CMPT 318

TERM PROJECT

ALI ALDEN 301302901

**Anomaly Detection using Hidden Markov Models**

**with Data Analysis**

For Dr. Uwe Glasser

Simon Fraser University

Fall 2021

Abstract:

-machine learning in today’s environment

-working with the right features and models

-using statistical analysis tools to find anomaly activities

-working with

1. Feature Engineering

-We did 3 analyses on the data with PCA.

1) Data as a whole:

The prcomp() fn was ran over 3 years of test data and here are the ff numbers:

Importance of components

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Standard deviation | 1.6911 | 0.9992 | 0.9698 | 0.9133 | 0.8777 | 0.68606 | 0.35513 |
| Proportion of Variance | 0.4086 | 0.1426 | 0.1343 | 0.1192 | 0.1101 | 0.06724 | 0.01802 |
| Cumulative Proportion | 0.4086 | 0.5512 | 0.6855 | 0.8047 | 0.9147 | 0.98198 | 1.00000 |

PC1 accounts for 40% of variability in the data.

Chart, bar chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | PC1 | | PC2 | | PC3 | | PC4 | | PC5 | | PC6 | | | PC7 |
| Global\_active\_power | | **-0.46872** | | 0.134757 | | -0.08736 | | -0.06854 | | 0.261449 | | -0.76918 | | | -0.29968 |
| Global\_reactive\_power | | -0.19475 | | -0.74423 | | 0.166002 | | 0.60787 | | 0.064466 | | -0.0329 | | | -0.07678 |
| Voltage | | 0.330526 | | -0.13246 | | -0.03506 | | -0.13266 | | 0.919 | | 0.081727 | | | 0.056028 |
| Global\_intensity | | **-0.5596** | | 0.019129 | | 0.001156 | | -0.06409 | | 0.138189 | | 0.081176 | | | 0.810364 |
| Sub\_metering\_1 | | -0.29884 | | -0.12875 | | 0.728446 | | -0.47839 | | 0.042941 | | 0.240646 | | | -0.27363 |
| Sub\_metering\_2 | | -0.28379 | | -0.41294 | | -0.65121 | | -0.42742 | | -0.05497 | | 0.286867 | | | -0.23846 |
| Sub\_metering\_3 | -0.38747 | | 0.472183 | | -0.09419 | | 0.438797 | | 0.242829 | | 0.503786 | | -0.33575 |

C) data for 6 months, JAN 2010 – to July 2010

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Global\_active\_power | **0.500356** | -0.15051 | 0.101917 | -0.13568 | -0.0053 | 0.77475 | -0.31294 |
| Global\_reactive\_power | 0.240051 | **0.595666** | -0.20484 | -0.02376 | -0.73251 | -0.05213 | -0.07575 |
| Voltage | -0.26098 | -0.05891 | 0.209997 | -0.92553 | -0.16294 | -0.02391 | 0.024278 |
| Global\_intensity | **0.554497** | -0.04634 | 0.034814 | -0.13284 | 0.059017 | -0.06547 | 0.814714 |
| Sub\_metering\_1 | 0.325075 | 0.178927 | -0.60464 | -0.32577 | 0.481271 | -0.26847 | -0.29479 |
| Sub\_metering\_2 | 0.271993 | 0.432722 | 0.72738 | 0.012405 | 0.280928 | -0.27697 | -0.23218 |
| Sub\_metering\_3 | 0.369848 | -0.63059 | 0.087577 | 0.022243 | -0.35046 | -0.49336 | -0.30196 |

Since PC1 has the most variability with 40% and PC2 with 17.2% also the fact that data is consistent, highlighted variables will be focused on for the model:

1. Global Active Power
2. Global intensity

Pc2 components hold very much value to global reactive power but not enough to justify the 40% variability of PC1.

**Part 2. Feature Selection and observation time window.**

-The test and train data set are every Monday from 7am to 10pm (3 hours inclusive).

-This time window is consistent in data, we ran simple moving average on data and checked anomaly data sets, the values are different and “odd” for the cycle of data for some weeks in the anomaly data sets.

**Part 3. HMM Training and Testing**

Data has been divided to two subsets: Test data of 3 years from December 2006 to December 2008 and training data from January 2008 to December 2009. The data is also a subset of Monday morning from 6 to 9am (3 hours).

This gives Test data to be 106 weeks of testing and 50 weeks of training.

1. HMM TRAINING:

-data has been ran for 4-24 states for find best model.

N states = 24.

Here are their values.

