



Term Project

Hidden Markov Models and Anomaly Detection
Group 15

Tasks

1. Principal Component Analysis
2. Training and Testing Hidden Markov Models
3. Anomaly Detection

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

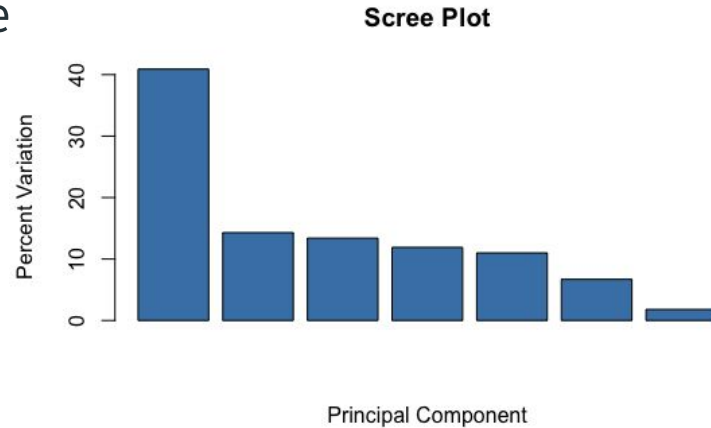
1.

Principal Component Analysis

1. Reduces dimensionality of data
2. Retains patterns and trends
3. Simplifies data analysis

Principal Component Analysis

1. Each component is responsible for certain percentage of variability
2. Pick the component with the highest percentage



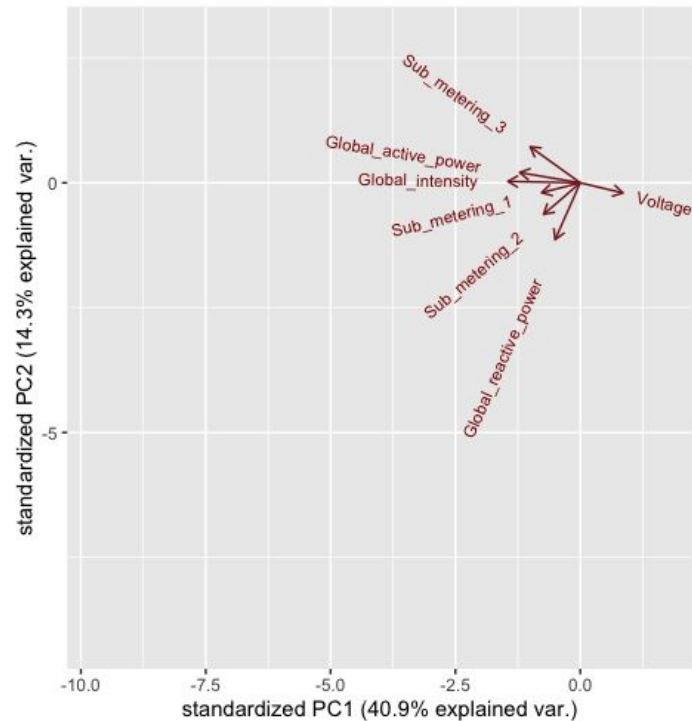
Principal Component Analysis

| Variables | Correlation |
|-----------------------|-------------|
| Global active power | -0.46 |
| Global reactive power | -0.19 |
| Voltage | 0.33 |
| Global Intensity | -0.55 |
| Sub metering 3 | -0.29 |
| Sub metering 2 | -0.28 |
| Sub metering 3 | -0.38 |

Loading values of PCA 1

Principal Component Analysis

1. Vectors represent the coefficient of the variable on the principal components
2. Vectors pointing in the same direction have similar effect on the component



