# 

Group member:

Shiya Ou, Tunan Jia, Qiandan Deng

Student number:

1003044449, 1002206463, 1002928905

,1002206463

Analysis on Hidden Markov Model

- Data Science for Risk Modelling Project 2

**Nov 8, 2020**

**Table of Contents**

[**Introduction** 3](#_Toc55476899)

[**Methodology** 3](#_Toc55476900)

[**2.1 Data preparation** 3](#_Toc55476901)

[**2.2 Modeling for one sequence** 3](#_Toc55476902)

[**2.3 Modeling for 50 sequences** 6](#_Toc55476903)

[**2.4 Model comparison** 8](#_Toc55476904)

[**Analysis** 8](#_Toc55476905)

[**3.1 Exploration of Random Variables** 8](#_Toc55476906)

[**3.1.1 Explore the univariate empirical distribution** 8](#_Toc55476907)

[ Distribution of X1 -Univariate distribution 8](#_Toc55476908)

[ Distribution of X2 -Univariate distribution 9](#_Toc55476909)

[ Distribution of X1 -Univariate distribution 9](#_Toc55476910)

[**3.2 EM- Algorithm 9**](#_Toc55476911)

[**3.3 Bayesian Information Criterion (BIC) 9**](#_Toc55476912)

[**3.4 Bootstrapping 10**](#_Toc55476913)

[**Conclusion 13**](#_Toc55476914)

[**References 14**](#_Toc55476915)

[**Appendix 15**](#_Toc55476916)

## **Introduction**

Hidden Markov Model (HMM) is a statistical model in which the system being modeled has invisible states. HMM assumes that in a given process, there is a relationship between a hidden objective Y and the following objective Y. Although each Y is unobservable, its behavior depends on X, which are observable. Our goal is to learn Y from X.

In this project, we are going to implement the Hidden Markov Model on a given dataset. The dataset includes 50 sequences of , and each is a 3-d data observation associated to at the step m. We assume each Y is a generative mixture model depend on three variables ,, and . They are mutually independent conditional on Y. is normal distributed, is exponential distributed and is a categorical distribution. We use to denote the prior probability on the initial Y in each sequence and matrix A to denote the transition probability matrix such that for each state, Those parameters contain all the information needed to necessarily determine the model. The objective is to find a best model with the corresponding parameters in a specific confidence level.

To accomplish this, firstly, we select a small integer as the number of regimes in the hidden states, then apply the Expectation Maximization method to find the parameters according to the information contained in the original dataset. Secondly, we’re going to try different models with different choices of regimes, then apply Bayesian Information Criteria to penalize the number of parameters and select a best model. Finally, to test the reliability of our results, we use bootstrap method to find a confidence band on those parameters and check whether the estimation lie in this interval.

## **Methodology**

### **2.1 Data preparation**

Before implementing the algorithms, we need to prepare the raw data so that it is ready to be modeled. The given data have 50 sequences with different length, representing the time step each process has experienced. Since the length differs among all sequences, we need to manually create a rectangle to store the information in a consistent order. To accomplish this, we establish a matrix, whose dimension is decided by the maximum length in the original sample. In this matrix, each column contains information in each sequence. But since not every sequence experiences the same amount of time, in this situation, we fill the missing times with zero, meaning there is no information to reference for that time, in the given sequence. In such a way, the algorithm According to the 50 sequences, the lengths of each sequence are different. We create three such rectangle, for each storing the information given by respectively.

### **2.2 Modeling for one sequence**

After preparing the data for to store all the information in each sequence. We first show the method to deal with single sequence . We apply the Expectation-Maximization (EM) algorithm for all parameters estimates in each sequence. The EM algorithm consists two main steps:

1. E-step: Derive the expectation over the hidden states by using posterior distribution from the initial observed data, mainly compute the expected completed log-likelihood. In HMM model, the and the smoother ( can be obtained through the forward-backward algorithm.
2. M-step: Maximize the log- likelihood over the parameters in order to get a new updated parameter.

Text

Description automatically generated

### Text Description automatically generated

### Text, letter Description automatically generated

### Text, letter Description automatically generated

Text, letter

Description automatically generated

### **2. 3 Modeling for 50 sequences**

Since the original sample has 50 time-sequence, we need to aggregate all observed information in our model. Since all sequence are independent with each other, the loglikelihood function is the sum of every sequence’s loglikelihood. The updated parameters for 50 sequences are given in the following:

A picture containing table

Description automatically generated

Parameters

### **2.4 Model Selection**

For the purposes of determining the optimal number of hidden states and the parameters, the Bayesian information criterion (BIC) is applied. BIC score is a criterion for model selection among models with different possible regimes. The lower BIC score indicates the better fit model and we are able to decide the final label number of . The BIC can be formally calculated for a given model HMM model M as:

where the loglikelihood function is given by:

, indicates the log-likelihood function with optimal number and k is the number of hidden states

is the number of free parameters:

Diagram

Description automatically generated

n is number of data we overserved from all 50 sequences.

## **Analysis**

### **3.1 Exploration of Random Variables**

### **3.1.1 Explore the univariate empirical distribution**

### Distribution of X1 -Univariate distribution

The random variable is a continuous random variable that follows a normal distribution. Since we have 50 sequences, we randomly choose 3 sequences out of 50. We plot the KDE distribution of from the 1st, 10th, 50th sequence by the given data set. The plots are shown below: Figure 1, we notice that each histogram shows two normal distributions with local maximum around x= 0.3, 0.25; 0.5,0.25; 0.35, 0.25. From those 3 sequences, we can clearly observe the peak of the data occurs around 0.45-0.6, it emphasizes that the concentration of data in that area.

Chart, histogram

Description automatically generated

Density

*Figure 1. Histogram and KDE distribution of X1*

### Distribution of X2 -Univariate distribution

The random variable is a continuous random variable that follows an exponential distribution. We plot the KDE for of the 1st, 10th, 50th sequence, the plot is illustrated in Figure 2.

Chart, histogram

Description automatically generated

Density

*Figure 2. Histogram and KDE distribution of X2*

### Distribution of X1 -Univariate distribution

Same as what we expect from the data set, the random variable of is a categorical variable and it has 5 categories with a multinomial distribution. From Figure 3 below, we plot the histogram for of the 1st, 10th, 50th sequence, we can see that most of the data points are in the category 3 from these 3 sequences.

### Chart, bar chart Description automatically generated

*Figure 3. Histogram of X3 on a categorical distribution*

### **3.2 Implement EM- Algorithm**

We implement the EM-algorithm derived as above using 50 sequences of data, in order to update all the parameters for each model with different number of hidden states. The attached code implementation would address this question.

### **3.3 Bayesian Information Criterion (BIC)**

BIC is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. We will use the BIC method to determine the optimal number of hidden states and the BIC is defined by the formula given above.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Regimes | 2 | 3 | 4 | 5 | 6 | 7 |
| **BIC score** | 7669.8 | 7005.4 | 7057.1 | 7117.5 | 7187.4 | 7253.4 |

*Figure 4. BIC score at different number of regimes*

Chart, line chart

Description automatically generated

*Figure 5. BIC score plot at different number of regimes*

From the plot, we can conclude that the optimal number of hidden states is 3 regimes because of the lowest BIC score among other regimes (BIC = 7005.4).

By testing, the values of each parameter will converge after iterating 50 times. To make sure the precision about our results, we iterated 100 times and the optimal parameters are illustrated in Figure 6 below

Table

Description automatically generated

*Figure 6. Optimal Parameters after Iteration*

### **3.4 Bootstrapping**

Bootstrap is an important method in statistical modeling. For empirical bootstrap, we can resample the original dataset to obtain a new sample, then use the new sample to obtain an estimation on the parameters. Repeating this process for thousands of times will give us a distribution on the parameter, and we’re able to apply normal test to derive a confidence interval. However, in this report, we’ve already given the explicit model, our focus is to test the availability about the derived parameter with the use of bootstrap. In the previous section, we’ve successfully selected a model and its parameters. We use this estimation to simulate 20 new samples according to the HMM process, with each sample has 50 sequences and each sequence has the same length as the original data. Then apply the EM algorithms on the new samples to obtain a new set of parameters. At the end, we collected 20 values on each parameter. In order to obtain a confidence interval, since 20 is a relatively small number, we prefer to use T-test instead of Z-test. On a 95% confident level, the interval is given by:

Where t is the t-statistics with 19 freedoms and 95% significant level.

The CI results can be illustrated in the following table:

Parameters confidence intervals (CI) at 95% significance level:

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameters** | | **Estimate** | **CI** |
|  | Y=1 | 0.1515 | [0.106, 0.167] |
| Y=2 | 0.6602 | [0.54, 0.6755] |
| Y=3 | 0.1882 | [0.169, 0.339] |
|  | Y=1 | -1.0012 | [-1.00362, -0.99945] |
| Y=2 | 0.013176 | [-0.2286, 0.065] |
| Y=3 | 1.0230 | [0.96, 1.027] |
|  | Y=1 | 0.1016 | [0.0987, 0.1017] |
| Y=2 | 1.03640 | [0.84, 1.06] |
| Y=3 | 0.4958 | [0.48, 0.561] |
|  | Y=1 | 0.9886 | [0.974, 0.998] |
| Y=2 | 5.0317 | [4.26, 5.194] |
| Y=3 | 19.3874 | [17.597, 19.924] |
|  |  | 0.79337 | [0.746195, 0.80576] |
|  | 0.092409 | [0.07661, 0.13117] |
|  | 0.11421 | [0.11100, 0.129257] |
|  |  | 0.50052 | [0.47666, 0.51042] |
|  | 0.19463 | [0.181496, 0.21652], |
|  | 0.30484 | [0.296307, 0.31859] |
|  |  | 0.1118 | [0.101367, 0.12290] |
|  | 0.18642 | [0.161556, 0.19219] |
|  | 0.70177 | [0.70, 0.71967] |
| **Parameters** | | **Estimate** | **CI** |
|  |  | 02 | [0.098, 0.104] |
|  | 195 | [0.19, 0.196] |
|  |  | [0.702, 0.709] |
|  |  | [-0.000004, 0.00024] |
|  |  | [-8.77e-06, 0.00156] |
|  |  |  | [-0.0028, 0.0157] |
|  |  | [-0.0077, 0.0364] |
|  |  | [0.236, 0.326] |
|  |  | [0.399, 0.496] |
|  |  | [0.22, 0.279] |
|  |  | 0.2024 | [0.192, 0.21] |
|  | 0.2089 | [0.19, 0.21] |
|  | 0.19067 | [0.188, 0.198] |
|  | 0.2101 | [0.207, 0.229] |
|  | 0.1877 | [0.182, 0.193] |

From the table above, we observe that ALL estimated parameters lie in the corresponding confidence intervals, which suggested that we are confident enough about our results.

## **Conclusion**

In this report, we implemented EM algorithms to find the unobserved patterns in the Hidden Markov Model, and use bootstrap to prove the availability about our results. According to the BIC criterion, we select t the model with three hidden states as the optimal model, labelling them as . After updating and optimization, we conclude that the transaction probability matrix is given by:

In terms of the generative mixture model, we also investigate the process with variables , which follow different distributions given . We conclude the followings:

* )
* 5.03)
* 19.388)

## 

## **References**

1. Yen, Lorna. “An Introduction to the Bootstrap Method.” *Medium*, Towards Data Science, 28 Jan. 2019, towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60.
2. “Bayesian Information Criterion.” *Wikipedia*, Wikimedia Foundation, 12 Oct. 2020, en.wikipedia.org/wiki/Bayesian\_information\_criterion.
3. “PH717 Module 6 - Random Error.” *Confidence Intervals for Sample Size Less Than 30*, sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module6-RandomError/PH717-Module6-RandomError11.html.
4. 13.0 Bootstrap Confidence Intervals. Retrieved from http://www2.stat.duke.edu/~banks/111-lectures.dir/lect13.pdf

## **Appendix**

* Perform forward and backward filter

Graphical user interface, text, application, email

Description automatically generated

* Implement the E-step

Graphical user interface, text, application

Description automatically generated

* Implement the M-step

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

* Implement the EM

Graphical user interface, text, application, email

Description automatically generated

* Perform simulate data

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated