CMPT318 Fall 2024

2024-10-30

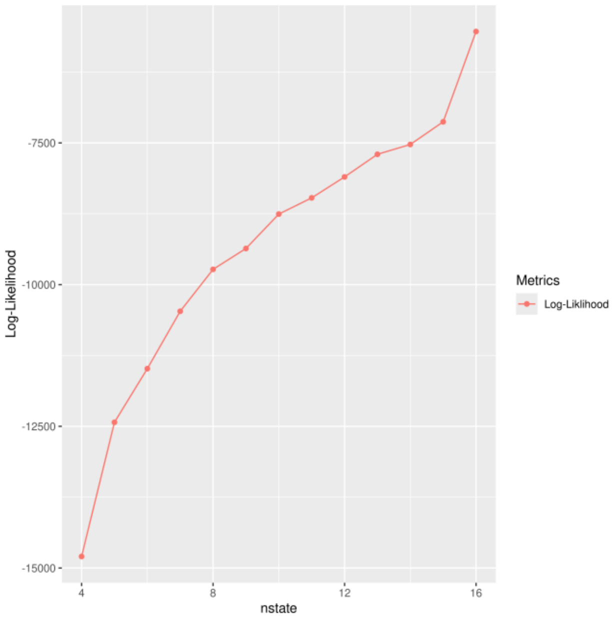
Assignment 3

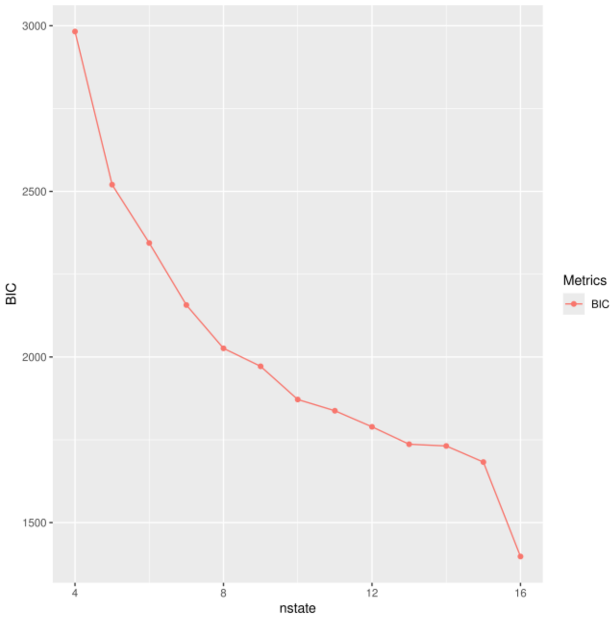
Group 8

Question 1: Continuous HMM

The preprocessed dataset from assignment 1 was used to train multiple univariate HMMs on **Global_active_power** using the *depmixS4* library. A time window of 8 hours from 9 am to 5 pm Monday was chosen which gives 481 observations per 52 weeks.

Each HMM was configured with a different **nstate** parameter from 4 to 16 and the following plots show the change in Log-likelihood and BIC over the number of states. From the graph, we observe a maximum Log-likelihood of –5535.1058 and a minimum BIC of 13976.6925 with 16 states.





Question 2: Discrete HMM

Theory Question: Positive Log-likelihood values with continuous HMM variables

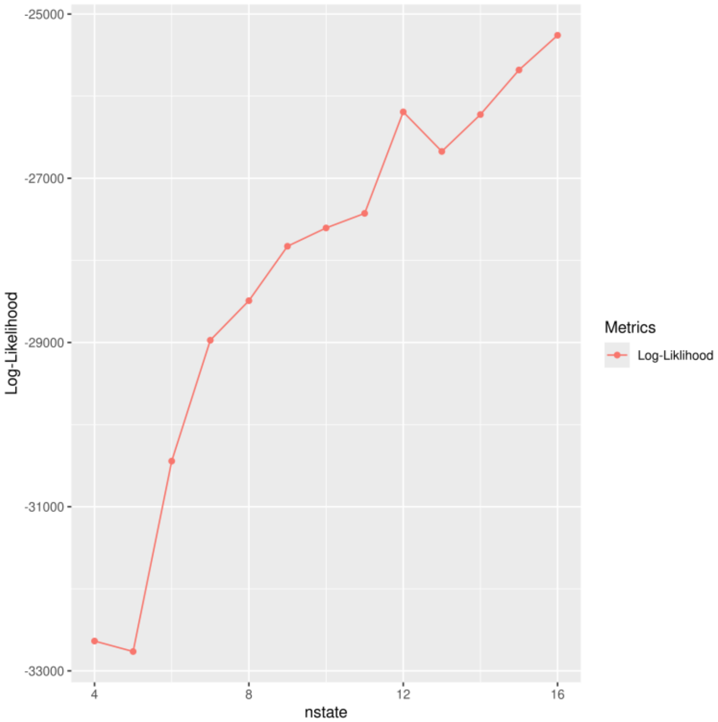
The appearance of positive log-likelihood values when an HMM is dealing with continuous variables is caused by the probability distribution used in the calculation of the Log-likelihood in *depmixS4.* For continuous variables, the probability density function (PDF) is used to compute the probability of an observation sequence, and since the probability density of a particular observation can be greater than 1, there is a chance that the log and subsequently the sum of logs of the probabilities can be positive. By contrast, for discrete variables, the probability mass function (pmf) is used, which for a given observation should always be less than or equal to 1, thus taking the log and the subsequent sum will be no greater than 0.

Practical Task: Address positive Log-likelihoods

To address positive log-likelihoods, the preprocessed dataset data frame was copied and the continuous values for **Global_active_power** in the copied data frame were rounded to the nearest half-decimal (0.5).

Model Training: Discrete HMM variables

Using the discretized **Global_active_power** values from the previous step, another 13 HMMs were trained with number of states ranging from 4 to 16 and the *multinomial*() distribution. Model fitting took considerably longer compared to using continuous variables, possibly because the pmf of a multinomial distribution is harder to compute. The plot below shows the log-likelihoods of the newly fitted models.



The maximum log-likelihood and minimum BIC were (-25257.9336, 54556.5844) using 16 states. Comparing the log-likelihoods of the discrete variable versus the continuous variable from question 1, we can see that the HMMs using the continuous variable have a much higher likelihood and lower BIC than the discrete variable HMMs given the same number of states. In this case, discretizing the **Global_active_power** has caused a decrease in the HMM's ability to capture the underlying patterns in the data, possibly because the rounding of the values to the nearest half-integer does not produce a good discretization of the continuous variable. As a small experiment, the rounding of **Global_active_power** was modified to the nearest quarter-decimal (0.25) and the nearest whole-integer, and traning HMMs for **nstate** = 4, the log-likelihood and BIC were better when rounding to the nearest whole-integer. This result suggests that poor discretization of the data will result in high loss of information, reducing HMM accuracy, while good discretization of the data will result in minimal loss of information.