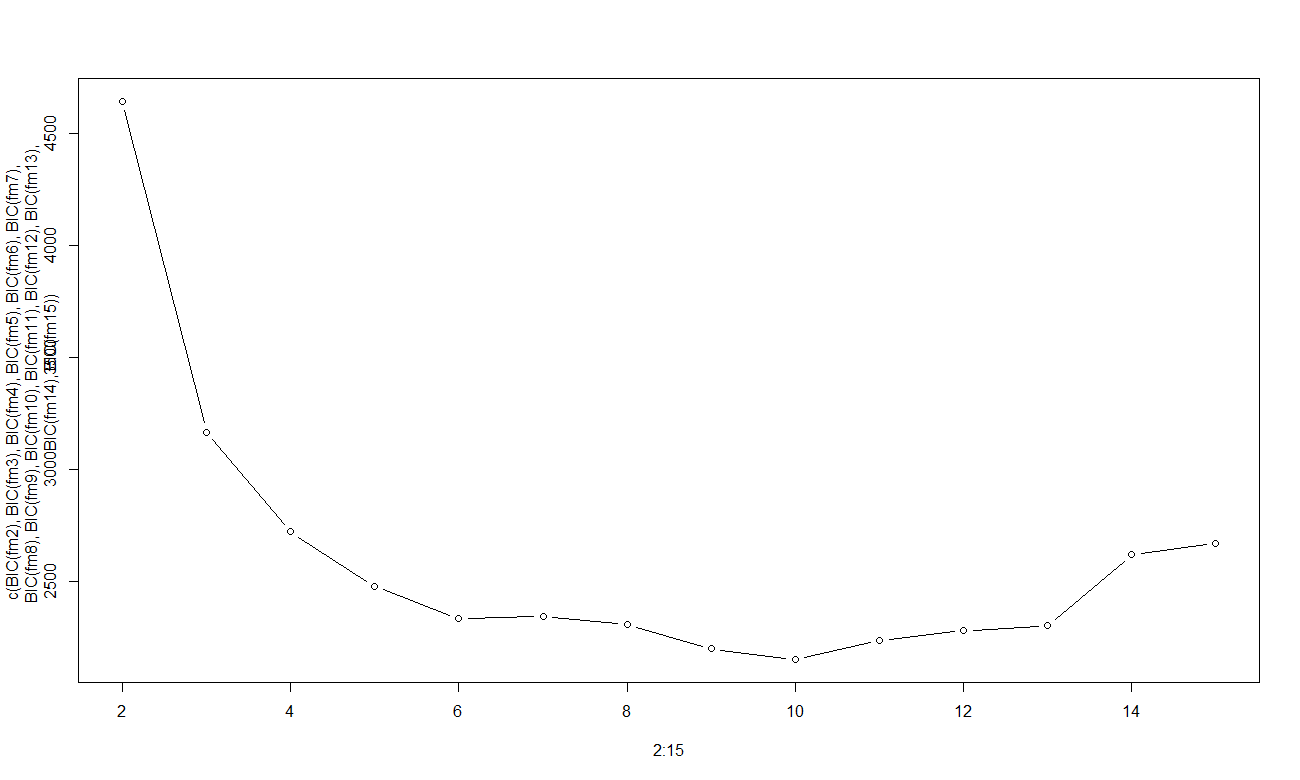
To provide some explanation, each vertical separation represents one Sunday starting from the 07/01/2007 to 30/12/2007 totaling 52 Sundays. And each point along the virtual vertical line represents the Global Active Power at a certain time in the format HH:MM:SS.

**Feature chosen:** Global Active Power

To train the HMM, we decided on focusing in one feature, the Global Active Power to represent the household electricity consumption. This is because multiple features increased the complexity in building HMM. In addition, when analyzing the dataset, our team noticed that other certain rows in all columns except Global Active Power, contained NA values. Although R provides functions to omit these values, we realized that building the HMM would still be much more difficult since our *ntimes* parameter in the *depmix* function would have to account for the omissions. Hence, our decision to focus in the Global Active Power.

As seen in our R code, we trained 15 HMMs by changing the number of states since we decided to focus in a univariate feature. There were two particular areas of interest to find the number of states that provided the best fit of the model to the data; log likelihood and BIC value. Essentially, we were looking for a local minimum BIC value and a higher log likelihood value (converging closest to 0).

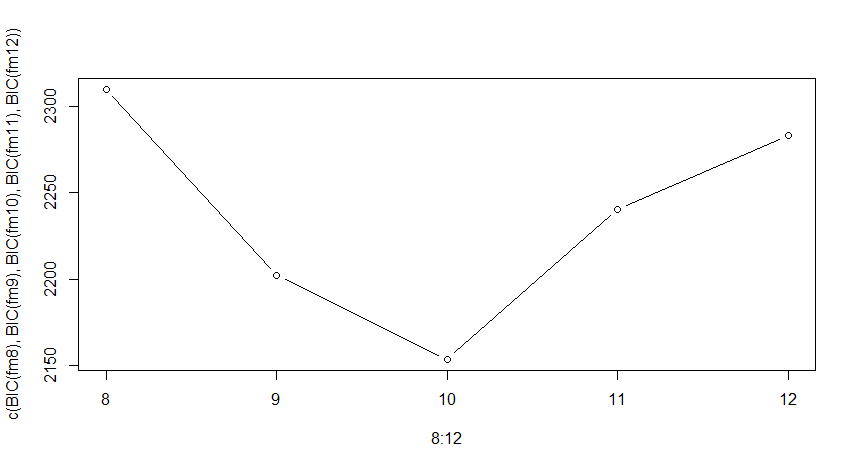
*Figure 2  
Plot of the number of states vs. BIC value of the HMM trained*



Based on figure 2, we noticed a local minimum (of the BIC value) forming at number of states = 10

Focusing on that area, we plotted a more concise graph:

*Figure 3  
Focused view of the number of states vs. BIC value of the HMM trained graph*



Evaluating this plot, we decided our “best” three HMM built were when the **number of states was 9, 10 and 11**. The BIC value was at its minimum when the number of states was equal to 10, however, the log-likelihood for 11 states was higher. This tells us that both have qualities representing a “best-fit” model.

*Table 1  
BIC values for each HMM*

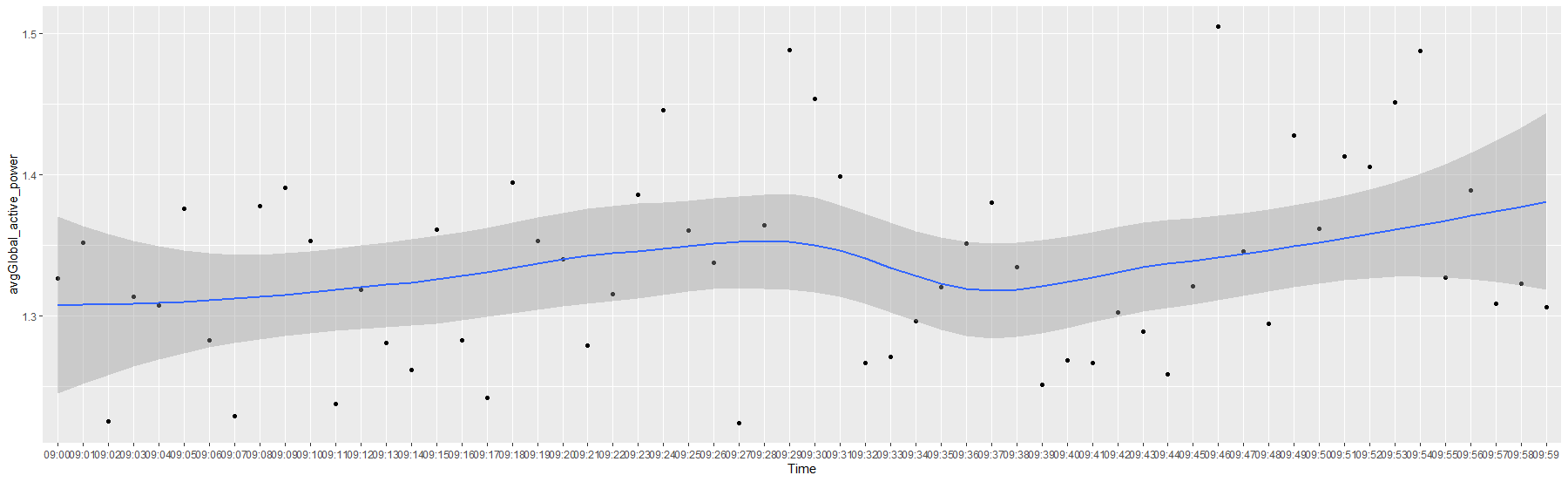
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number of states** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| **BIC Value** | 4642.777 | 3167.232 | 2727.849 | 2479.326 | 2337.619 | 2347.692 | 2310.064 | 2202.089 | 2153.462 | 2240.43 | 2283.081 | 2305.57 | 2620.866 | 2673.586 |

However, for the final say, our group agrees that at 10 states, the model best avoids both overfitting and underfitting.

# Task 2

**Part A**

*Figure 3  
Time series plotting the Average Global Active Power vs Time of Day (09:00 to 10:00)*



The graph experiences an increase in the average global active power until approximately 9:30, where there is a slight dip until approximately 9:45, where then the power increases again.

We believe that the slight increase is due to people waking up on Sundays, and then using utilities such as the cooking, television, etc. One potential reason for the slight dip at approximately 9:30 is because multiple religions host events on Sundays at around 10:00. So people leaving their premises and travelling to the event could cause the household power consumption to experience a short dip.