Term Project Report

Exploring Anomaly-Detection Based Intrusion Detection Methods on Electricity

CMPT 318 Spring 2023

Abstract

Improved efficiency and service quality have resulted from the growing focus on automation in vital infrastructure, including electric power grids, public water utilities, and smart transportation networks. The protection of such infrastructure, however, must be given top attention because doing so also exposes these systems to various adversarial situations and vulnerabilities. Identifying unexpected patterns that can point to cyberattacks or system breakdowns depends on a good anomaly detection system. The primary objective of this project is to develop an effective anomaly detection method for normal electricity consumption data, using multivariate Hidden Markov Models (HMMs). The project aims to address several challenges, including imperfections in the data, varying types of anomalies depending on the application context, and striking a balance between precision and recall while reducing the false alarm rate. To achieve this, the project will involve feature engineering using Principal Component Analysis (PCA), HMM training and testing, and evaluating the model's performance on three different datasets with injected anomalies.

Table of Contents

1. Introduction

The increasing integration of automation in critical infrastructure has resulted in a pressing need for effective anomaly detection methods to ensure the security and reliability of these systems. This report presents an in-depth exploration of a multivariate Hidden Markov Model-based approach to anomaly detection in normal electricity consumption data. The study comprises three main tasks: feature engineering, HMM training and testing, and anomaly detection evaluation. Principal Component Analysis is used in feature engineering to choose the best response variables for HMM training (PCA) [1]. Data normalization and a justification for the final response variable selection based on PCA results are included in this stage. Partitioning the scaled data, choosing an acceptable time window, and training several multivariate HMMs with varied numbers of states are all part of the HMM training and testing procedure. To identify the best-fitting model, the effectiveness of these models is compared using log-likelihood and Bayesian Information Criterion (BIC) metrics. The degree of anomalies present in each dataset is studied and compared after the chosen model is used for anomaly identification on three datasets with injected anomalies. This report offers a thorough summary of the experimental analysis and results from interpreting the datasets on electric energy usage, including diagrams, graphs, and tables to show the experiments and results. The goal of the study is to improve critical infrastructure cybersecurity by creating a reliable and accurate anomaly detection model.

2. Methodology Overview

The methodology used in this project involves several steps aimed at analyzing and detecting anomalies in electricity consumption datasets using principal component analysis and hidden Markov models. The methods used to detect anomalies are further explained below.

2.1 Data Extraction and Preparation

The dataset for power usage must first be extracted and prepared for analysis. Use of the read.table() method loads the dataset into the R environment. The lubridate library's as.Date() and parse hms() methods are used to convert the Date and Time columns to the correct data formats. The Date column's wday() and strftime() operations are used to retrieve the weekday and week numbers. The scale() function in base R is then used to scale the data. By utilizing the na.omit() method, the NA values are removed from the data.

2.2 Principal Component Analysis (PCA)

Principal component analysis is a statistical method for reducing the number of dimensions in data while maintaining as much of the original data as possible [1]. The number of variables in the dataset for power usage is decreased in this project using principal component analysis. The time frame selected was on Fridays from 18:00 to 21:00 and is used as the training data. This time window is used to subset the data, after which the prcomp() method is used on the subsetted data. The states are displayed against the second main component. By examining the scatter plot of the second principal component against the states, one can observe if there are any anomalies present in the data. Finding patterns in the data that can point to anomalies can be done using the principal component analysis results.

2.3 Hidden Markov Model

A popular category of probabilistic models for sequence analysis is the Hidden Markov Model [2]. The power consumption dataset is modelled in this study using hidden Markov models. To train numerous univariate hidden Markov models on the training set using the depmix() method from the depmixS4 package. The number of states varies with each hidden Markov model and ranges from 3 to a maximum of 16. The Gaussian distribution is used as the emission probability distribution. For fitting the hidden Markov models and printing a summary of the fitted hidden Markov models, the fit() and summary() functions are used. Plotted against the states for visualization, the log-likelihood and BIC are calculated. The HMMs are trained on the training data to learn the normal patterns in the electricity consumption.

2.4 Anomaly Detection

In data, anomalies are deviations from the regular trends. An anomaly detection system can be used to detect irregularities early to give a warning [3]. Three anomalous datasets were used for this project. The same standards used for the training data are applied to set the time window for the normal data. On each anomalous dataset, the preprocessing operations, such as scaling and subset selection, are carried out. The log-likelihood of the anomalous data is calculated using the trained hidden Markov models from the preceding phase. The log-likelihood is calculated using the Hidden Markov Model library's forwardbackward() method. The data point is considered an anomaly if the log-likelihood is less than a predetermined threshold. To determine the reason for the anomaly, further investigation can be done on the anomalous data.

3. Feature Engineering

The goal is to choose a subset of the existing response variables that will be the best ones to train our Hidden Markov Models. For this project, our team performed a principal component analysis i.e PCA to find those variables that shall be used for the training of our multivariate Hidden Markov models.

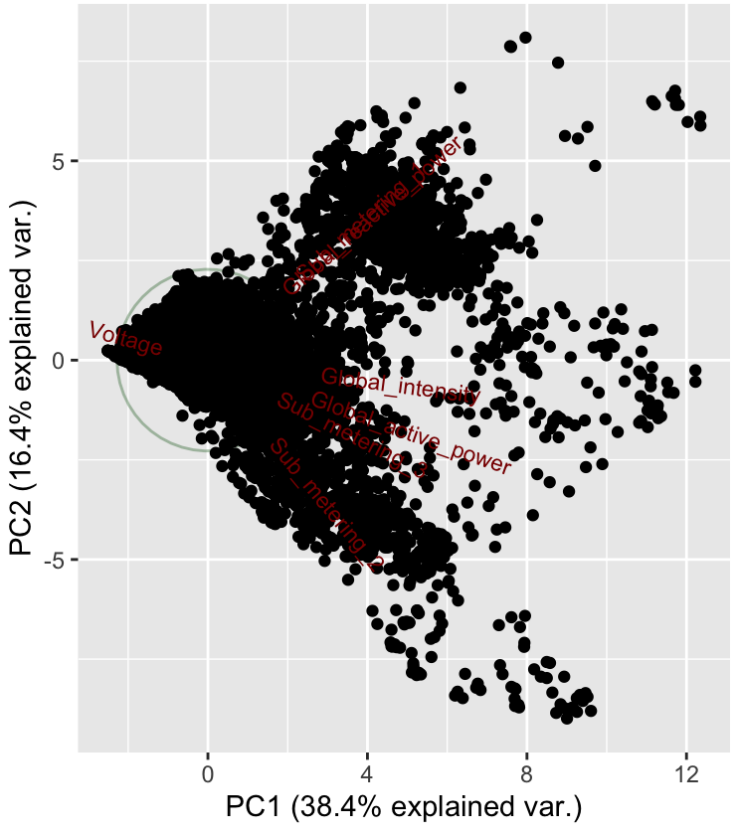

Figure 1 *PCA biplot based on samples from Fridays (18:00 - 21:00)*

As can be observed from the PCA biplot shown in Figure 1, the variance for the principal component 1 (PC1) turned out to be 38.4% while the variance for the principal component 2 (PC2) the variance of 16.4% which is even lesser than the half of the PC1's variance.