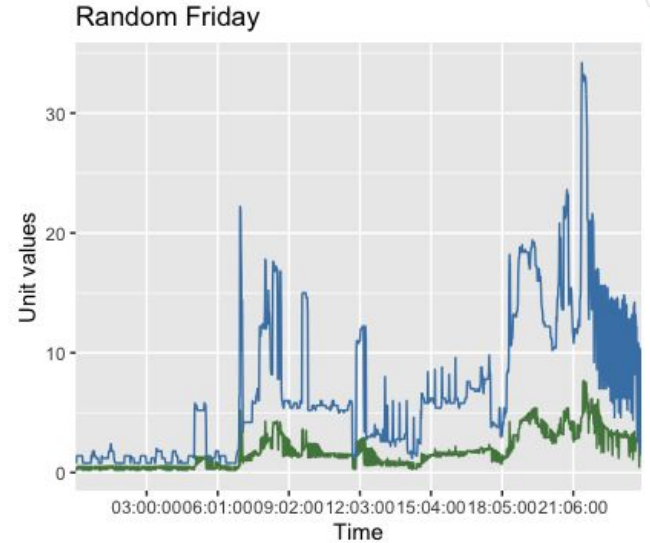
2. Hidden Markov Model

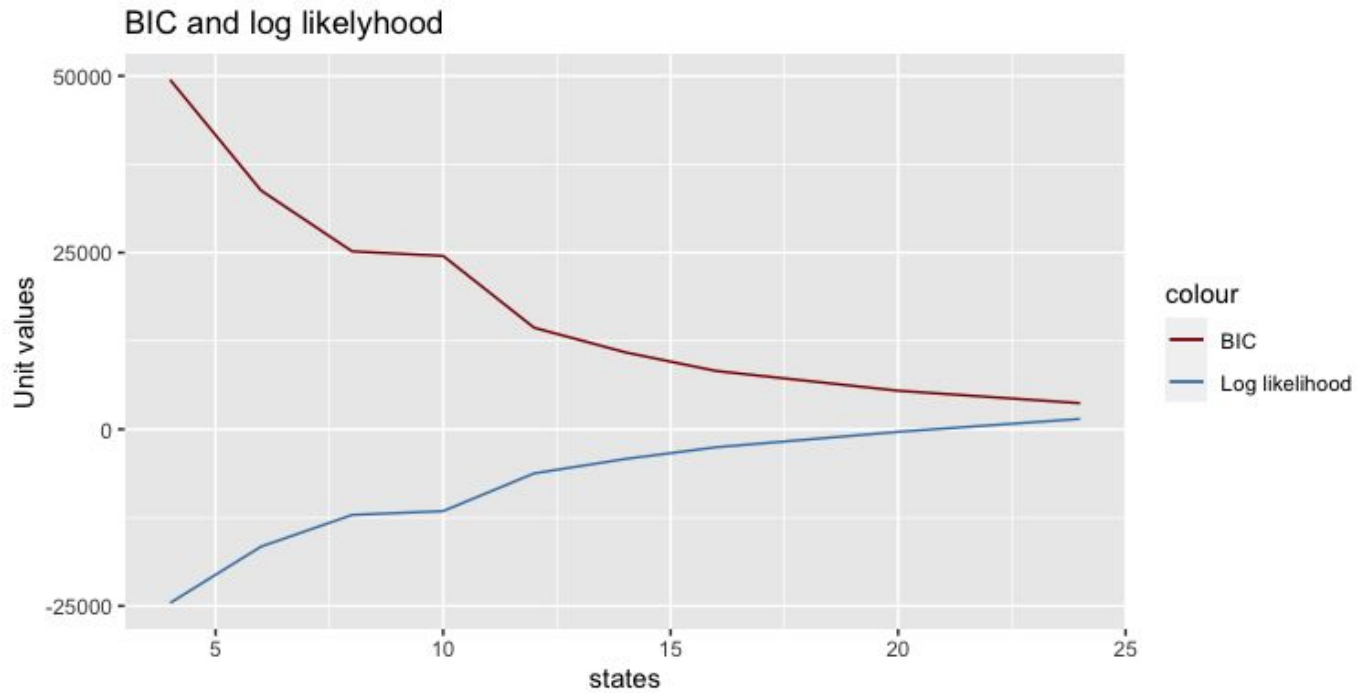
Stochastic processes where the states are unobservable and then next state relies only on the present state



Data Preprocessing

1. We picked Fridays from 6:00 PM to 10:00 PM
 - I. Peaks
 - II. Pattern observed
2. Divided dataset into 70% for training and 30% for testing





The log likelihood and BIC reduces as the number of state increases.

Hidden Markov Model

1. Picked state 20
 - I. Had the highest log likelihood and the lowest BIC
 - II. No overfitting

| States | Test log likelihood | Train log likelihood |
|--------|---------------------|----------------------|
| 16 | -34.88603 | -23.82227 |
| 20 | -18.44902 | -3.38566 |

State 20 is picked as it has a high log likelihood for test and train data

3. **Anomaly Detection**

Identification of unexpected events or observations that deviate from the norm



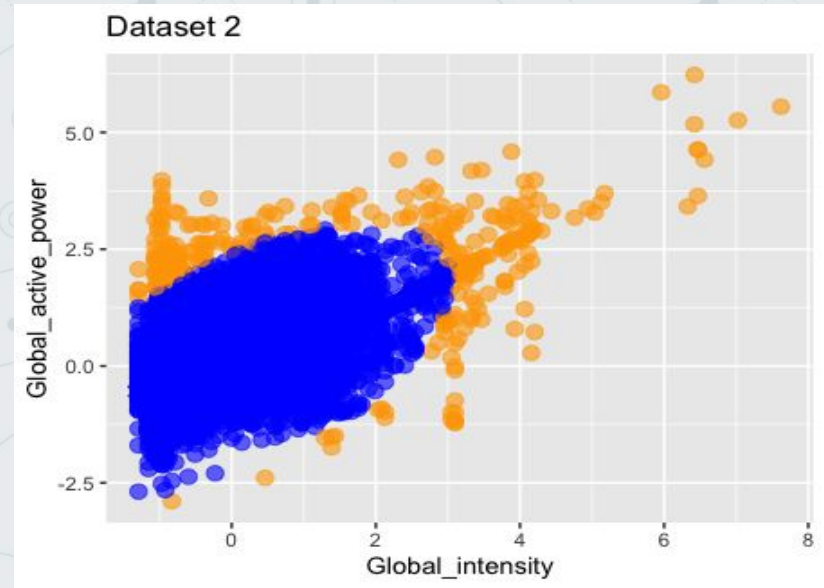
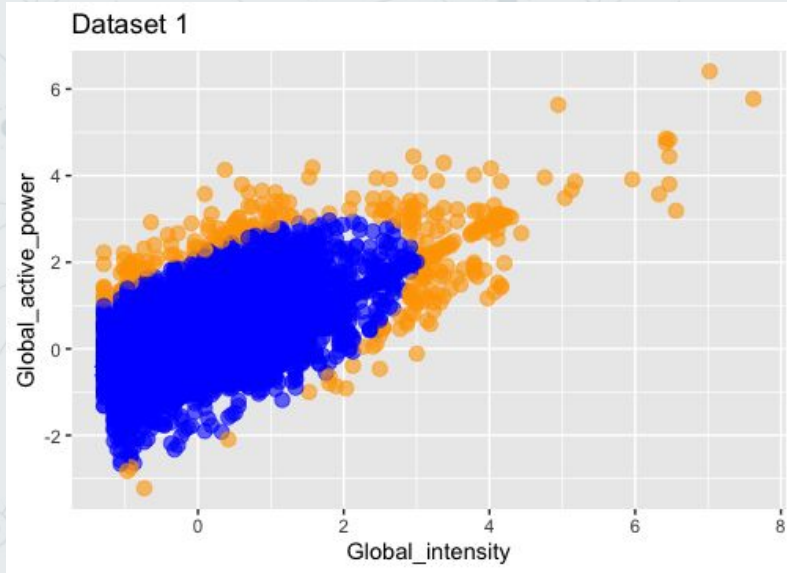
Anomaly Detection

1. Presence of outliers gives a lower log likelihood at the chosen state

| Dataset | BIC | Log likelihood |
|---------|-----------|----------------|
| 1 | 9677.169 | -2647.221 |
| 2 | 11434.964 | -3526.119 |
| 3 | 10286.065 | -3683.649 |

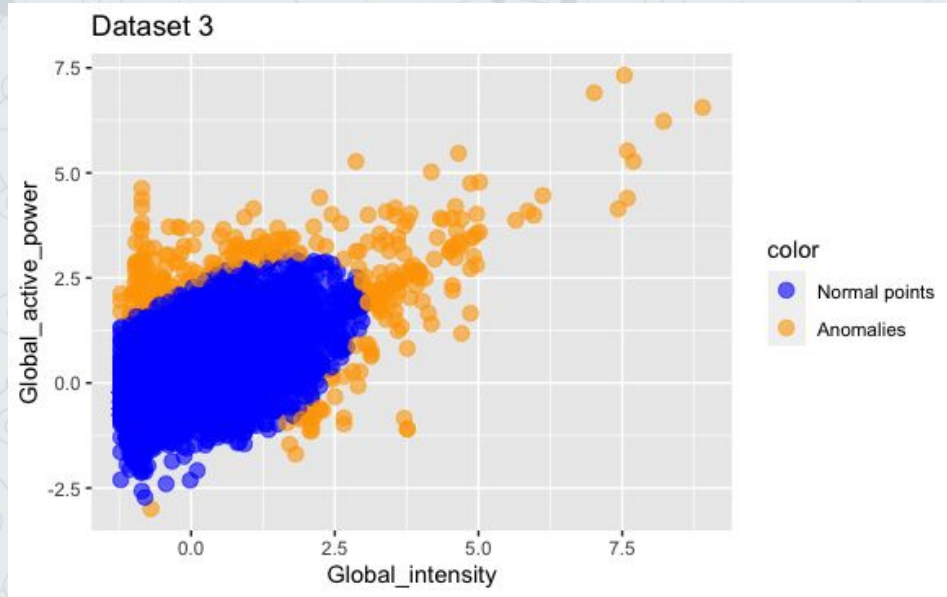
Hypothesis: Dataset 3 is the most anomalous

Anomaly Detection



- Normal Data points
- Anomalous Data points

Anomaly Detection



1. Mahalanobis Distance was used to compute the distance of each point
2. If distance > cutoff distance then its marked as an anomaly
3. Dataset 3 is the most anomalous
4. Aligns with our previous hypothesis

- Normal Data points
- Anomalous Data points

Lessons Learned

1. PCA analysis to reduce complexity
2. Training and testing HMM models
3. Anomaly detection in a multivariate model



The background of the slide is a light blue-grey color with a complex network pattern. This pattern consists of numerous small circles, some of which are solid grey and others are hollow with grey outlines. These circles are interconnected by a web of thin, light grey lines, creating a dense, interconnected mesh that covers the entire slide area.

Thanks!

Any questions?