Assignment 3 R language Report

Group 14

Junchen Li (301385486)

Kwok Yee Cheung (301367833)

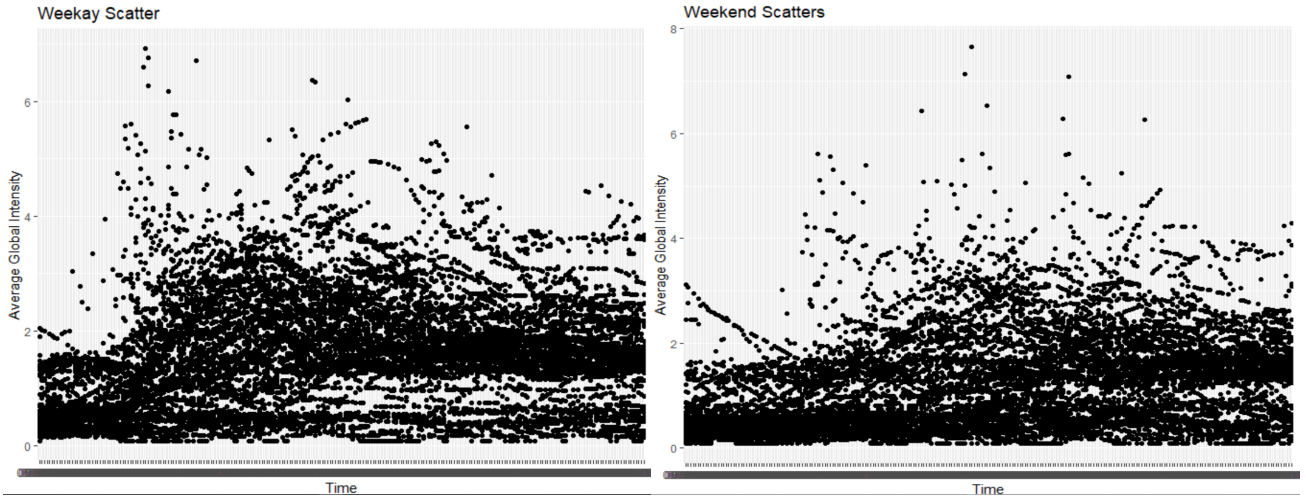
1 Data Exploration

Time windows:

1. On weekdays: Each Thursday, 6am to 9am, throughout the year of 2008
2. On weekend days: Each Saturday, 6am to 9am, throughout the year of 2008

Visualization:

By visualizing the two time windows respectively, it can be found that the power consumption patterns of the two time Windows are different.



2 Model Training

Spite dataset:

The two datasets were divided, and the size of the divided data set was as follows

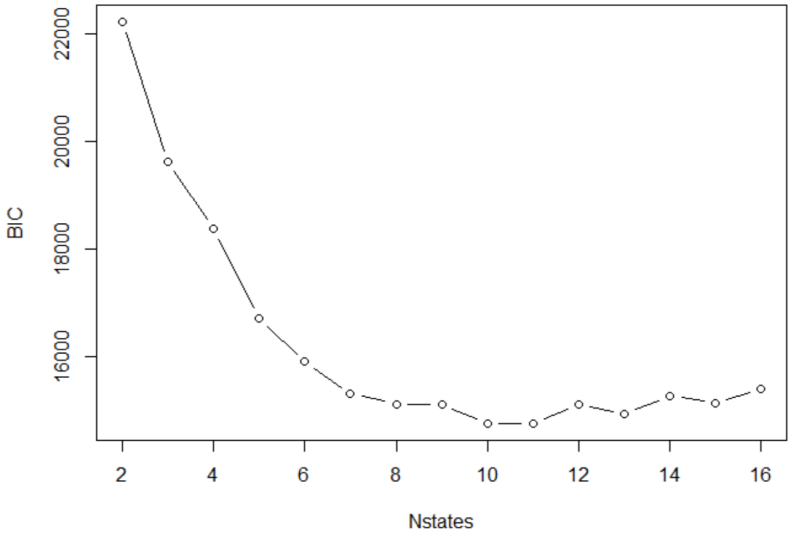
1. On weekdays: train dataset has 10061 samples and test dataset has 2419 samples
2. On weekend: train dataset has 9797 samples and test dataset has 2443 samples

Variable selected:

In order to build multivariate HMMs, we use Sub_metering_1, Sub_metering_2 and Sub_metering_3 to fit Global_active_power in this study.

Find the best model

We experimented with global active power on weekdays mentioned in question2. We chose Nstate from 2 to 16(15 observations) as per requirement, and observed given the timeframe, we obtained the lowest BIC value when NSTATES=10, and obtained a value of 14754.22.



Output

converged at iteration 331 with logLik: -6736.365

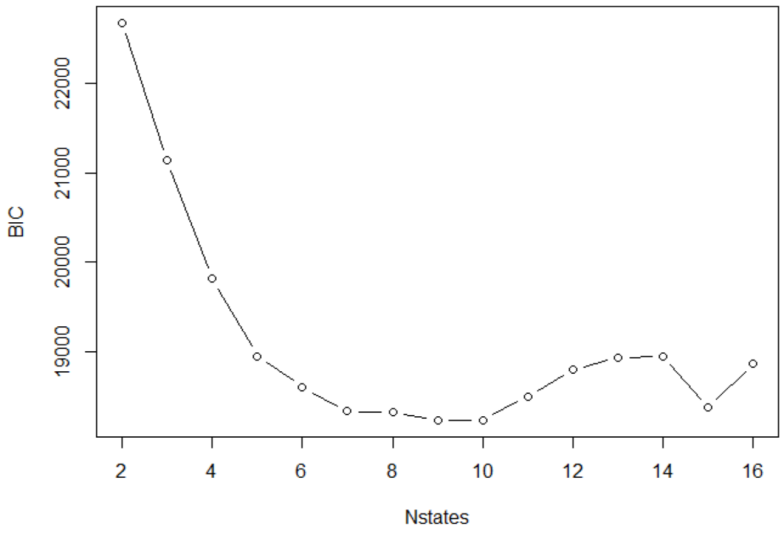
Convergence info: Log likelihood converged to within tol. (relative change)

'log Lik.' -6736.365 (df=139)

AIC: 13750.73

BIC: 14753.81

Then, we experimented with global active power on weekend mentioned in question2. We chose Nstate from 2 to 16(15 observations) as per requirement, and observed given the timeframe, we obtained the lowest BIC value when NSTATES=9, and obtained a value of 18235.04



Output:

Convergence info: 'maxit' iterations reached in EM without convergence.

'log Lik.' -8584.508 (df=116)

AIC: 17401.02

BIC: 18235.04

3 Model Testing

First, test on the Weekday test dataset, where nStates =10. I get BIC and log-likelihood value as follows

We added scale for standardization, and then compared with HMM model, BIC and loglik, the results are as follows

```
> BIC(fm_new)
[1] 5217.30
> logLik(fm_new)
'log Lik.' -1851.77 (df=139)
```

Test on the Weekend test dataset,

```
> BIC(fm_new)
[1] 5772.02
> logLik(fm_new)
'log Lik.' -2133.51 (df=139)
```

After data comparison, I found that there was little difference before and after standardization, which indicated that data dimension had little influence on data.

4 Anomaly Detection

After reading in the data, we select the same time window and use the model in Question 3 for fitting.

On the Weekday test dataset with anomaly, BIC and log-likelihood value as follows

```
> BIC(fm_new)
[1] 1537.668
> logLik(fm_new)
'log Lik.' -387.9294 (df=139)
```

On the Weekend test dataset with anomaly, BIC and log-likelihood value as follows

```
> BIC(fm_new)
[1] 1526.215
> logLik(fm_new)
'log Lik.' -382.2029 (df=139)
```

For an HMM model, you can simply determine the maximum likelihood of your and based on the same time Window data. Specifically saying, if the maximum likelihood difference between the new data and the test data is large, it means that the model on the test data is not applicable to the new data, indicating that there are data anomalies on the new data. The maximum likelihood estimate is smaller than the test set, so it can be inferred that an exception occurs.