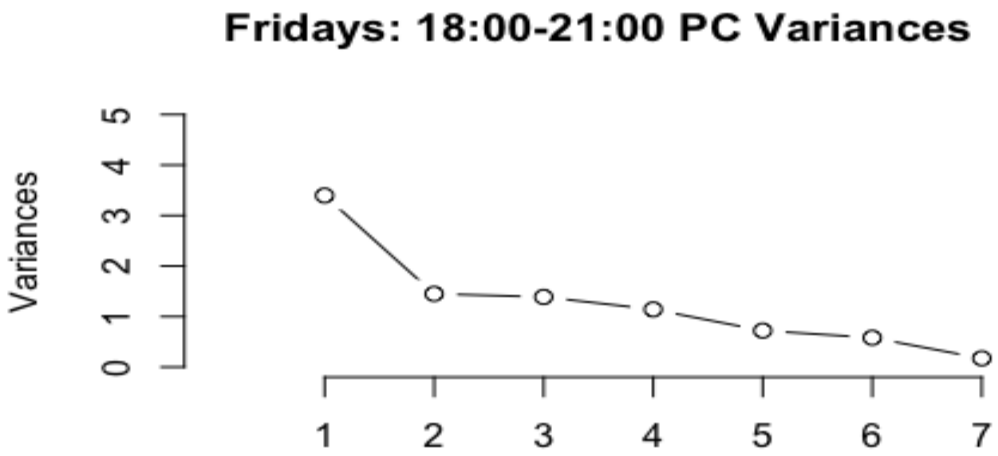

Figure 2 *Variances of PC 1-7 plotted on a line graph*

Furthermore, it can be seen from Figure 2 that the PC Variance for each principal component has been plotted on a line graph. This line graph clearly shows that there is a huge drop in variance as we go from PC1 to PC2 and this trend continues as the variance keeps on declining as we go from PC2 to PC7. Therefore, this line graph helped our team in deciding that we should choose our response variables from PC1 without having any second thoughts.

```
Standard deviations (1, .., p=7):
[1] 1.8429085 1.2038360 1.1781028 1.0681014 0.8493335 0.7616456 0.4153044

Rotation (n x k) = (7 x 7):
                          PC1         PC2         PC3         PC4         PC5         PC6         PC7
Global_active_power    0.4185993 -0.22637522  0.31363118 -0.15302252  0.74529787 -0.12116529 -0.28565461
Global_reactive_power  0.2660333  0.35927277 -0.80648779 -0.29444212  0.18276417 -0.15163922 -0.08144582
Voltage               -0.2049029  0.08625286 -0.10664269 -0.04725782  0.37952877  0.87631151  0.15813076
Global_intensity       0.5689249 -0.11089896  0.05911721 -0.07335092 -0.07884669  0.03611258  0.80475444
Sub_metering_1         0.4721897  0.65801806  0.26125003  0.37446346 -0.16092633  0.21307547 -0.25352746
Sub_metering_2         0.3097770 -0.56961935 -0.39667828  0.55361590 -0.12447602  0.20718894 -0.23938725
Sub_metering_3         0.2634676 -0.20611944  0.10363485 -0.65997033 -0.46848468  0.32368213 -0.34285644
```

Figure 3 *Loadings for every principal component*

After choosing PC1 for the response variables, we had to select features with a higher loading value for our multivariate HMM. The top two features with the highest loading values were Gloabal_intensity and Sub_metering_1 but our multivariate HMM was not converging with Sub_metering_1 being one of the features. So we had to consider the feature with the next highest loading value which was Global_active_power (Figure 3). After selecting Global_intensity and Global_active_power as the final features, our multivariate HMM started to converge.

4. Training and Testing with Hidden Markov Models

Our second task of hidden Markov model training and testing required us to choose a weekday or weekend and a time window of 2 to 6 hours for the chosen day. With the chosen time period, our task was to train multiple multivariate Hidden Markov Models on the train data with a different number of states to compare the trained model's log-likelihood with BIC so that the most suitable model with a good fit on the train data can be selected. Once selecting the best model from the training dataset, we must calculate the log-likelihood of the test data for our selected models to choose the best candidate. To accomplish this task, our code partitioned the scaled data into train and test data with the first 2 years as train data and the last year as test data. Then we trained various multivariate Hidden Markov Models on the train data with a different number of states. We used this trained model to find the log-likelihood of test data.

4.1 Result Comparison and Model Selection

The graph below (*figure 4*), depicts the BIC and log-likelihood values for each number of states for the trained Hidden Markov Model. By examining the graph, our team can select the best model since the model with the highest log-likelihood value with the lowest BIC value is considered to be the most suitable model [4]. According to the graph, the optimal number of states is 24 states since the log-likelihood value is the highest and the BIC value is the lowest compared to other models.

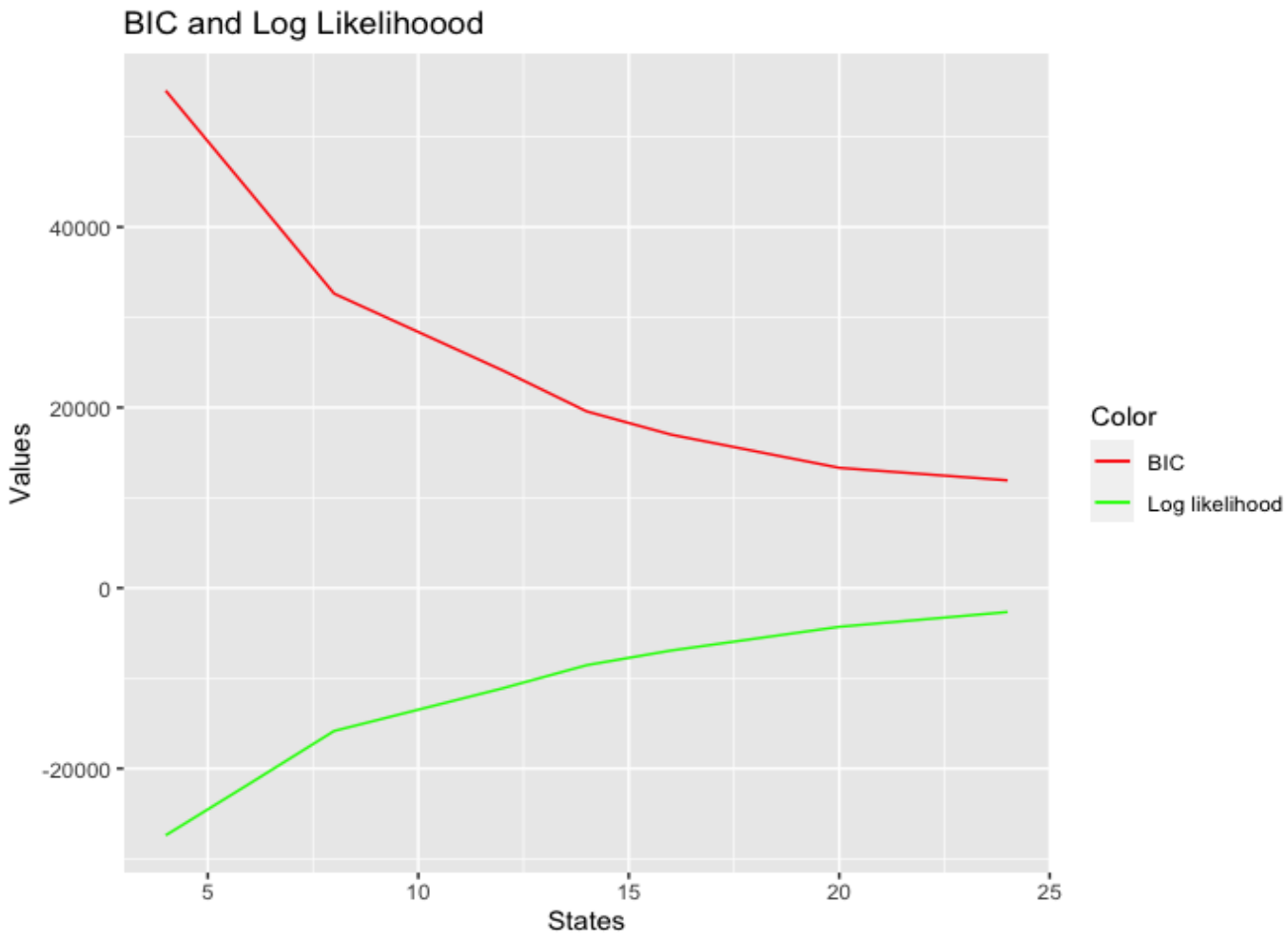

Figure 4 *BIC and Log-Likelihood graph for a given number of states*

The BIC values, normalized log-likelihood for the training data values, and normalized log-likelihood values for the test data are plotted on the y-axis of the second graph (*figure 5*), below, along with the number of states on the x-axis. We can determine whether the chosen HMM model overfits or under fits the data by just looking at the graph (*figure 5*). The model will perform well on the training data but poorly on the test data if it is overfitting to the data, and we will notice a significant discrepancy in the normalized log-likelihood values of the two datasets [3]. On the other hand, if the model under fits the data, it will perform poorly on both the training and test data, and we will observe low normalized log-likelihood values for both datasets.

According to the graph below (*figure 5*), the test data and train data share similar log-likelihood values, signalling that the model is not overfitting to the training data. In addition, since the log-likelihood values for the test and train data are similar, it suggests that the model that our group chose has good performance since the model was able to capture the underlying pattern of the data and make accurate predictions on novel data.

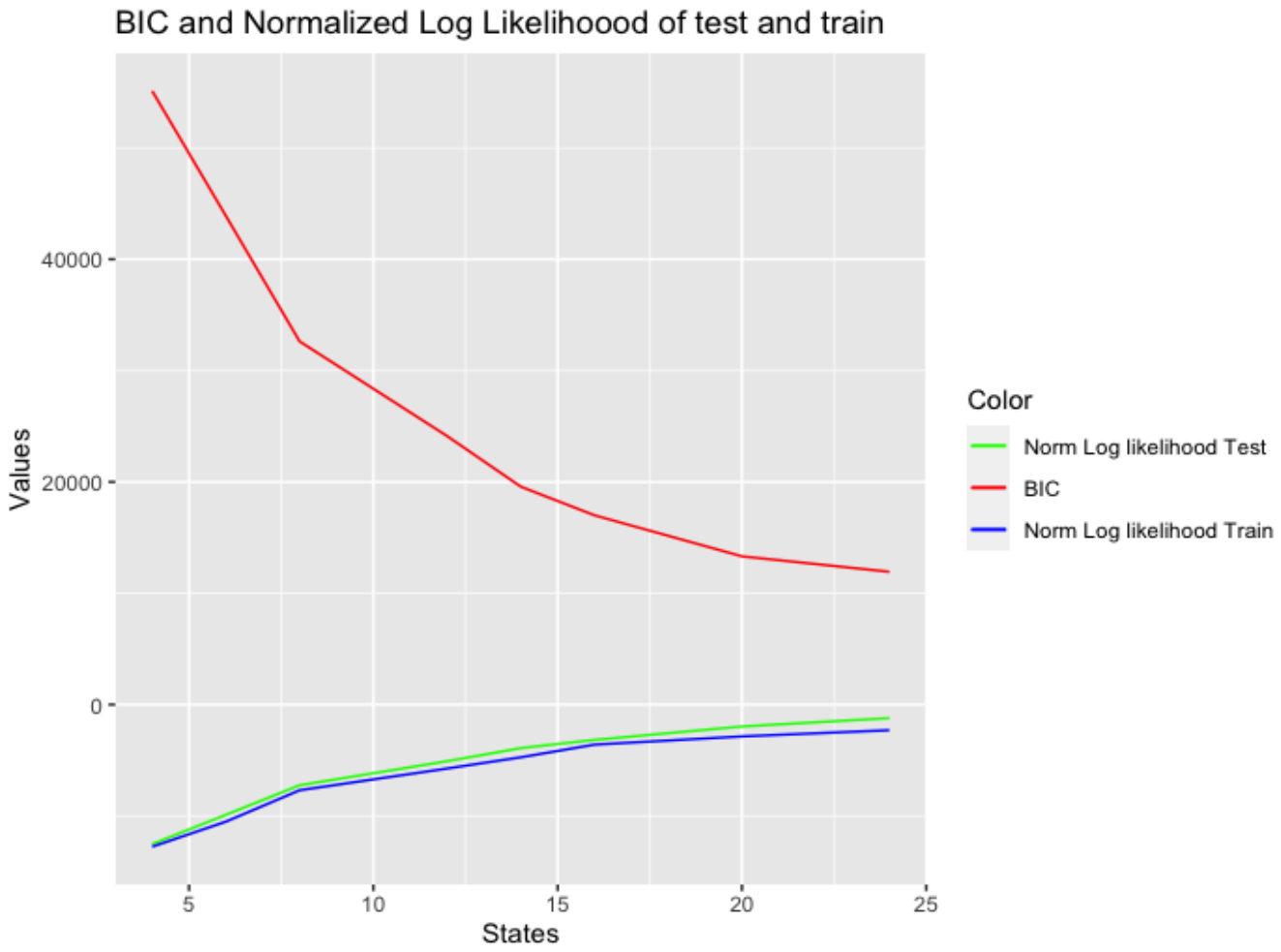

Figure 5 *Log-Likelihood values for train and test data and BIC for a given number of states*

5. Anomaly Detection

Anomaly Detection is the identification of unexpected events, observations, or items that differ significantly from standard behaviours or patterns [4]. Anomalies in data are also called standard deviations, outliers, noise, novelties and exceptions.

5.1 Datasets with Injected Anomalies

We read the three dataset files with anomalies in them and filtered them out so that they contain values only for Fridays from 18:00 to 21:00. The data was scaled before making the Hidden Markov Models for each of the three datasets with anomalies. The parameters or features chosen for the multivariate model were the same as the ones in Part 2. All the 3 data sets were trained on the selected fitted model of the trained data from Part 2. This gives us the Log likelihoods of 3 datasets with anomalies.

5.2 Degree of Anomalies Present

The Log Likelihoods for the 3 datasets with anomalies came out to be =>

Data set 1 with anomalies => -10936.82

Data set 2 with anomalies => -19242.80

Data set 3 with anomalies => -10936.68

According to the results, Data set 2 has higher anomalies as compared to Data set 1 and Data set

6. Reinforcement Learning Paradigm

Reinforcement Learning is a paradigm of Machine Learning and this paradigm is used by various AI models. This technique has numerous well-known applications in the field of robotics and natural language processing (NLP). It's also used in various other fields of AI like computer vision systems and healthcare systems.

6.1 Key Characteristics of Reinforcement Learning

Basically Reinforcement Learning is all about learning or training a model based upon the previous experience which is how an agent got rewarded while it went through a change in state. This technique involves simple steps starting with observing the environment, deciding how to respond based on some strategy, then taking an action according to the strategy, then receiving a reward or penalty based upon the action taken in the previous step, then learning from the experience and changing the strategy accordingly, the last step is to keep on iterating until an optimal strategy is found. Q-learning is one of the simple reinforcement learning algorithms that enable the agent to keep on learning from its environment as after taking each action causing a change of state, the agent receives a reward and therefore can keep on taking actions based on previous experience. For each action taken by the agent, a Q-value is stored in the Q-table and this Q-value basically is the magnitude of the kind of rewards the agent got when it did a certain action. A better Q-value implies that there are better chances of getting greater rewards. So, the Q-value is essentially the experience that is being gathered by the agent, used to take better actions for higher rewards and is being stored in the Q-table simultaneously. Q-table is a matrix which is continuously updated after each action of the agent based on its state.

6.2 Advantages of Reinforcement Learning over the classical ML approach our team used for this term project

As we have used the Hidden Markov Models technique for our project, Reinforcement Learning takes a different approach to solve problems like extracting anomalies to know about any intrusions happening. The fact that reinforcement learning learns from its experience gives it a big edge over our supervised learning technique that uses HMM. The Q-learning method of reinforcement learning can detect anomalies in the data by interacting with the anomalies in the system and gaining experience from the response of its interaction with the anomalies. As anomalies can take different forms, the Q-learning method is way better to detect those anomalies efficiently in comparison to our supervised learning method. The second advantage is that the Q-learning method does not require a fixed set of training data like our HMMs. This is a huge advantage in the terms of application in real-world AI systems as sometimes it's very difficult to obtain a well-labelled set of data, this limits the supervised learning methods. The Q-learning method can simply learn from the experience without the presence of labelled data being a requirement.

7. Conclusion

7.1 Report Results

This project taught us how to use Principal Component Analysis and Hidden Markov Model to learn more about anomalies. We were able to select the best features for multivariate HMM convergence by observing the PCA's variation and loading values. Furthermore, we learned how to divide the data into training and testing and how to choose the best Hidden Markov Model. We used the fitted model of train data to find the log-likelihood of test data. By selecting the best Hidden Markov Model, we were able to conclude that Data set 2 had higher anomalies than data set 1 and data set 3.

7.2 Lessons Learned

The biggest challenge we faced while working on this project was training the Hidden Markov Models. We had the following warning message with every state:

1: In em.depmix(object = object, maxit = emcontrol$maxit, tol = emcontrol$tol, :

Log-likelihood decreased on iteration 6 from 516019.030988919 to 471007.092347213

This further showed us in the graph that our multivariate HMM was not converging even when we selected the features with higher loading values from PCA1. To tackle this, we had to do so many trials on different features. We also changed the set.seed() value repeatedly and ran the code multiple times. Eventually, we were able to get the converged multivariate HMM with the right chosen features. Moreover, we also faced some issues while finding the log-likelihood of test data using fitted models of trained data. But after carefully examining/debugging our code,

we were able to find the pinpoint of the code that was causing an issue. In the end, with teamwork and dedication, we were successfully able to complete the project and code as needed.

Works Cited

[1] Eddy, Sean R. “What Is a Hidden Markov Model?” *Nature Biotechnology*, vol. 22, no. 10, 2004, pp. 1315–1316., https://doi.org/10.1038/nbt1004-1315.

[2] Ghafir, Ibrahim, et al. “Hidden Markov Models and Alert Correlations for the Prediction of Advanced Persistent Threats.” *IEEE Access*, vol. 7, 2019, pp. 99508–99520., https://doi.org/10.1109/access.2019.2930200.

[3] “Principal Component Analysis for Special Types of Data.” *Principal Component Analysis*, pp. 338–372., https://doi.org/10.1007/0-387-22440-8_13.

[4] Ten, Chee-Wooi, et al. “Anomaly Detection for Cybersecurity of the Substations.” *IEEE Transactions on Smart Grid*, vol. 2, no. 4, 2011, pp. 865–873., https://doi.org/10.1109/tsg.2011.2159406.

[5] “What Is Anomaly Detection? Definition & Faqs.” *Avi Networks*, 25 Oct. 2022, https://avinetworks.com/glossary/anomaly-detection/.