We ran the model 2 times on the same model, the values are then compared, this gave us these following numbers:

[1] "TRAIN"

[1] 4

converged at iteration 40 with logLik: -18023.89

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

'log Lik.' -18023.89 (df=31)

AIC: 36109.78

BIC: 36353.33

[1] "TEST"

[1] 4

converged at iteration 57 with logLik: -7878.144

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

'log Lik.' -7878.144 (df=31)

AIC: 15818.29

BIC: 16037.92

[1] "TRAIN"

[1] 5

converged at iteration 33 with logLik: -12781.91

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

'log Lik.' -12781.91 (df=44)

AIC: 25651.82

BIC: 25997.51

[1] "TEST"

[1] 5

converged at iteration 66 with logLik: -5875.928

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

'log Lik.' -5875.928 (df=44)

AIC: 11839.86

BIC: 12151.59

[1] "TRAIN"

[1] 6

converged at iteration 45 with logLik: -11196.91

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

'log Lik.' -11196.91 (df=59)

AIC: 22511.81

BIC: 22975.34

[1] "TEST"

[1] 6

converged at iteration 42 with logLik: -3889.521

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

'log Lik.' -3889.521 (df=59)

AIC: 7897.042

BIC: 8315.044

[1] "TRAIN"

[1] 7

converged at iteration 58 with logLik: -7224.733

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

'log Lik.' -7224.733 (df=76)

AIC: 14601.47

BIC: 15198.55

[1] "TEST"

[1] 7

converged at iteration 77 with logLik: -2872.481

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

'log Lik.' -2872.481 (df=76)

AIC: 5896.961

BIC: 6435.404

[1] "TRAIN"

[1] 8

converged at iteration 63 with logLik: -5049.951

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

'log Lik.' -5049.951 (df=95)

AIC: 10289.9

BIC: 11036.26

[1] "TEST"

[1] 8

converged at iteration 105 with logLik: -2915.289

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

'log Lik.' -2915.289 (df=95)

AIC: 6020.579

BIC: 6693.633

[1] "TRAIN"

[1] 9

converged at iteration 89 with logLik: -3513.379

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

'log Lik.' -3513.379 (df=116)

AIC: 7258.758

BIC: 8170.1

[1] "TEST"

[1] 9

converged at iteration 79 with logLik: -2090.329

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

'log Lik.' -2090.329 (df=116)

AIC: 4412.657

BIC: 5234.492

[1] "TRAIN"

[1] 4

converged at iteration 40 with logLik: -18023.89

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

'log Lik.' -18023.89 (df=31)

AIC: 36109.78

BIC: 36353.33

[1] "TEST"

[1] 4

converged at iteration 57 with logLik: -7878.144

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

'log Lik.' -7878.144 (df=31)

AIC: 15818.29

BIC: 16037.92

[1] "TRAIN"

[1] 5

converged at iteration 33 with logLik: -12781.91

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

'log Lik.' -12781.91 (df=44)

AIC: 25651.82

BIC: 25997.51

[1] "TEST"

[1] 5

converged at iteration 66 with logLik: -5875.928

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

'log Lik.' -5875.928 (df=44)

AIC: 11839.86

BIC: 12151.59

[1] "TRAIN"

[1] 6

converged at iteration 45 with logLik: -11196.91

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

'log Lik.' -11196.91 (df=59)

AIC: 22511.81

BIC: 22975.34

[1] "TEST"

[1] 6

converged at iteration 42 with logLik: -3889.521

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

'log Lik.' -3889.521 (df=59)

AIC: 7897.042

BIC: 8315.044

[1] "TRAIN"

[1] 7

converged at iteration 58 with logLik: -7224.733

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

'log Lik.' -7224.733 (df=76)

AIC: 14601.47

BIC: 15198.55

[1] "TEST"

[1] 7

converged at iteration 77 with logLik: -2872.481

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

'log Lik.' -2872.481 (df=76)

AIC: 5896.961

BIC: 6435.404

[1] "TRAIN"

[1] 8

converged at iteration 63 with logLik: -5049.951

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

'log Lik.' -5049.951 (df=95)

AIC: 10289.9

BIC: 11036.26

[1] "TEST"

[1] 8

converged at iteration 105 with logLik: -2915.289

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

'log Lik.' -2915.289 (df=95)

AIC: 6020.579

BIC: 6693.633

[1] "TRAIN"

[1] 9

converged at iteration 89 with logLik: -3513.379

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

'log Lik.' -3513.379 (df=116)

AIC: 7258.758

BIC: 8170.1

[1] "TEST"

