Third Group Assignment

The value of this assignment is 7%. Please analyze the electricity consumption dataset available on the course home page using the R language and environment.

This assignment addresses the training of Hidden Markov models for the purpose of **unsupervised intrusion detection** by analyzing stream data from a supervisory control system. For household electricity consumption data, we can generally assume that normal instances occur far more frequent than anomalies; that is, except for the inevitable noise, the datasets considered are largely comprised of "normal" data instances. Thus, the resulting models learned during training should be robust to the relatively few anomalies a dataset may contain (e.g., caused by technical system component malfunctions).

Please complete the tasks described below and submit an electronic copy of your solution through CourSys by Thursday, March 10, 2022.

1. Scale the dataset provided, e.g. by means of the R command **`scale()`**. For one of the response variables, `Global_active_power`, `Global_reactive_power` or `Global_intensity`, determine a time window for a specific weekday that shows a clearly recognizable electricity consumption pattern over a time period of not less than 120 and not more than 240 minutes. Extract the same time window for each week of the dataset and concatenate the extracted time windows to build a dataset for the training of HMMs.

2. Use your training dataset for training a number of univariate HMMs that each have a different number of states across a range from not less than 3 states to not more than 16 states. For each HMM, compute the log-likelihood measure on the training dataset. In addition, compute the *Bayesian information criterion*[1], or BIC, as a measure of the complexity of your model. The goal is to find the intercept of the two plots for log-likelihood and BIC values respectively so as to determine the best model (avoiding overfitting). You may not need to train HMMs for each and every number of states within the range by making smart choices.

   How likelihood and BIC measures are used:

   Likelihood of a sequence of observations for a given model indicates how likely the model can produce this sequence of observations. In other words, a high likelihood means that the model is a good representation of the dataset and can produce the observation sequence with a relatively high chance. It is important to point out that we are actually

[1] The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC). When fitting models, it is possible to increase the likelihood by adding parameters but doing so may result in overfitting.

calculating **log-likelihood** (instead of likelihood), which is always a negative number. On the other hand, BIC is a measurement of the complexity of our model. Increasing the number of states of an HMM, may increase log-likelihood but, at the same time, also increases the complexity of the model. This trade-off is inescapable. We aim to strike a sensible balance between log-likelihood and BIC values for any good model.

In this project we train HMMs using the **depmixS4** package in R. Training an HMM in depmixS4 has 2 steps: first, you specify the HMM parameters, and second, you fit the model to the data. One important parameter you should specify in the first step is the number of HMM states as explained earlier.

Please note that you should use the "**ntimes**" command in the training process, because you are training an HMM for *n* occurrences (52) of a specific time window.

Below are the commands you need to train an HMM and get log-likelihood and BIC values as well as the links to the documentations of the depmixS4 package.

```
model <- depmix(response =, data =, nstates =, ntimes = )
fitModel <- fit(model)
summary(fitModel)
```

https://cran.r-project.org/web/packages/depmixS4/vignettes/depmixS4.pdf
https://cran.r-project.org/web/packages/depmixS4/depmixS4.pdf

Please submit a report with your solution and the R code through CourSys by March 10.

Thank you!