[1] 9

converged at iteration 79 with logLik: -2090.329

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

'log Lik.' -2090.329 (df=116)

AIC: 4412.657

BIC: 5234.492

**[1] "TRAIN"**

**[1] 10**

**converged at iteration 230 with logLik: -2975.834**

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

**'log Lik.' -2975.834 (df=139)**

**AIC: 6229.667**

**BIC: 7321.706**

**[1] "TEST"**

**[1] 10**

**converged at iteration 122 with logLik: -1056.556**

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

**'log Lik.' -1056.556 (df=139)**

**AIC: 2391.113**

**BIC: 3375.897**

[1] "TRAIN"

[1] 11

converged at iteration 265 with logLik: -1152.013

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

'log Lik.' -1152.013 (df=164)

AIC: 2632.026

BIC: 3920.474

[1] "TEST"

[1] 11

converged at iteration 105 with logLik: -1116.486

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

'log Lik.' -1116.486 (df=164)

AIC: 2560.972

BIC: 3722.876

[1] "TRAIN"

[1] 12

converged at iteration 154 with logLik: -206.6025

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

'log Lik.' -206.6025 (df=191)

AIC: 795.205

BIC: 2295.777

[1] "TEST"

[1] 12

converged at iteration 268 with logLik: -1056.944

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

'log Lik.' -1056.944 (df=191)

AIC: 2495.889

BIC: 3849.081

[1] "TRAIN"

[1] 13

converged at iteration 384 with logLik: 365.8019

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

'log Lik.' 365.8019 (df=220)

AIC: -291.6038

BIC: 1436.803

[1] "TEST"

[1] 13

converged at iteration 213 with logLik: -593.247

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

'log Lik.' -593.247 (df=220)

AIC: 1626.494

BIC: 3185.145

[1] "TRAIN"

[1] 14

converged at iteration 311 with logLik: 1882.959

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

'log Lik.' 1882.959 (df=251)

AIC: -3263.918

BIC: -1291.963

[1] "TEST"

[1] 14

converged at iteration 212 with logLik: 1013.177

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

'log Lik.' 1013.177 (df=251)

AIC: -1524.354

BIC: 253.925

[1] "TRAIN"

[1] 11

converged at iteration 265 with logLik: -1152.013

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

'log Lik.' -1152.013 (df=164)

AIC: 2632.026

BIC: 3920.474

[1] "TEST"

[1] 11

converged at iteration 105 with logLik: -1116.486

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

'log Lik.' -1116.486 (df=164)

AIC: 2560.972

BIC: 3722.876

[1] "TRAIN"

[1] 12

converged at iteration 154 with logLik: -206.6025

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

'log Lik.' -206.6025 (df=191)

AIC: 795.205

BIC: 2295.777

[1] "TEST"

[1] 12

converged at iteration 268 with logLik: -1056.944

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

'log Lik.' -1056.944 (df=191)

AIC: 2495.889

BIC: 3849.081

[1] "TRAIN"

[1] 13

converged at iteration 384 with logLik: 365.8019

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

'log Lik.' 365.8019 (df=220)

AIC: -291.6038

BIC: 1436.803

[1] "TEST"

[1] 13

converged at iteration 213 with logLik: -593.247

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

'log Lik.' -593.247 (df=220)

AIC: 1626.494

BIC: 3185.145

[1] "TRAIN"

[1] 14

converged at iteration 311 with logLik: 1882.959

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

'log Lik.' 1882.959 (df=251)

AIC: -3263.918

BIC: -1291.963

[1] "TEST"

[1] 14

converged at iteration 212 with logLik: 1013.177

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

'log Lik.' 1013.177 (df=251)

AIC: -1524.354

BIC: 253.925

The value starts diverging between 11 to 14 states (we are overfitting the model.)

In terms of model selection, state 10 seems to be the one making sense in the test data and training data with the value of.

Training Log Likelihood (Normalized with N observations): (N is in minutes)

**TRAIN data**

**States = 10**

**logLik (Normalized): 0.15596614255765199161425576519916**

**BIC (normalized): 0.38373721174004192872117400419287**

**TEST data**

**logLik (Normalized): 0.11979092970521541950113378684807**

**BIC (normalized): 0.38275476190476190476190476190476**

As you can see, the normalized values are very close together, meaning that the data is not overfit.

PART 3)

ANOMALY DETECTION:

-we ran the data into the HMM model and the loglikelihood is way off for the 1st Dataset with Anomalies.

Using an actual year comparison and a normalized likelihood of the Model there is something wrong as the value is close to the model.

[1] "MODEL"

[1] 10

converged at iteration 96 with logLik: -1086.82

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

'log Lik.' -1086.82 (df=139)

AIC: 2451.639

BIC: 3436.423

[1] "ANOMALY DATASET 1 "

converged at iteration 108 with logLik: -6733.942

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

'log Lik.' -6733.942 (df=139)

AIC: 13745.88

BIC: 14736.23

Here the LOG likelihood of the Model and the Anomaly data 1.

Assuming N states = 10

MODEL data (AVG of Train and Test)

logLik (Normalized): **0.137878**

BIC (normalized): **0.383246**

Anomaly dataset 1

logLik (Normalized): **0.733544**

BIC: **1.6052538**

This means that given the likelihood of these output Data to prop up is 3 times less likely than the model.

Anomaly Data 2:

[1] "MODEL"

converged at iteration 263 with logLik: -1739.688

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

'log Lik.' -1739.688 (df=139)

AIC: 3757.376

BIC: 4742.16

[1] "ANOMALY DATASET 2"

converged at iteration 82 with logLik: -6004.436

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

'log Lik.' -6004.436 (df=139)

AIC: 12286.87

BIC: 13277.22

Here the LOG likelihood of the Model compared to Anomaly data 2.

Assuming N states = 10

MODEL data (AVG of Train and Test)

logLik (Normalized): **0.137878**

BIC (normalized): **0.383246**

Anomaly dataset 2

logLik (Normalized): **0.624077**

BIC: **1.446320**

Anomaly Data 3:

[1] "MODEL"

converged at iteration 152 with logLik: -1194.151

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

'log Lik.' -1194.151 (df=139)

AIC: 2666.301

BIC: 3651.085

[1] "ANOMALY DATASET 3"

converged at iteration 165 with logLik: -6294.694

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

'log Lik.' -6294.694 (df=139)

AIC: 12867.39

BIC: 13857.73

Here the LOG likelihood of the Model compared to Anomaly data 3.

Anomaly dataset 3

logLik (Normalized): **0.68569**

BIC: **1.509556**

**What are affecting the values?**

**We ran a simple moving average (to not overfit the data to see what is going on.)**

**Chart

Description automatically generated**Model data from every Monday with ROLLING MEAN OF 30 MINUTES PER OBSERVATION. (Global power)

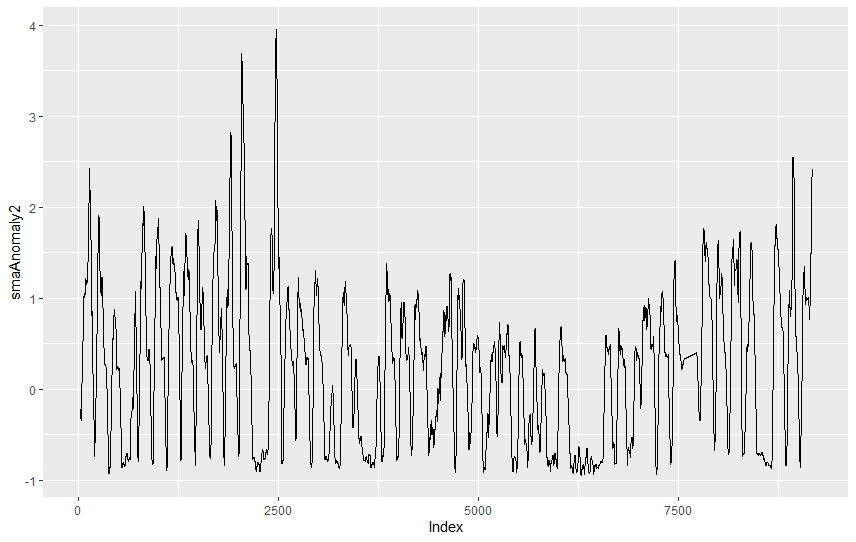
Model data from every Monday with ROLLING MEAN OF 30 MINUTES PER OBSERVATION. (Global intensity) **Chart

Description automatically generated**

Anomaly Dataset 1 with Rolling mean of 30 minutes (global intensity)

**A picture containing text, group, different, several

Description automatically generated**

****Anomaly Dataset 2 with Rolling mean of 30 minutes (global intensity)

Anomaly Dataset 3 with Rolling mean of 30 minutes (global intensity)

**Graphical user interface, chart

Description automatically generated**

All dataset and Code can also be found here:

<https://github.com/mandog2005/CMPT318_FALL2021>

Resources: