

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**HIDDEN MARKOV MODELS BASED ON
DIFFERENT DEPENDENCY ASSUMPTIONS**

by
Özgür DANIŞMAN

January, 2021
İZMİR

HIDDEN MARKOV MODELS BASED ON DIFFERENT DEPENDENCY ASSUMPTIONS

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of
Philosophy in Statistics**

**by
Özgür DANIŞMAN**

**January, 2021
İZMİR**

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**HIDDEN MARKOV MODELS BASED ON DIFFERENT DEPENDENCY ASSUMPTIONS**” completed by **ÖZGÜR DANIŞMAN** under supervision of **ASSOC. PROF. DR. UMay UZUNOĞLU KOÇER** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

.....
Assoc. Prof. Dr. Umay UZUNOĞLU KOÇER

Supervisor

.....
Prof. Dr. Burcu HÜDAVERDİ

Thesis Committee Member

.....
Assoc. Prof. Dr.
Femin YALÇIN KÜÇÜKBAYRAK

Thesis Committee Member

.....
Prof. Dr. Muhammet BEKÇİ

Examining Committee Member

.....
Assoc. Prof. Dr. Halil TANIL

Examining Committee Member

.....
Prof. Dr. Özgür ÖZÇELİK

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Assoc. Prof. Dr. Umay UZUNOĞLU KOÇER, for the patient guidance, encouragement and advice she has provided throughout my time as her student. I have been extremely lucky to have a supervisor who cared so much about my research. Her positive outlook and immense knowledge in my research inspired me and gave me confidence. Her guidance and careful editing contributed enormously to the production of this thesis.

I would like to thank to Prof. Dr. Burcu HÜDAVERDİ and Assoc. Prof. Dr. Femin YALÇIN KÜÇÜKBAYRAK for their suggestions and constructive comments as a member of my thesis committee.

Special thanks to Dr. Hanifi Okan İŞGÜDER for his invaluable advices, endless support, and friendship.

I would like to acknowledge to TÜBİTAK (The Scientific and Technical Research Council of Turkey) for the financial support via “2211-E Direct National Scholarship Programme for Ph.D. Students” during my Ph.D. education.

I would like to send my most sincere and loving thanks to my life partner Fahriye AYRIÇ for her endless support and for being with me every time throughout the thesis.

Finally, I would like to express my deepest love, respect, and gratitude to my family for their continuous and unparalleled love, help and support. They gave me the opportunities and experiences that have made me who I am. Their selfless encouragement is the most important factor to explore new directions in life and to seek my own density. This thesis would not have been possible if not for them, and I dedicate this milestone to them.

Özgür DANIŞMAN

HIDDEN MARKOV MODELS BASED ON DIFFERENT DEPENDENCY ASSUMPTIONS

ABSTRACT

Hidden Markov models are widely used to model the probabilistic structures with latent random variables. The main assumption of hidden Markov models is that; observations are conditionally independent and identically distributed random variables. There may exist some cases where this assumption may not be valid in practice. That is, an observation that occurs in the current state may depend on the previous observation symbol that occurred in the previous state. In this thesis, two types of hidden Markov models are introduced which differ from the classical hidden Markov model based on different first-order Markov dependence assumptions. The introduced models are capable of capturing a possible first-order Markov dependence between the successive observations or successive system informations. They can provide better representation for the appropriate real-life problems where, if the observations have some conditional dependencies among them.

The two proposed models are defined with their assumptions and using appropriate notation. Modifications to the algorithms used for parameter estimates and hidden state sequence estimates are explained. In addition, an experimental study is conducted to show the performance of the introduced models compared to the classical hidden Markov model. According to the results of the experimental study, the proposed models outperform in generated observation sequences that have appropriate assumptions. Besides, two different case studies are conducted namely the occurrences of strong earthquakes and daily stock prices. They are modelled with both the classical hidden Markov model and the proposed models, and the results are compared.

Keywords: Hidden Markov model, dependent observations, parameter estimation, Baum-Welch algorithm, strong earthquake, stock closing price

FARKLI BAĞIMLILIK VARSAYIMLARINA DAYALI SAKLI MARKOV MODELLERİ

ÖZ

Saklı Markov modelleri, saklı rastgele değişkenler içeren olasılıksal yapıları modellemede yaygın olarak kullanılmaktadır. Saklı Markov modellerinin temel varsayımı gözlemlerin koşullu bağımsız ve özdeş dağılıma sahip rastgele değişkenler olmalarıdır. Bu varsayımın pratikte geçerli olmayabileceği bazı durumlar olabilir. Yani, mevcut durumda üretilen bir gözlem, önceki durumda meydana gelen gözleme bağlı olabilir. Bu tezde, farklı birinci dereceden Markov bağımlılık varsayımlarına bağlı olarak klasik saklı Markov modelinden farklı iki tür saklı Markov modeli sunulmuştur. Sunulan modeller, ardışık gözlemler veya ardışık sistem bilgileri arasındaki olası birinci dereceden Markov bağımlılığı yakalayabilir. Eğer gözlemler arasında bazı koşullu bağımlılıklar varsa, bu modeller uygun gerçek hayat problemlerini daha iyi temsil edebilirler.

Önerilen iki model, varsayımlarıyla ve uygun notasyon kullanılarak tanımlanmıştır. Parametre tahminleri ve durum zinciri tahminleri için kullanılan algoritmalarda yapılan modifikasyonlar açıklanmıştır. Ayrıca, sunulan modellerin klasik saklı Markov modeline göre performansını değerlendirmek için deneysel bir çalışma yapılmıştır. Deneysel çalışmanın sonuçlarına göre, sunulan modeller, uygun varsayımlara göre üretilen gözlem serilerinde daha iyi performans göstermektedir. Bunlara ek olarak, kuvvetli deprem oluşumları ve günlük hisse senedi fiyatları olmak üzere iki farklı vaka çalışması gerçekleştirilmiştir. Hem klasik saklı Markov modeli hem de sunulan modeller ile modellenerek sonuçlar karşılaştırılmıştır.

Anahtar kelimeler: Saklı-Markov modeli, bağımlı gözlemler, parametre tahmini, Baum-Welch algoritması, şiddetli deprem, hisse senedi kapanış fiyatı

CONTENTS

	Page
Ph.D. THESIS EXAMINATION RESULT FORM	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZ	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
CHAPTER 1 - INTRODUCTION.....	1
CHAPTER 2 – HIDDEN MARKOV MODEL	7
2.1 Basic Preliminaries.....	7
2.2 Model Definition	10
2.3 Fundamental Problems of HMM.....	13
2.3.1 Evaluation Problem	13
2.3.2 Learning Problem	17
2.3.3 Decoding Problem	23
2.4 Illustrative Example	25
CHAPTER 3 – MARKOV-EMITTED HIDDEN MARKOV MODEL	37
3.1 Model Definition	38
3.2 Fundamental Problems of MEHMM.....	42
3.2.1 Evaluation Problem	42
3.2.2 Learning Problem	45
3.2.3 Decoding Problem	50
3.3 Experimental Study	52
3.4 A Case Study on Stock Price Trend Probability Forecasting.....	60
3.4.1 Model Selection	63

3.4.2 Bootstrap Confidence Intervals for the 2-State-2-Obs. MEHMM Parameters.....	69
3.4.3 Results	71
CHAPTER 4 – BINARY-DEPENDENT HIDDEN MARKOV MODEL.....	73
4.1 Model Definition	74
4.2 Fundamental Problems of BDHMM	78
4.2.1 Evaluation Problem	79
4.2.2 Learning Problem	81
4.2.3 Decoding Problem	86
4.3 Experimental Study	88
4.4 An Application for the Strongest Earthquake Occurences.....	95
4.4.1 Model Selection	100
4.4.2 Bootstrap Confidence Intervals for the 2-State-3-Obs. BDHMM Parameters.....	106
4.4.3 Results	108
CHAPTER 5 - CONCLUSION.....	110
REFERENCES.....	113

LIST OF FIGURES

	Page
Figure 2.1 Graphical illustration of HMMs	10
Figure 2.2 Flow chart of the forward probabilities for time t	14
Figure 2.3 Flow chart of the backward probabilities at time t	16
Figure 3.1 Graphical illustration of MEHMMs	39
Figure 3.2 Time-series plot of stock price closing values of Microsoft Corporation.	61
Figure 3.3 Time-series plot of the first-difference values.....	62
Figure 3.4 Time-series plot of the relative change (%) values	62
Figure 4.1 Graphical illustration of BDHMMs.....	75
Figure 4.2 Distribution of the decoded hidden state sequence between 1845 and 2010	104
Figure 4.3 Distribution of the actual observation sequence between 1845 and 2010	104

LIST OF TABLES

	Page
Table 3.1 Complete parameter sets of HMM and MEHMM.....	42
Table 3.2 Types of observation sequences.....	53
Table 3.3 Results obtained by modeling the observation sequences which are generated based on MEHMM assumptions	54
Table 3.4 Results obtained by modeling the observation sequences which are generated based on HMM assumptions	56
Table 3.5 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of MEHMM.....	58
Table 3.6 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of HMM	58
Table 3.7 Comparison of the number of successes and the expected number of more likely model for MEHMM	59
Table 3.8 Types of observation sequences.....	64
Table 3.9 Comparison of the models	64
Table 3.10 Matching of hidden decoded states and observation sequence.....	67
Table 3.11 Estimate values and estimated standard deviations of the parameter $\mathbf{B}$..	70
Table 3.12 Estimate values and estimated standard deviations of the parameter $\mathbf{P}$..	71
Table 3.13 Estimate values and estimated standard deviations of the parameter ϕ ..	71
Table 4.1 Complete parameter sets of HMM and BDHMM	78
Table 4.2 Types of observation sequences.....	89
Table 4.3 Results obtained by modeling the observation sequences which are generated based on BDHMM assumptions.....	90
Table 4.4 Results obtained by modeling the observation sequences which are generated based on HMM assumptions	91
Table 4.5 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of BDHMM	92
Table 4.6 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of HMM	93

Table 4.7 Comparison of the number of successes and the expected number of more likely model for BDHMM.....	94
Table 4.8 Types of observation sequences.....	100
Table 4.9 Comparison of the models	101
Table 4.10 Matching of hidden decoded states and observation sequence.....	104
Table 4.11 Estimate values and estimated standard deviations of the parameter Ψ	107
Table 4.12 Estimate values and estimated standard deviations of the parameter ϕ	108



CHAPTER 1

INTRODUCTION

In probability theory, a *stochastic process* is usually defined as a family of random variables. It has quite wide application area especially in Statistics and Mathematics. Assume that $X(t)$ represents the state of the process at time t , where t belongs to an index set T . Depending on how the index set T is defined, two different types of stochastic processes can occur. If T is a countable or finite set, it is called as a discrete-time stochastic process. If T is a continuous set, it is called as a continuous-time stochastic process.

Markov processes (MP), which satisfy the Markovian property, are one of the most popular type of stochastic processes. The process makes transition between the states and, if the conditional transition probability of next state depends only on the current state, it is called the Markov process. Thus, the entire history of the process is no more helpful to make a probabilistic statement about the future behavior of the Markov process. In the literature, Markov process is used for modeling in many different settings such as; semi Markov process, Markov decision process, Markov random field, tolerant Markov model or hidden Markov model. In this thesis, the *hidden Markov model*, which has applications in many different areas in the literature, will be discussed in detail.

Hidden Markov Model (HMM) is a statistical modeling technique to capture hidden information from observable random variables. HMM was first introduced by Baum & Petrie (1966) and further developed by Rabiner (Baum, 1972; Rabiner, 1989). It is a specific form of dynamic Bayesian networks which are based on the Bayes' theory. Besides, HMMs can be considered as an extension of Markov models and can be defined as a two-dimensional stochastic process. The first stochastic process is related to the states and it is modeled by finite-state Markov chains. The second stochastic process, on the other hand, is related to the observations emitted from the states of the Markov chain. Observations are conditionally independent and identically distributed (i.i.d.) random variables. Each observation depends only on the related hidden state in

which it is emitted. The observations appear explicitly whereas the states are not usually observable apparently. Hence the name ‘hidden Markov model’ is used. Hidden Markov models have been successfully applied to variety of areas such as speech recognition (Rabiner, 1989; Juang & Rabiner, 1991), text recognition (Bose & Kuo, 1994; AlKhateeb, Ren, Jiang, & Al-Muhtaseb, 2011), bioinformatics (Durbin, Eddy, Krogh, & Mitchison, 1998), climatology (Zucchini & Guttorp, 1991), signal processing (Bryan & Levinson, 2015; Rasku, Juhola, Toppila, & Pyykkö, 2007), and seismology (Granat & Donnellan, 2002).

The simplicity of the assumptions of HMM encourage researchers to study on this modeling technique more often. However, it is possible that these assumptions may be inappropriate for some problems. For some cases, it is impossible to assume either the observations are conditionally independent or conditionally i.i.d. random variables among themselves. Dependency structure may occur in different ways between observations. On the other hand, when transitions between hidden states occur, the value of the next hidden state may also depend on the previous observation. These kinds of dependency structures, which are not covered by a classical HMM, have required the development of different expansions or modifications in HMM, in terms of their assumptions. In this thesis, based on this motivation, two different types of HMMs, which are called Markov-emitted hidden Markov model (MEHMM) and binary-dependent hidden Markov model (BDHMM), have been developed under the assumption of Markov dependence between the observations.

In the literature, there exist a few studies that extend classical HMM incorporating the information about the history of the system. One extension is that integrating higher-order Markov dependence between hidden states and assuming a dependence between observations (Ching, Fung, & Ng, 2003; du Preez, 1998). An extension of the classical HMM on the molecular biology area is given in which the current observation symbol depends on both the current state and a group of previous observation symbols (Tamposis, Theodoropoulou, Tsirigos, & Bagos, 2018). Furthermore, there are some existing studies in the literature with the assumption of observation dependency. One of the closest work to our study is Li (2005). An HMM is introduced by Li (2005)

where each hidden state of the system depends on not only the preceding state but also the preceding observation symbol. State dependence is still represented by a first-order Markov chain. The current state distribution is conditional on both the previous state and the previous observation. Another study that is close to our work is given by Berchtold (1999) where there is Markov dependence between hidden states and also between observations. Transitions between hidden states occur independently of observations. The system consists of two different Markov chains and their model is a generalization of both Markov process and HMM.

Both MEHMM and BDHMM, introduced in this thesis, have assumptions different from classical HMM. The assumption of conditionally i.i.d. observations is not valid for the proposed models. The main assumption of MEHMM is that each observation depends not only on the related hidden state but also the previous observation. The hidden states are still independent of the observations and make transitions according to a first order Markov chain. On the other hand, the basic assumption of BDHMM is different from HMM and MEHMM. It is assumed that the conditional distribution of the current state and observation depends on both the previous state and previous observation. In other words, a binary dependence is proposed between observations and hidden states. The most recent hidden state and emitted observation are Markov dependent on the previous hidden state and observation. The most important point in the proposed model is that, unlike HMM and MEHMM, transitions between hidden states do not occur independently of observation symbols. Hidden states and observations make interrelated transitions depending on a joint probability distribution. If there is a real Markov dependence between the observations, modeling with classical HMM would not be a correct way and MEHMM or BDHMM will be an alternative way of modeling.

In addition, for the proposed models, dependency structures include more information. That is the conditional distribution depends on much more information relative to classical HMM and this is an advantage of our models in some more areas where HMM is inadequate in terms of its assumptions. A few examples are listed

below for the practical justification of MEHMM and BDHMM for the real-life problems where HMM is used:

- *Stock prices.* Since the market conditions, which cannot be known directly during the day, affect the closing price of a stock, HMM appears as one of the existing modeling techniques. Existing models in the literature usually include different number of hidden states related to market conditions and competition. And also, observation sequences are usually defined as “stock price” or “stock market price” in these models. However, the closing price of a stock in the previous day may have an effect on the same stock's behavior and price for the current day. For this reason, MEHMM and BDHMM may be the models that can be preferred for probabilistic estimation of stock prices.
- *Earthquake occurrences.* Modeling of earthquake data with HMM appears in the literature. In almost all of these studies, the states of the hidden Markov chain are defined as the unobservable stress level of the crust. Observations are usually defined as the magnitude, the location or the inter-event time between two earthquakes. According to classical HMM assumptions, observations such as earthquake magnitudes are assumed to be conditionally i.i.d. and the hidden states transitions occur independently of the observations. However, according to the seismicity theory, there exists a dependence between the earthquake magnitudes. It is very likely that the magnitude of an earthquake and its related stress level, which corresponds to the hidden states, are affected by those of the previous earthquake. In this context, MEHMM and BDHMM can be preferred for modeling and probabilistic forecasting of earthquake data.
- *Flow of Traffic.* MEHMM and BDHMM can be used to estimate the traffic flow for each intersection, if we assume there are T intersections that exist on the main road. In this case, the observation sequence corresponds to the traffic flow rate at any time for the i -th intersection, where $i = 1, 2, \dots, T$. The traffic condition on the main road can be considered as hidden states, due to factors that cannot be observed, such as pedestrians, bikers, or road conditions. It is assumed that

there is a probabilistic transition between hidden states which correspond to the level of traffic condition. In such a problem, the level of traffic condition and the related traffic flow rate at the $(i-1)$ st intersection may probably have an impact on the upcoming i -th intersections condition and flow rate, or there can be a first order Markov dependence only between the successive flow rates. So, MEHMM and BDHMM can be an appropriate modeling technique to estimate the future traffic flow rate.

- *Wind speed and direction.* Another real-life application is about the forecasting of future observations of the speed and also the direction of the wind in a given region for a time interval. The hidden states can be defined as the levels of “weather condition” effect, such that the pressure or the rain intensity on the wind speed and direction. This type of problem can be solved with MEHMM or BDHMM, as both the weather type and the observable wind for the current time unit is possible to have a Markov-emitted or binary Markov dependence with the current weather type and the wind observation of the previous time unit.
- *Soccer game results.* The game results of a football team have been modeled with Markov chains and probabilistic predictions have been made about winning, drawing, or losing future games. Another point of view to this problem may be modeling with MEHMM or BDHMM. Performance and readiness of a team before the game cannot be observed by an outside observer and the hidden states of the process correspond to the level of “team compatibility” before the game and the observations are one of three possible outputs which are winning, drawing or losing. It is very likely that the result of the previous game or poor performance from the previous game will have an impact on the pre-match compatibility level and the result of the current game. In this context, MEHMM and BDHMM can be expected to outperform than classical HMM.

The main contribution of this thesis is two fold:

- The aim is to develop different types of HMMs with extended Markov assumptions, especially the assumption of conditionally i.i.d. observations are not valid for classical HMM.
- The introduced models may be more appropriate for the relevant real life problems in order to minimize the loss of information that will occur from modeling with classical HMM.

Classical HMM is presented in detail in Chapter 2. Model definition, model assumptions, and the complete parameter set of the model is given. The three main problems which are the evaluation problem, the learning problem, and the decoding problem are shown in detail. Then, an illustrative example is given to absorb the functioning of HMM. In Chapter 3, MEHMM, one of the two proposed models, is introduced in detail. Model definition and model assumptions are given. The complete parameter set is defined according to MEHMM assumptions. The three fundamental problems are given again and the modified versions of the forward-backward algorithm, the Baum-Welch algorithm, and the Viterbi algorithm are introduced. An experimental study is conducted to show the performance of the model according to HMM. Besides, an application for the daily stock closing prices is presented. In Chapter 4, BDHMM, the other proposed model, is introduced in detail. Model definition and model assumptions are given. The complete parameter set is defined according to the binary dependency assumptions. The three fundamental problems are given for the new setting and the modified versions of the forward-backward algorithm, the Baum-Welch algorithm, and the Viterbi algorithm are introduced for BDHMM. An experimental study is conducted to show the performance of the model according to HMM. In addition, a case study for the occurrences of the yearly strongest earthquakes in a given time period is presented. In Chapter 5, some discussions, concluding remarks, and possible further studies are given briefly.

CHAPTER 2

HIDDEN MARKOV MODEL

HMMs are discrete-time stochastic processes and consist of two sets of random variables. The first set of random variables is represented by a finite-state Markov chain. The transitions of the Markov chain are invisible (hidden) to an observer outside of the system and the current state is always dependent on the previous state because of the Markovian property. The second set of random variables is related with the observations in each state of the Markov chain. The observations appear explicitly to the observer whereas the states in which the observation occur are “hidden”. The observations are independent from each other and only depend on the current hidden state. It is not possible to observe the states of the system directly and someone is able to record only the observations. Hence the name “hidden Markov model” is used. HMM is a specific form of dynamic Bayesian networks, which is based on the Bayes’ theory. The problem is modeled as a Markov process and the main challenge is to estimate the hidden parameters from the observable parameters. HMMs are different from traditional Markov models due to their assumptions. The observed data is not the actual state of the system for HMMs, but these observations are emitted by the underlying hidden states. On the other hand, the observations correspond directly to the states for traditional Markov models. So, the term “hidden” comes from the first order Markov chain behind the observations.

In this chapter, after giving some basic preliminaries about the MP and the definition of HMM, three fundamental problems of HMM will be discussed. Finally, in section 2.4, the fitting of the model to an observation sequence will be shown with an illustrative example.

2.1 Basic Preliminaries

Markov Process (MP) is a special type of stochastic process which is indexed by time, and with the property that the conditional probability distribution of the current state depends only on the previous state. Markov processes, developed by Russian

mathematician Andrei Markov (1856-1922) in 1906, are one of the most important of all probabilistic models. Therefore, MP has applications in many areas such as financial modeling, flow of traffic, engineering or seismology.

It was stated that the difference of the MP from a classical stochastic process is that the conditional probability distribution of any state depends only on the previous state. This is also known as the Markovian property. Let $\{X(t), t \geq 0\}$ be a Markov process. Then, the mathematical definition of the Markovian property is given as follows:

$$\begin{aligned} P(X_{t+s} = j | X_s = i, X_u = x_u, 0 \leq u < s) \\ = P(X_{t+s} = j | X_s = i), \end{aligned} \quad (2.1)$$

where $s, t \geq 0$ and $0 \leq u < s$.

MPs can be classified depending on whether the parameter space and the state space are discrete or continuous. If the parameter space and the state space are both discrete, then the process is called discrete-time Markov chains (DTMC). Let $\{X_t, t = 0, 1, 2, \dots\}$ be the random variable that represents the state of the process is

$$\begin{aligned} P(X_t = j | X_0 = i_0, X_1 = i_1, X_2 = i_2, \dots, X_{t-1} = i) \\ = P(X_t = j | X_{t-1} = i), \end{aligned} \quad (2.2)$$

where i, j are the states of the process at time t , where $t = 0, 1, 2, \dots$.

Let $\Omega = \{1, 2, 3, \dots\}$ be the state space and p_{ij} the one-step transition probability from state i to state j . Then,

$$p_{ij} = P(X_{t+1} = j | X_t = i) \quad (2.3)$$

for $t = 0, 1, 2, \dots$ and $i, j \in \Omega$.

The transition probabilities for all $i, j \in \Omega$ form one-step transition probability matrix. If the transition probabilities of a DTMC do not change over time and remain constant through time, then the process is said to be time-homogeneous. The one-step transition probability matrix ($\mathbf{P}$) for a time-homogeneous Markov chain is shown below,

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2j} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3j} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{i1} & p_{i2} & p_{i3} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

All possible transition probabilities for a time-homogeneous DTMC are greater than or equal to zero. In addition, the sum of the transition probabilities from one state to other possible states is one. In other words, all elements of the $\mathbf{P}$ are greater than or equal to zero, and the sum of each row is equal to one. That is,

$$p_{ij} \geq 0, \quad i, j \in \Omega; \quad (2.4)$$

$$\sum_{j \in \Omega} p_{ij} = 1, \quad i \in \Omega. \quad (2.5)$$

Although it is widely used and has many advantages for probabilistic modeling, there are some real-life problems where the MP may be limited in terms of its assumptions. Over time, there has been a need to be extended into different types of MPs with extra assumptions. Markov decision process, hierarchical Markov model, tolerant Markov model, semi-Markov model and hidden Markov model are the most used examples of MPs. Each one is a special case of MPs and differs from each other according to their assumptions. The discrete hidden Markov model will be explained in detail in the following section.

2.2 Model Definition

Let v_k be the k th observation and there are M possible emitted observations per state. The system makes a transition and emits one of the observation symbols in the set of $\mathcal{O} = \{v_1, v_2, \dots, v_M\}$. There are N number of hidden states and transitions occur between these states. The hidden state sequence and the observation sequence are finite sets and T is the length of these sets. Let X_t be the random variable that represents the state of the hidden Markov chain at time t , where $X_t \in \Omega$, $t = 0, 1, \dots, T$. Furthermore, let Y_t be the random variable that represents the observation in a state at time t , where $Y_t \in \mathcal{O}$ and $t = 1, 2, \dots, T$. So, the sequence of hidden states and observations are denoted by $\{X_1, X_2, \dots, X_T\}$ and $\{Y_1, Y_2, \dots, Y_T\}$, respectively. Any hidden state $X_t, t = 0, 1, \dots, T$ of HMM depends only on the previous state. And also, observations $\{Y_1, Y_2, \dots, Y_T\}$ are conditionally independent and identically distributed (i.i.d.) random variables. Observation $Y_t, t = 1, 2, \dots, T$ conditionally depend only on the current related hidden state X_t . The dependency structure between the hidden states and the observations are shown in Figure 2.1.

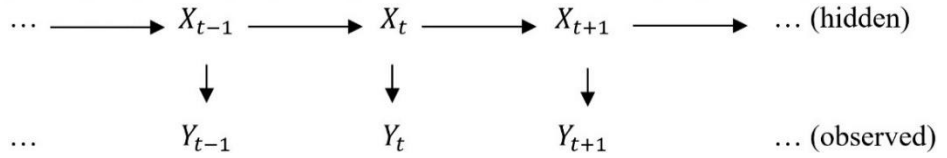


Figure 2.1 Graphical illustration of HMMs

The conditional probability distribution of a current state depends only on the most recent state because of the Markovian property; given below.

$$P(X_t | X_{t-1}, X_{t-2}, \dots, X_1, Y_{t-1}, Y_{t-2}, \dots, Y_1) = P(X_t | X_{t-1}), \quad (2.6)$$

where $t = 0, 1, \dots, T$.

In the classical HMM, the conditional probability distribution of observation symbols depends only on the current state and observations are conditionally independent of other variables given the current state. That is,

$$P(Y_t|X_t, X_{t-1}, \dots, X_1, Y_{t-1}, Y_{t-2}, \dots, Y_1) = P(Y_t|X_t), \quad (2.7)$$

where $t = 1, 2, \dots, T$.

Besides, the conditional probability of the observation sequence $(Y_t, Y_{t+1}, \dots, Y_T)$ depends only on the previous hidden state X_{t-1} ,

$$\begin{aligned} P(Y_t, Y_{t+1}, \dots, Y_T | X_1, X_2, \dots, X_{t-1}, Y_1, Y_2, \dots, Y_{t-1}) \\ = P(Y_t, Y_{t+1}, \dots, Y_T | X_{t-1}), \end{aligned} \quad (2.8)$$

where $t = 1, 2, \dots, T$.

In order to characterize HMM, following elements are needed to be described completely. The complete parameter set of HMM is $\lambda = (P, B, \pi)$. Let $\mathbf{P}$ be the one-step transition probability matrix, $\mathbf{B}$ be the observation emission distribution, and π be the initial state distribution. Then the complete parameter set is defined as follows:

The transition probability distribution. Transition probability from the current hidden state to the next hidden state can be obtained from the transition probability distribution. Besides, transitions occur according to an identical probability distribution. The transition probability matrix is denoted by $\mathbf{P} = \{p_{ij}\}$ and consists of $N \times N$ elements. Each element of the $\mathbf{P}$ are the probabilities of transitions from hidden state i to state j . That is,

$$p_{ij} = P\{X_{t+1} = j | X_t = i\}, \quad (2.9)$$

where $i, j \in \Omega$. Note that, p_{ij} provides the properties given in (2.4) and (2.5).

The observation emission distribution. The model emits an observation after each transition by the observation probability distribution. The probability of emitting the t -th observation $\{Y_t\}$ may vary according to hidden state. There are N hidden states and there are M possible observations in each state. So, the observation probability matrix is denoted by $\mathbf{B} = \{b_j(v_k)\}$ and consists of $N \times M$ elements. This matrix gives the probability of emitting an observation given that the system is at any hidden state. That is,

$$b_j(v_k) = P(Y_t = v_k \mid X_t = j), \quad (2.10)$$

where $j \in \Omega$, $v_k \in \mathcal{O}$, $k = 1, 2, \dots, M$, and $t = 0, 1, 2, \dots, T$.

The matrix representation of the observation emission distribution is as follows

$$\mathbf{B} = \begin{bmatrix} b_1(v_1) & b_1(v_2) & \cdots & b_1(v_M) \\ b_2(v_1) & b_2(v_2) & \cdots & b_2(v_M) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(v_1) & b_N(v_2) & \cdots & b_N(v_M) \end{bmatrix}.$$

All the elements of $\mathbf{B} = \{b_j(v_k)\}$, for all $j \in \Omega$ and $k = 1, 2, \dots, M$, are greater than or equal to zero. In addition, the sum of the probabilities of emitting all the possible observations in a state is equal to one. In other words, the sum of each row is equal to one. That is,

$$b_j(v_k) \geq 0 \text{ for all } j \in \Omega, v_k \in \mathcal{O}; \quad (2.11)$$

$$\sum_{k=1}^M b_j(v_k) = 1 \text{ for all } j \in \Omega. \quad (2.12)$$

The initial state distribution. The starting state of the system has also a distribution called “initial state distribution”. It is usually shown by an N -dimensional vector and gives the starting probabilities. Let π_i be the probability of starting at state i . Let X_0 be the initial hidden state of the process. Then

$$\pi_j = P(X_0 = j), \quad j \in \Omega. \quad (2.13)$$

Since π_j is a probability and the process is certain to start in one of the states, the sum of the initial probabilities are equal to one, that is, $\sum_{j=1}^N \pi_j = 1$.

2.3 Fundamental Problems of HMM

There are three fundamental problems for Hidden Markov models and some probabilistic algorithms are used to solve these problems.

2.3.1 Evaluation Problem

The aim of the evaluation is to compute the probability that an observation sequence is produced by the given HMM. In other words, the likelihood of the observation sequence $P(Y_1, Y_2, \dots, Y_T | \lambda)$ is computed. The evaluation problem is solved by forward or backward probabilities (Stratonovich, 1960).

Let $\alpha_j(t)$ be the forward probability which is defined as the probability of being in state j at time t having seen the observation sequence $\{Y_1, Y_2, \dots, Y_t\}$ where $t \leq T$. Forward probabilities calculated for each step are greater than or equal to zero ($\alpha_j(t) \geq 0, j \in \Omega$).

$$\alpha_j(t) = P(Y_1, Y_2, \dots, Y_t, X_t = j | \lambda) \quad (2.14)$$

$$\begin{aligned} &= \sum_{i=1}^N P(Y_1, Y_2, \dots, Y_{t-1}, Y_t, X_{t-1} = i, X_t = j) \\ &= \sum_{i=1}^N P(Y_1, Y_2, \dots, Y_{t-1}, X_{t-1} = i) \\ &\quad \cdot P(X_t = j, Y_t = v_k | Y_1, Y_2, \dots, Y_{t-1}, X_{t-1} = i) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N \alpha_i(t-1) P(X_t = j, Y_t = v_k | X_{t-1} = i) \\
&= \sum_{i=1}^N \alpha_i(t-1) \frac{P(X_{t-1} = i, X_t = j, Y_t = v_k)}{P(X_{t-1} = i)} \\
&= \sum_{i=1}^N \alpha_i(t-1) \frac{P(X_{t-1} = i) P(X_t = j | X_{t-1} = i) \cdot P(Y_t = v_k | X_t = j, X_{t-1} = i)}{P(X_{t-1} = i)} \\
&= \sum_{i=1}^N \alpha_i(t-1) p_{ij} P(Y_t = v_k | X_t = j) \\
&= \sum_{i=1}^N \alpha_i(t-1) p_{ij} b_j(v_k) \\
\Rightarrow \alpha_j(t) &= \begin{cases} \pi_j b_j(v_k) & , t = 0 \\ \sum_{i=1}^N \alpha_i(t-1) p_{ij} b_j(v_k) & , t = 1, 2, \dots, T \end{cases} \quad , \quad (2.15)
\end{aligned}$$

where $i, j \in \Omega$, $v_k \in \mathcal{O}$, $k = 1, 2, \dots, M$.

Suppose that HMM has N hidden states with a finite state space $\Omega = \{1, 2, 3, \dots\}$. In this case, the calculation of the forward probabilities for time t is given in Figure 2.2.

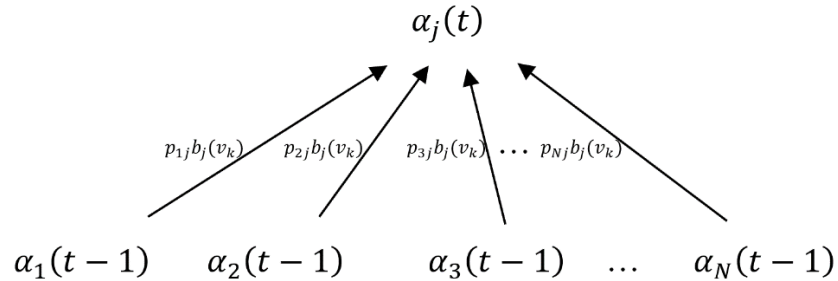


Figure 2.2 Flow chart of the forward probabilities for time t

Figure 2.2 above only gives the forward probability calculation for the hidden state j . Considering that there are N hidden states, number of N forward probabilities are calculated at each step of the observation sequence.

Based on the definition of the forward probability, it is possible to calculate the likelihood value of the model in terms of $\alpha_j(t)$. Let $L(\lambda)$ be the likelihood of HMM.

$$L(\lambda) = P(Y_1, Y_2, \dots, Y_T) \quad (2.16)$$

$$\begin{aligned} &= \sum_{j=1}^N P(Y_1, Y_2, \dots, Y_T, X_T = j) \\ &= \sum_{j=1}^N \alpha_j(T) \end{aligned}$$

where T is the length of the observation sequence and $j \in \Omega$.

Let $\beta_j(t)$ be the backward probability which is defined as the probability of observing $\{Y_{t+1}, Y_{t+2}, \dots, Y_T\}$, given that the process is in state j at time t , where $t \leq T$. Backward probabilities calculated for each step are greater than or equal to zero; that is $\beta_j(t) \geq 0$ for all $j \in \Omega$.

$$\beta_j(t) = P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, \lambda) \quad (2.17)$$

$$\begin{aligned} &= \sum_{i=1}^N P(Y_{t+1}, Y_{t+2}, \dots, Y_T, X_{t+1} = i | X_t = j) \\ &= \sum_{i=1}^N P(Y_{t+2}, \dots, Y_T | X_t = j, X_{t+1} = i, Y_{t+1}) \\ &\quad \cdot P(Y_{t+1} = v_k | X_t = j, X_{t+1} = i) \\ &\quad \cdot P(X_{t+1} = i | X_t = j) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N P(Y_{t+2}, \dots, Y_T | X_{t+1} = i) \\
&\quad \cdot P(Y_{t+1} = v_k | X_{t+1} = i) P(X_{t+1} = i | X_t = j) \\
&= \sum_{i=1}^N \beta_i(t+1) p_{ji} b_i(v_k) \\
\Rightarrow \beta_j(t) &= \begin{cases} \sum_{i=1}^N \beta_i(t+1) p_{ji} b_i(v_k) & , t = 1, 2, \dots, T-1, \\ 1 & , t = T \end{cases} \quad (2.18)
\end{aligned}$$

where $i, j \in \Omega$, $v_k \in \mathcal{O}$, $k = 1, 2, \dots, M$.

The computational space complexity of HMM is $\mathcal{O}(NT)$, because of the number of N ; and forward or backward probabilities are calculated for a length of $t = 1, 2, \dots, T$. In addition, each forward or backward probability consists of the sum of the number of components as many as the number of hidden states. So, the time complexity of the algorithm for the evaluation problem is $\mathcal{O}(N^2T)$.

Suppose that HMM has N hidden states with a state space $\Omega = \{1, 2, 3, \dots, N\}$. In this case, the calculation of the backward probabilities for time t is given in the Figure 2.3.

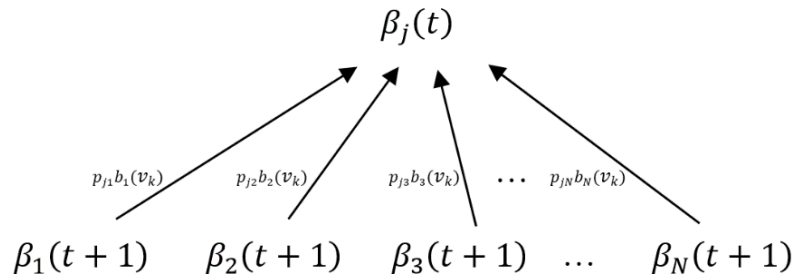


Figure 2.3 Flow chart of the backward probabilities at time t

The Figure 2.3 above only gives the backward probability calculation for the hidden state j . Considering that there are N hidden states, number of N backward probabilities are calculated at each step of the observation sequence.

It is possible to obtain the posterior marginal probabilities of hidden states $\{X_1, X_2, \dots, X_T\}$ at any time, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$. Forward-backward algorithm (Rabiner, 1989) is used to calculate the related posteriors.

$$\begin{aligned}
 P(X_t = j | Y_1, Y_2, \dots, Y_T) &\underset{X_t}{\propto} P(X_t = j, Y_1, Y_2, \dots, Y_T) & (2.19) \\
 &= P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_1, Y_2, \dots, Y_t) \\
 &\quad \cdot P(X_t = j, Y_1, Y_2, \dots, Y_t) \\
 &= P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j) \\
 &\quad \cdot P(X_t = i, Y_1, Y_2, \dots, Y_t) \\
 &= \beta_j(t) \alpha_j(t)
 \end{aligned}$$

2.3.2 Learning Problem

The aim of the learning procedure is to find the most likely set of state transition and observation emission probabilities, given a set of output sequences and general structure of HMM (number of hidden states). The learning problem and estimation of parameters are solved by the Baum-Welch algorithm (Baum, Petrie, Soules, & Weiss, 1970). It is an iterative expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977) which determines HMM parameters $\lambda = (P, B, \pi)$ that best fit to observation training data, which maximizes $P(Y_1, Y_2, \dots, Y_T | \lambda)$.

The first thing to do is to determine arbitrarily the initial values of the complete parameter set because initial values are essential to start the algorithm. In order for the Baum-Welch algorithm to run, some probabilistic ideas must first be defined. These general ideas help to obtain the equations required for the estimation of the parameters using forward and backward probabilities.

The general idea for calculating the new estimates of the state transition probabilities p_{ij} is given below,

$$p_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions out of state } i}.$$

The general idea for calculating the new estimates of the observation emission probabilities $b_j(v_k)$ is given below,

$$b_j(v_k) = \frac{\text{expected number of times observation } v_k \text{ occurs in state } j}{\text{expected number of times in state } j}.$$

Finally, the general idea for calculating the new estimates of the initial state probabilities π_j is given below,

$$\pi_j = \text{expected frequency in state } j \text{ at time } t = 1.$$

Baum-Welch algorithm consists of two steps. In the first step, called the expectation-step, two probabilities must be defined. Let the random variable $\xi_{ij}(t)$ be the probability of being in state i at time t and in state j at time $t + 1$, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$.

$$\xi_{ij}(t) = P(X_t = i, X_{t+1} = j | Y_1 Y_2 \dots Y_T) \quad (2.20)$$

$$\begin{aligned} &= \frac{P(X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \\ &= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1}) \cdot P(X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1})}{L(\lambda)} \\ &= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j) \cdot P(Y_{t+1} = v_k | X_{t+1} = j, X_t = i, Y_1, Y_2, \dots, Y_t) \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i, X_{t+1} = j)}{L(\lambda)} \end{aligned}$$

$$\begin{aligned}
& \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j) P(Y_{t+1} = v_k | X_{t+1} = j)}{.P(X_{t+1} = j | Y_1, Y_2, \dots, Y_t, X_t = i) P(Y_1, Y_2, \dots, Y_t, X_t = i)} \\
&= \frac{L(\lambda)}{L(\lambda)} \\
&= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j) P(Y_{t+1} = v_k | X_{t+1} = j)}{.P(X_{t+1} = j | X_t = i) P(Y_1, Y_2, \dots, Y_t, X_t = i)} \\
&= \frac{L(\lambda)}{L(\lambda)} \\
&= \frac{\alpha_i(t) p_{ij} b_j(v_k) \beta_j(t+1)}{L(\lambda)}
\end{aligned}$$

Note that the probabilities $\xi_{ij}(t)$ for each step are greater than or equal to zero and the sum of the probabilities for all states at time t is equal to one. That is,

$$\xi_{ij}(t) \geq 0, \quad i, j \in \Omega; \quad (2.21)$$

$$\sum_{i=1}^N \sum_{j=1}^N \xi_{ij}(t) = 1, \quad (2.22)$$

where $t = 1, 2, \dots, T-1$.

Let the random variable $\gamma_j(t)$ defined as the probability of being in state j at time t , given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$.

$$\gamma_j(t) = P(X_t = j | Y_1, Y_2, \dots, Y_T) \quad (2.23)$$

$$\begin{aligned}
&= \frac{P(X_t = j, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \\
&= \frac{P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_1, Y_2, \dots, Y_t)}{.P(X_t = j, Y_1, Y_2, \dots, Y_t)} \\
&= \frac{L(\lambda)}{L(\lambda)} \\
&= \frac{P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j)}{.P(X_t = j, Y_1, Y_2, \dots, Y_t)} \\
&= \frac{L(\lambda)}{L(\lambda)}
\end{aligned}$$

$$= \frac{\alpha_j(t)\beta_j(t)}{L(\lambda)}$$

Note that sum of the probabilities $\gamma_j(t)$ for all states at time t is equal to one, and each are greater than or equal to zero. That is,

$$\gamma_j(t) \geq 0 \text{ for all } j \in \Omega ; \quad (2.24)$$

$$\sum_{j=1}^N \gamma_j(t) = 1, \quad (2.25)$$

where $j \in \Omega$, $t = 1, 2, \dots, T$.

It is possible to write the expected values defined in the general probability ideas in terms of probabilities $\xi_{ij}(t)$ and $\gamma_j(t)$. By this way, general ideas can be formulated and expectation-step can be completed.

Let $E(N_{i \rightarrow j})$ be the expected number of transitions from state i to state j . It can be calculated as follows,

$$E(N_{i \rightarrow j}) = \sum_{t=1}^{T-1} \xi_{ij}(t) . \quad (2.26)$$

Let $E(N_{i \rightarrow})$ be the expected number of transitions out of state i . It can be calculated as follows,

$$E(N_{i \rightarrow}) = \sum_{t=1}^{T-1} \gamma_i(t) . \quad (2.27)$$

There is an important relationship between the expected values $E(N_{i \rightarrow j})$ and $E(N_{i \rightarrow})$. The sum of the values $E(N_{i \rightarrow j})$ from one state to all other states are equal to the value of $E(N_{i \rightarrow})$. That is,

$$\sum_{j=1}^N E(N_{i \rightarrow j}) = E(N_{i \rightarrow}). \quad (2.28)$$

Let $E(N_j^{v_k})$ be the expected number of times that observation v_k occurs in state j . It can be calculated as follows,

$$E(N_j^{v_k}) = \sum_{\substack{t=1 \\ Y_t=v_k}}^T \gamma_j(t). \quad (2.29)$$

Let $E(N_j)$ be the expected number of times that the system visits state j . It can be calculated as follows,

$$E(N_j) = \sum_{t=1}^T \gamma_j(t). \quad (2.30)$$

And finally, let $E(N_j(1))$ be the expected frequency in state j at time ($t = 1$). It can be calculated as follows,

$$E(N_j(1)) = \gamma_j(1). \quad (2.31)$$

In the second step, called the maximization-step, new estimated values of the parameters are obtained by using the expected values. The estimated value of the transition probability distribution for the n th iteration is

$$p_{ij}^{(n)} = \frac{E(N_{i \rightarrow j})}{E(N_{i \rightarrow})} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}. \quad (2.32)$$

The estimated value of the observation emission distribution for the n th iteration is

$$b_j^{(n)}(v_k) = \frac{E(N_j^{v_k})}{E(N_j)} = \frac{\sum_{t=1}^T \gamma_j(t)}{\sum_{t=1}^T \gamma_j(t)} \quad (2.33)$$

And finally, the estimated value of the initial state distribution for the n th iteration is

$$\pi_j^{(n)} = E(N_j(1)) = \gamma_j(1) \quad (2.34)$$

Expectation-maximization iterations are continued until a predetermined stopping criteria is achieved. Usually, two types of stopping criteria are used for the Baum-Welch algorithm. One is by using the parameter estimates whereas the other one is by using the log-likelihood value of the model.

Let ϵ be a predetermined value ($10^{-3}, 10^{-4}$ etc.) which is small enough. If the absolute difference between successive parameter estimates is smaller than or equal to ϵ , algorithm stops and final estimates are recorded. It should be noted that the complete parameter set must hold this condition. That is,

$$|\hat{\lambda}^{(n)} - \hat{\lambda}^{(n-1)}| \leq \epsilon \quad (2.35)$$

The next method is taking the absolute difference of the log-likelihood values computed by the parameter estimates between two successive iterations. The stopping criteria in terms of log-likelihood values is given as follows,

$$|\ln L(\hat{\lambda})^{(n)} - \ln L(\hat{\lambda})^{(n-1)}| \leq \epsilon \quad (2.36)$$

2.3.3 Decoding Problem

The aim of the decoding procedure is to estimate the most probable state transition sequence, given the observation sequence. In other words, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$, and the model parameters $\lambda = (P, B, \pi)$, how do we choose a corresponding state sequence path which is optimal in some meaningful sense (i.e., best “explains” the observations)? The most common and effective way to solve this problem is the Viterbi algorithm (Viterbi, 1967).

Let S be hidden state sequence and S^* be the “best” possible hidden state sequence. Then the goal is to maximize $P(S^*|Y_1, Y_2, \dots, Y_T)$ by choosing the “best” S . That is,

$$S^* = \underset{S}{\operatorname{argmax}} [P(S|Y_1, Y_2, \dots, Y_T)]. \quad (2.37)$$

The following equation can be achieved by using the conditional probability and the chain rule.

$$\begin{aligned} P(S|Y_1, Y_2, \dots, Y_T) &= \frac{P(S, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \\ &= \frac{P(S)P(Y_1, Y_2, \dots, Y_T|S)}{L(\lambda)} \end{aligned} \quad (2.38)$$

$P(Y_1, Y_2, \dots, Y_T)$ is the likelihood $L(\lambda)$ of the model and remain constant over time, so $P(S|Y_1, Y_2, \dots, Y_T)$ is proportional to the joint probability distribution $P(S, Y_1, Y_2, \dots, Y_T)$. That is,

$$P(S|Y_1, Y_2, \dots, Y_T) \underset{S}{\propto} P(S) \cdot P(Y_1, Y_2, \dots, Y_T|S). \quad (2.39)$$

So, the “best” possible hidden state sequence (S^*) can be estimated as follows,

$$S^* = \underset{S}{\operatorname{argmax}} [P(S)P(Y_1, Y_2, \dots, Y_T|S)]. \quad (2.40)$$

The joint probability of the hidden state sequence $P(S)$ can be calculated as follows, note that X_0 is the initial state of the Markov chain.

$$\begin{aligned}
P(S) &= P(X_0, X_1, \dots, X_T) \\
&= P(X_0)P(X_1|X_0)P(X_2|X_0, X_1) \\
&\quad \dots P(X_T|X_0, X_1, \dots, X_{T-1}) \\
&= P(X_0)P(X_1|X_0)P(X_2|X_1) \dots P(X_T|X_{T-1})
\end{aligned} \tag{2.41}$$

The conditional probability of the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$ given the hidden state sequence $\{X_0, X_1, \dots, X_T\}$ can be calculated as given in (2.42).

$$\begin{aligned}
P(Y_1, Y_2, \dots, Y_T|S) &= \frac{P(Y_1, Y_2, \dots, Y_T, S)}{P(S)} \\
&= \frac{P(S)P(Y_1|S)P(Y_2|S, Y_1) \dots \\
&\quad \cdot P(Y_T|S, Y_1, Y_2, \dots, Y_{T-1})}{P(S)} \\
&= P(Y_1|X_1)P(Y_2|X_2) \dots P(Y_T|X_T)
\end{aligned} \tag{2.42}$$

Therefore, the term in equation (2.40) can be expressed as follows:

$$\begin{aligned}
&P(S)P(Y_1, Y_2, \dots, Y_T|S) \\
&= P(X_0)P(X_1|X_0)P(X_2|X_1) \dots P(X_T|X_{T-1}) \\
&\quad \cdot P(Y_1|X_1)P(Y_2|X_2) \dots P(Y_T|X_T) \\
&= P(X_0) \prod_{t=1}^T P(X_t|X_{t-1}) \prod_{t=1}^T P(Y_t|X_t) \\
&= P(X_0 = r) \prod_{t=1}^T P(X_t = j|X_{t-1} = i)P(Y_t = v_k|X_t = j)
\end{aligned} \tag{2.43}$$

$$= \pi_r \prod_{t=1}^T p_{ij} b_j(v_k),$$

where $i, j, r \in \Omega$, $v_k \in \mathcal{O}$ and $k = 1, 2, \dots, M$.

The most likely hidden state is estimated for each step of the observation sequence by using the above formula. The “best” possible hidden state sequence is the most likely state sequence that should emit the observation sequence.

2.4 Illustrative Example

Assume that there is a HMM with two states, where $\Omega = \{1, 2\}$. In addition there are two possible observations $\mathcal{O} = \{v_k, v_l\}$ which can be emitted after each state transitions. Let us assume that the process emits an observation sequence with the length of $T = 3$ and the observation sequence is $\{v_k, v_k, v_l\}$. Let us further assume that the complete parameter set $\lambda = (P, B, \pi)$ is defined as follows,

$$\mathbf{P} = \begin{bmatrix} 0.25 & 0.75 \\ 0.60 & 0.40 \end{bmatrix}, \quad (2.44)$$

$$\mathbf{B} = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix}, \quad (2.45)$$

$$\boldsymbol{\pi} = [0.50 \quad 0.50]. \quad (2.46)$$

The probabilistic definitions of each element of the parameter matrices are given below,

$$P(X_t = 1 | X_{t-1} = 1) = 0.25$$

$$P(X_t = 2 | X_{t-1} = 1) = 0.75$$

$$P(X_t = 1 | X_{t-1} = 2) = 0.60$$

$$P(X_t = 2|X_{t-1} = 2) = 0.40$$

$$P(Y_t = v_k|X_t = 1) = 0.80$$

$$P(Y_t = v_l|X_t = 1) = 0.20$$

$$P(Y_t = v_k|X_t = 2) = 0.30$$

$$P(Y_t = v_l|X_t = 2) = 0.70$$

$$P(X_1 = 1) = 0.50$$

$$P(X_1 = 2) = 0.50$$

First, assume that $\lambda = (P, B, \pi)$ given above is the initial values chosen arbitrarily. The goal is to find the parameter estimates that fit best to the given observation sequence.

To start the Baum-Welch algorithm, forward and backward probabilities must be calculated first. That is,

$$\hat{\lambda}^{(0)};$$

$$\hat{\mathbf{P}}^{(0)} = \begin{bmatrix} 0.25 & 0.75 \\ 0.60 & 0.40 \end{bmatrix},$$

$$\hat{\mathbf{B}}^{(0)} = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix},$$

$$\hat{\boldsymbol{\pi}}^{(0)} = [0.50 \quad 0.50].$$

$$\alpha_1(1) = \pi_1 b_1(v_k) = (0.50)(0.80) = 0.40$$

$$\alpha_2(1) = \pi_2 b_2(v_k) = (0.50)(0.30) = 0.15$$

$$\alpha_1(2) = \sum_{i=1}^2 \alpha_i(1) p_{i1} b_1(v_k)$$

$$\begin{aligned}
&= \alpha_1(1)p_{11}b_1(v_k) + \alpha_2(1)p_{21}b_1(v_k) \\
&= (0.40)(0.25)(0.80) + (0.15)(0.60)(0.80) = 0.152
\end{aligned}$$

$$\begin{aligned}
\alpha_2(2) &= \sum_{i=1}^2 \alpha_i(1) p_{i2} b_2(v_k) \\
&= \alpha_1(1)p_{12}b_2(v_k) + \alpha_2(1)p_{22}b_2(v_k) \\
&= (0.40)(0.75)(0.30) + (0.15)(0.40)(0.30) = 0.108
\end{aligned}$$

$$\begin{aligned}
\alpha_1(3) &= \sum_{i=1}^2 \alpha_i(2) p_{i1} b_1(v_l) \\
&= \alpha_1(2)p_{11}b_1(v_l) + \alpha_2(2)p_{21}b_1(v_l) \\
&= (0.152)(0.25)(0.20) + (0.108)(0.60)(0.20) = 0.02056
\end{aligned}$$

$$\begin{aligned}
\alpha_2(3) &= \sum_{i=1}^2 \alpha_i(2) p_{i2} b_2(v_l) \\
&= \alpha_1(2)p_{12}b_2(v_l) + \alpha_2(2)p_{22}b_2(v_l) \\
&= (0.152)(0.75)(0.70) + (0.108)(0.40)(0.70) = 0.11004
\end{aligned}$$

$$\beta_1(3) = 1$$

$$\beta_2(3) = 1$$

$$\begin{aligned}
\beta_1(2) &= \sum_{i=1}^N \beta_i(3) p_{1i} b_i(v_l) \\
&= \beta_1(3)p_{11}b_1(v_l) + \beta_2(3)p_{12}b_2(v_l) \\
&= (1)(0.25)(0.20) + (1)(0.75)(0.7) = 0.575
\end{aligned}$$

$$\begin{aligned}
\beta_2(2) &= \sum_{i=1}^N \beta_i(3) p_{2i} b_i(v_l) \\
&= \beta_1(3)p_{21}b_1(v_l) + \beta_2(3)p_{22}b_2(v_l)
\end{aligned}$$

$$= (1)(0.60)(0.20) + (1)(0.40)(0.70) = 0.40$$

$$\begin{aligned}\beta_1(1) &= \sum_{i=1}^N \beta_i(2) p_{1i} b_i(v_k) \\ &= \beta_1(2) p_{11} b_1(v_k) + \beta_2(2) p_{12} b_2(v_k) \\ &= (0.575)(0.25)(0.80) + (0.40)(0.75)(0.30) = 0.205\end{aligned}$$

$$\begin{aligned}\beta_2(1) &= \sum_{i=1}^N \beta_i(2) p_{2i} b_i(v_k) \\ &= \beta_1(2) p_{21} b_1(v_k) + \beta_2(2) p_{22} b_2(v_k) \\ &= (0.575)(0.60)(0.80) + (0.40)(0.40)(0.30) = 0.324\end{aligned}$$

Note that the likelihood and log-likelihood values of the model are,

$$L(\hat{\lambda}^{(0)}) = \sum_{j=1}^2 \alpha_j(3) = \alpha_1(3) + \alpha_2(3) = 0.02056 + 0.11004 = 0.1306,$$

$$\ln L(\hat{\lambda}^{(0)}) \cong -2.0356$$

Since forward and backward probabilities are calculated, $\xi_{ij}(t)$ and $\gamma_j(t)$ can now be calculated,

$$\xi_{11}(1) = \frac{\alpha_1(1) p_{11} b_1(v_k) \beta_1(2)}{L(\hat{\lambda})} = \frac{(0.40)(0.25)(0.80)(0.575)}{(0.1306)} \cong 0.3522$$

$$\xi_{12}(1) = \frac{\alpha_1(1) p_{12} b_2(v_k) \beta_2(2)}{L(\hat{\lambda})} = \frac{(0.40)(0.75)(0.30)(0.40)}{(0.1306)} \cong 0.2757$$

$$\xi_{21}(1) = \frac{\alpha_2(1) p_{21} b_1(v_k) \beta_1(2)}{L(\hat{\lambda})} = \frac{(0.15)(0.60)(0.80)(0.575)}{(0.1306)} \cong 0.3170$$

$$\xi_{22}(1) = \frac{\alpha_2(1) p_{22} b_2(v_k) \beta_2(2)}{L(\hat{\lambda})} = \frac{(0.15)(0.40)(0.30)(0.40)}{(0.1306)} \cong 0.0551$$

$$\xi_{11}(2) = \frac{\alpha_1(2)p_{11}b_1(v_l)\beta_1(3)}{L(\hat{\lambda})} = \frac{(0.152)(0.25)(0.20)(1)}{(0.1306)} \cong 0.0582$$

$$\xi_{12}(2) = \frac{\alpha_1(2)p_{12}b_2(v_l)\beta_2(3)}{L(\hat{\lambda})} = \frac{(0.152)(0.75)(0.70)(1)}{(0.1306)} \cong 0.6110$$

$$\xi_{21}(2) = \frac{\alpha_2(2)p_{21}b_1(v_l)\beta_1(3)}{L(\hat{\lambda})} = \frac{(0.108)(0.60)(0.20)(1)}{(0.1306)} \cong 0.0992$$

$$\xi_{22}(2) = \frac{\alpha_2(2)p_{22}b_2(v_l)\beta_2(3)}{L(\hat{\lambda})} = \frac{(0.108)(0.40)(0.70)(1)}{(0.1306)} \cong 0.2316$$

Note that, sum of the probabilities $\xi_{ij}(t)$ at time t is,

$$\sum_{i=1}^2 \sum_{j=1}^2 \xi_{ij}(1) = 0.3522 + 0.2757 + 0.3170 + 0.0551 = 1,$$

$$\sum_{i=1}^2 \sum_{j=1}^2 \xi_{ij}(2) = 0.0582 + 0.6110 + 0.0992 + 0.2316 = 1.$$

$$\gamma_1(1) = \frac{\alpha_1(1)\beta_1(1)}{L(\hat{\lambda})} = \frac{(0.40)(0.205)}{(0.1306)} \cong 0.6279$$

$$\gamma_2(1) = \frac{\alpha_2(1)\beta_2(1)}{L(\hat{\lambda})} = \frac{(0.15)(0.324)}{(0.1306)} \cong 0.3721$$

$$\gamma_1(2) = \frac{\alpha_1(2)\beta_1(2)}{L(\hat{\lambda})} = \frac{(0.152)(0.575)}{(0.1306)} \cong 0.6692$$

$$\gamma_2(2) = \frac{\alpha_2(2)\beta_2(2)}{L(\hat{\lambda})} = \frac{(0.108)(0.40)}{(0.1306)} \cong 0.3308$$

$$\gamma_1(3) = \frac{\alpha_1(3)\beta_1(3)}{L(\hat{\lambda})} = \frac{(0.02056)(1)}{(0.1306)} \cong 0.1574$$

$$\gamma_2(3) = \frac{\alpha_2(3)\beta_2(3)}{L(\hat{\lambda})} = \frac{(0.11004)(1)}{(0.1306)} \cong 0.8426$$

Note that, sum of the probabilities $\gamma_j(t)$ at time t for all $t \in T$ is,

$$\sum_{j=1}^2 \gamma_j(1) = 0.6279 + 0.3721 = 1,$$

$$\sum_{j=1}^2 \gamma_j(2) = 0.6692 + 0.3308 = 1,$$

$$\sum_{j=1}^2 \gamma_j(3) = 0.1574 + 0.8426 = 1.$$

The expected number of transitions from state i to state j for each $i, j \in \Omega$ is,

$$E(N_{1 \rightarrow 1}) = \sum_{t=1}^2 \xi_{11}(t) = \xi_{11}(1) + \xi_{11}(2) = 0.3522 + 0.0582 = 0.4104$$

$$E(N_{1 \rightarrow 2}) = \sum_{t=1}^2 \xi_{12}(t) = \xi_{12}(1) + \xi_{12}(2) = 0.2757 + 0.6110 = 0.8867$$

$$E(N_{2 \rightarrow 1}) = \sum_{t=1}^2 \xi_{21}(t) = \xi_{21}(1) + \xi_{21}(2) = 0.3170 + 0.0992 = 0.4162$$

$$E(N_{2 \rightarrow 2}) = \sum_{t=1}^2 \xi_{22}(t) = \xi_{22}(1) + \xi_{22}(2) = 0.0551 + 0.2316 = 0.2867$$

The expected number of transitions out of state i for each $i \in \Omega$ is,

$$E(N_{1 \rightarrow}) = \sum_{t=1}^2 \gamma_1(t) = \gamma_1(1) + \gamma_1(2) = 0.6279 + 0.6692 = 1.2971$$

$$E(N_{2 \rightarrow}) = \sum_{t=1}^2 \gamma_2(t) = \gamma_2(1) + \gamma_2(2) = 0.3721 + 0.3308 = 0.7029$$

The expected number of times that observation v_k or v_l occurs in state j is given below. Note that $Y_1 = v_k$, $Y_2 = v_k$ and $Y_3 = v_l$.

$$E(N_1^{v_k}) = \sum_{\substack{t=1 \\ Y_t=v_k}}^3 \gamma_1(t) = \gamma_1(1) + \gamma_1(2) = 0.6279 + 0.6692 = 1.2971$$

$$E(N_1^{v_l}) = \sum_{\substack{t=1 \\ Y_t=v_l}}^3 \gamma_1(t) = \gamma_1(3) = 0.1574$$

$$E(N_2^{v_k}) = \sum_{\substack{t=1 \\ Y_t=v_k}}^3 \gamma_2(t) = \gamma_2(1) + \gamma_2(2) = 0.3721 + 0.3308 = 0.7029$$

$$E(N_2^{v_l}) = \sum_{\substack{t=1 \\ Y_t=v_l}}^3 \gamma_2(t) = \gamma_2(3) = 0.8426$$

The expected number of times that the system visits state j , for each $j \in \Omega$ is:

$$\begin{aligned} E(N_1) &= \sum_{t=1}^3 \gamma_1(t) = \gamma_1(1) + \gamma_1(2) + \gamma_1(3) \\ &= 0.6279 + 0.6692 + 0.1574 = 1.4545 \end{aligned}$$

$$\begin{aligned} E(N_2) &= \sum_{t=1}^3 \gamma_2(t) = \gamma_2(1) + \gamma_2(2) + \gamma_2(3) \\ &= 0.3721 + 0.3308 + 0.8426 = 1.5455 \end{aligned}$$

The expected frequency in state j , for each $j \in \Omega$ at time $t = 1$ is:

$$E(N_1(1)) = \gamma_1(1) = 0.6279$$

$$E(N_2(1)) = \gamma_2(1) = 0.3721$$

It is time to update the estimated values of the parameters for the maximization step, as the expected values are calculated. That is,

$$p_{11}^{(1)} = \frac{E(N_{1 \rightarrow 1})}{E(N_{1 \rightarrow})} = \frac{0.4104}{1.2971} \cong 0.3164$$

$$p_{12}^{(1)} = \frac{E(N_{1 \rightarrow 2})}{E(N_{1 \rightarrow})} = 1 - p_{11}^{(1)} \cong 0.6836$$

$$p_{21}^{(1)} = \frac{E(N_{2 \rightarrow 1})}{E(N_{2 \rightarrow})} = \frac{0.4162}{0.7029} \cong 0.5921$$

$$p_{22}^{(1)} = \frac{E(N_{2 \rightarrow 2})}{E(N_{2 \rightarrow})} = 1 - p_{21}^{(1)} \cong 0.4079$$

$$b_1^{(1)}(v_k) = \frac{E(N_1^{v_k})}{E(N_1)} = \frac{1.2971}{1.4545} \cong 0.8918$$

$$b_1^{(1)}(v_l) = \frac{E(N_1^{v_l})}{E(N_1)} = 1 - b_1^{(1)}(v_k) \cong 0.1082$$

$$b_2^{(1)}(v_k) = \frac{E(N_2^{v_k})}{E(N_2)} = \frac{0.7029}{1.5455} \cong 0.4548$$

$$b_2^{(1)}(v_l) = \frac{E(N_2^{v_l})}{E(N_2)} = 1 - b_2^{(1)}(v_k) \cong 0.5452$$

$$\pi_1^{(1)} = E(N_{1_1}) = \gamma_1(1) = 0.6279$$

$$\pi_2^{(1)} = E(N_{2_1}) = \gamma_2(1) = 0.3721$$

So, as a result of the first step of Baum-Welch algorithm, the new estimates are obtained as follows,

$$\hat{\lambda}^{(1)};$$

$$\hat{\mathbf{P}}^{(1)} = \begin{bmatrix} 0.3164 & 0.6836 \\ 0.5921 & 0.4079 \end{bmatrix},$$

$$\hat{\mathbf{B}}^{(1)} = \begin{bmatrix} 0.8918 & 0.1082 \\ 0.4548 & 0.5452 \end{bmatrix},$$

$$\hat{\boldsymbol{\pi}}^{(1)} = [0.6279 \quad 0.3721].$$

Assume that the stopping criteria for the algorithm is defined as $\epsilon = 10^{-4}$. Then these iterations are continued until the difference between the previous and current estimates are less than or equal to $\epsilon = 10^{-4}$. The second method is to compute the differences of log-likelihood values. The second iteration is started by assigning $\hat{\lambda}^{(1)}$ as the new initial estimate values, and $\ln L(\hat{\lambda}^{(1)})$ is obtained by using the new forward probabilities. Finally, if the difference $|\ln L(\hat{\lambda}^{(1)}) - \ln L(\hat{\lambda}^{(0)})| \leq 10^{-4}$, the algorithm stops, otherwise continues.

Now, assume that $\lambda = (P, B, \pi)$ given in (2.44), (2.45), and (2.46) is the final values of the Baum-Welch algorithm. The goal is to find the most likely hidden state sequence given the observation sequence $\{Y_1 = v_k, Y_2 = v_k, Y_3 = v_l\}$; That is, $\hat{\lambda}^{(final)}$,

$$\hat{\mathbf{P}}^{(final)} = \begin{bmatrix} 0.25 & 0.75 \\ 0.60 & 0.40 \end{bmatrix},$$

$$\hat{\mathbf{B}}^{(final)} = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix},$$

$$\hat{\boldsymbol{\pi}}^{(final)} = [0.5 \quad 0.5].$$

The logic of the Viterbi algorithm is based on finding the conditional probability of all hidden state sequence combinations given observation sequence and choose the most likely (best) among them. Recall that the length of the observation sequence is $T = 3$, and each observation is emitted right after a state transition. Therefore, the length of the hidden state sequence is also $T = 3$ and a sum of $2^3 = 8$ different hidden state state sequence combinations can occur. The probabilities of the estimates of hidden state sequences are obtained by equation (2.45).

$$\begin{aligned}
& P(X_1 = 1, X_2 = 1, X_3 = 1 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i1} b_1(Y_1 = v_k) p_{11} b_1(Y_2 = v_k) p_{11} b_1(Y_3 = v_l) \\
&= (0.50)(0.25)(0.80)(0.25)(0.80)(0.25)(0.20) \\
&\quad + (0.50)(0.60)(0.80)(0.25)(0.80)(0.25)(0.20) = 0.0034
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 1, X_2 = 1, X_3 = 2 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i1} b_1(Y_1 = v_k) p_{11} b_1(Y_2 = v_k) p_{12} b_2(Y_3 = v_l) \\
&= (0.50)(0.25)(0.80)(0.25)(0.80)(0.75)(0.70) \\
&\quad + (0.50)(0.60)(0.80)(0.25)(0.80)(0.75)(0.70) = 0.0357
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 1, X_2 = 2, X_3 = 1 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i1} b_1(Y_1 = v_k) p_{12} b_2(Y_2 = v_k) p_{21} b_1(Y_3 = v_l) \\
&= (0.50)(0.25)(0.80)(0.75)(0.30)(0.60)(0.20) \\
&\quad + (0.50)(0.60)(0.80)(0.75)(0.30)(0.60)(0.20) = 0.0092
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 2, X_2 = 1, X_3 = 1 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i2} b_2(Y_1 = v_k) p_{21} b_1(Y_2 = v_k) p_{11} b_1(Y_3 = v_l) \\
&= (0.50)(0.75)(0.30)(0.60)(0.80)(0.25)(0.20) \\
&\quad + (0.50)(0.40)(0.30)(0.60)(0.80)(0.25)(0.20) = 0.0041
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 1, X_2 = 2, X_3 = 2 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i1} b_1(Y_1 = v_k) p_{12} b_2(Y_2 = v_k) p_{22} b_2(Y_3 = v_l) \\
&= (0.50)(0.25)(0.80)(0.75)(0.30)(0.40)(0.70) \\
&\quad + (0.50)(0.60)(0.80)(0.75)(0.30)(0.40)(0.70) = 0.0214
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 2, X_2 = 1, X_3 = 2 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i2} b_2(Y_1 = v_k) p_{21} b_1(Y_2 = v_k) p_{12} b_2(Y_3 = v_l) \\
&= (0.50)(0.75)(0.30)(0.60)(0.80)(0.75)(0.70) \\
&\quad + (0.50)(0.40)(0.30)(0.60)(0.80)(0.75)(0.70) = 0.0435
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 2, X_2 = 2, X_3 = 1 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i2} b_2(Y_1 = v_k) p_{22} b_2(Y_2 = v_k) p_{21} b_1(Y_3 = v_l) \\
&= (0.50)(0.75)(0.30)(0.40)(0.30)(0.60)(0.20) \\
&\quad + (0.50)(0.40)(0.30)(0.40)(0.30)(0.60)(0.20) = 0.0025
\end{aligned}$$

$$\begin{aligned}
& P(X_1 = 2, X_2 = 2, X_3 = 2 | Y_1 = v_k, Y_2 = v_k, Y_3 = v_l) \\
&= \sum_{i=1}^2 \pi_i p_{i2} b_2(Y_1 = v_k) p_{22} b_2(Y_2 = v_k) p_{22} b_2(Y_3 = v_l) \\
&= (0.50)(0.75)(0.30)(0.40)(0.30)(0.40)(0.70) \\
&\quad + (0.50)(0.75)(0.30)(0.40)(0.30)(0.40)(0.70) = 0.0076
\end{aligned}$$

The most likely hidden state sequence given the observation sequence is, $S^* = \{X_1 = 2, X_2 = 1, X_3 = 2\}$. This is the expected result as the final parameters actually give some information about the state pattern. According to the $\hat{\mathbf{B}}^{(final)}$, the

observation v_k is more likely to be emitted in *state 1* (0.80) and the observation v_l is more likely to be emitted in *state 2* (0.70).

The insufficiency of classical HMM which exists in the literature in some real life problems has motivated the study of different types of HMMs with extended assumptions. For this purpose, in the following chapter 3 and chapter 4, two different types of expanded and modified HMMs with some extra Markov assumptions are proposed. One of them is the Markov-emitted HMM (MEHMM) which the observations are assumed to have a Markov dependence. And, the other is the binary-dependent HMM (BDHMM) which is based on the binary dependency of (X_t, Y_t) to (X_{t-1}, Y_{t-1}) .

CHAPTER 3

MARKOV-EMITTED HIDDEN MARKOV MODEL

The main contribution of the introduced Markov-emitted hidden Markov model (MEHMM) is to develop an alternative model to decrease the loss of information that may arise from the assumption of conditionally i.i.d. observations. MEHMM is capable of capturing a possible first order Markov dependence between observations and may provide better representation for some real-life problems rather than HMM where observations are not independent random variables. MEHMM is proposed with the assumption that each observation depends not only on the related hidden state but also the previous observation. And also, the hidden state transitions are independent of the observations. State transitions occur according to a Markov chain. So, the proposed model takes into account more dependency information than classical HMM. The parameter estimation procedure that is used for estimating the most likely hidden state sequence and the structure of MEHMM are similar to classical HMM. However, the proposed model has one extra parameter and hence, the algorithms used for estimation need to be modified according to the Markov dependence assumption between observations.

The proposed model considers the transitions between successive observations. When in addition to the transitions between states, a Markov dependence between successive observations really exist, modeling with classical HMM would not be a correct approach and MEHMM will be an alternative way of modeling. Moreover, one of the most important point of MEHMM is that; the proposed model allows us to identify if there is a first order Markov dependence between observations that affected by the state transitions. For an observation sequence of interest, if MEHMM outperforms than HMM based on the AIC and BIC values, it is concluded that there is a first order Markov dependence between observations which is ignored by modeling with HMM.

The contribution of this section is three-fold.

1. A modified HMM, which uses more information from the past is proposed.
2. Baum-Welch and Viterbi algorithms are modified under the assumption of the first order Markov dependence between observations.
3. The proposed model allows us to reveal the first-order Markov dependence between the current and the previous observations of the system if there is.

3.1 Model Definition

Let X_t be the random variable that represents the hidden state of the Markov chain at time t , with discrete state space $\Omega = \{1, 2, \dots, N\}$. Let Y_t be the random variable that represents the observation symbol in a state at time t , where $Y_t \in \{v_1, v_2, \dots, v_M\}$ and v_k be the k -th observation and M possible observations can be emitted for each state. So, the assumptions of Markov dependence and independence from the observations are also valid for MEHMM. The conditional probability distribution of a current state depends only on the most recent state because of the Markovian property. That is,

$$P(X_t | X_{t-1}, X_{t-2}, \dots, X_1, Y_{t-1}, Y_{t-2}, \dots, Y_1) = P(X_t | X_{t-1}), \quad (3.1)$$

where $t = 1, 2, \dots, T$.

Since the observations are assumed to have first order Markov dependence, the conditional probability distribution of observations depends on both the previous observation symbol and the successive hidden states. That is,

$$P(Y_t | X_t, X_{t-1}, \dots, X_1, Y_{t-1}, Y_{t-2}, \dots, Y_1) = P(Y_t | X_t, X_{t-1}, Y_{t-1}), \quad (3.2)$$

where $t = 1, 2, \dots, T$.

Besides, the conditional probability of the observation sequence $\{Y_t, Y_{t+1}, \dots, Y_T\}$ is dependent only on the previous hidden state and observation $\{X_{t-1}, Y_{t-1}\}$,

$$\begin{aligned} P(Y_t, Y_{t+1}, \dots, Y_T | X_1, X_2, \dots, X_{t-1}, Y_1, Y_2, \dots, Y_{t-1}) \\ = P(Y_t, Y_{t+1}, \dots, Y_T | X_{t-1}, Y_{t-1}), \end{aligned} \quad (3.3)$$

where $t = 1, 2, \dots, T$.

The transition structure of MEHMM between the hidden states and the observations are shown in Figure 3.1.

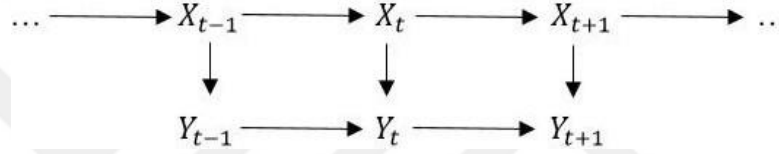


Figure 3.1 Graphical illustration of MEHMMs

In order to characterize MEHMM, following elements are needed to describe completely. The complete parameter set of HMM is $\lambda = (P, B, \pi, \phi)$;

The transition probability distribution. The transition probability matrix is denoted by $\mathbf{P} = \{p_{ij}\}$ and consists of $N \times N$ elements. Each element of $\mathbf{P}$ are transition probabilities from hidden state i to state j . That is,

$$p_{ij} = P(X_{t+1} = j | X_t = i), \quad (3.4)$$

where $i, j \in \Omega$. Note that p_{ij} provides the properties given in (2.4) and (2.5).

The observation emission distribution. The observation distribution matrix is denoted by $\mathbf{B}$ and consists of $N \times N$ sub-matrices. Each sub-matrix of $\mathbf{B}$ is an M -dimensional square matrix. Each entry of this matrix is the probability of emitting

observation v_l at time t , given that the process makes a transition from state i to state j and observation v_k is emitted at time $t - 1$. That is,

$$b_{ij}(v_k, v_l) = P\{Y_t = v_l | X_{t-1} = i, X_t = j, Y_{t-1} = v_k\}, \quad (3.5)$$

where $i, j \in \Omega$, $v_k, v_l \in \mathcal{O}$, $k, l = 1, 2, \dots, M$, and $t = 1, 2, \dots, T$.

The matrix representation of the observation emission distribution is presented below. Let sub-matrix $\mathbf{B}_{i,j}$ defined as follows:

$$\mathbf{B}_{i,j} = \begin{bmatrix} b_{i,j}(v_1, v_1) & b_{i,j}(v_1, v_2) & \cdots & b_{i,j}(v_1, v_M) \\ b_{i,j}(v_2, v_1) & b_{i,j}(v_2, v_2) & \cdots & b_{i,j}(v_2, v_M) \\ \vdots & \vdots & \ddots & \vdots \\ b_{i,j}(v_M, v_1) & b_{i,j}(v_M, v_2) & \cdots & b_{i,j}(v_M, v_M) \end{bmatrix}.$$

Then the matrix $\mathbf{B}$ takes the form of

$$\mathbf{B} = \begin{bmatrix} [\mathbf{B}_{11}] & [\mathbf{B}_{12}] & \cdots & [\mathbf{B}_{1N}] \\ [\mathbf{B}_{21}] & [\mathbf{B}_{22}] & \cdots & [\mathbf{B}_{2N}] \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{B}_{N1}] & [\mathbf{B}_{N2}] & \cdots & [\mathbf{B}_{NN}] \end{bmatrix}.$$

All the elements of $\mathbf{B}$ are greater than or equal to zero. In addition, the sum of each row of each matrix $\mathbf{B}_{i,j}$ is equal to one. That is,

$$b_{ij}(v_k, v_l) \geq 0 \text{ for all } i, j \in \Omega, v_k, v_l \in \mathcal{O}; \quad (3.6)$$

$$\sum_{l=1}^M b_{ij}(v_k, v_l) = 1 \text{ for all } i, j \in \Omega, v_k, v_l \in \mathcal{O}. \quad (3.7)$$

The initial state distribution. The probabilities related to the initial states of the system can be represented by N dimensional row vector. Let π_j be the probability of starting at state j . Then

$$\pi_j = P(X_0 = j). \quad (3.8)$$

since π_j is a probability and the process is certain to start in one of the states, the sum of the initial probabilities are equal to one, that is, $\sum_{j=1}^N \pi_j = 1$.

The initial observation distribution. It is related to the initial emission probabilities of observations with the condition of the initial state distribution. M possible observations can be emitted for each state; and there are N different hidden states so, let ϕ be an $N \times M$ dimensional matrix which gives the probabilities of emitting observations for the initial state. That is,

$$\phi_j(v_k) = P\{Y_0 = v_k | X_0 = j\}, \quad (3.9)$$

where $j \in \Omega$, $v_k \in \mathcal{O}$, and $k = 1, 2, \dots, M$.

The matrix representation of the initial observation distribution is as follows,

$$\phi = \begin{bmatrix} \phi_1(v_1) & \phi_1(v_2) & \cdots & \phi_1(v_M) \\ \phi_2(v_1) & \phi_2(v_2) & \cdots & \phi_2(v_M) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N(v_1) & \phi_N(v_2) & \cdots & \phi_N(v_M) \end{bmatrix}.$$

Note that, all elements of ϕ are greater than or equal to zero, and the sum of the row probabilities equals to one. That is,

$$\phi_j(v_k) \geq 0 \text{ for all } j \in \Omega, v_k \in \mathcal{O}; \quad (3.10)$$

$$\sum_{k=1}^M \phi_j(v_k) = 1 \text{ for all } j \in \Omega, v_k \in \mathcal{O}. \quad (3.11)$$

The differences between complete parameter sets for classical HMM and proposed MEHMM are given in Table 3.1.

Table 3.1 Complete parameter sets of HMM and MEHMM

Parameter	HMM	MEHMM
P	$p_{ij} = P(X_{t+1} = j X_t = i)$	$p_{ij} = P(X_{t+1} = j X_t = i)$
B	$b_j(k) = P(Y_t = v_k X_t = j)$	$b_{ij}(v_k, v_l) = P\{Y_t = v_l X_t = j, X_{t-1} = i, Y_{t-1} = v_k\}$
π	$\pi_j = P(X_0 = j)$	$\pi_j = P(X_0 = j)$
ϕ	—	$\phi_j(v_k) = P\{Y_0 = v_k X_0 = j\}$

3.2 Fundamental Problems of MEHMM

As it is known from classical HMM, the model has three fundamental problems namely the evaluation, learning and decoding problems. However, according to the assumptions of proposed MEHMM, the algorithms used to solve these problems need to be modified.

3.2.1 Evaluation Problem

The aim of the evaluation is to compute the probability that an observation sequence is produced by given MEHMM. In other words, the likelihood of the observation sequence $P(Y_1, Y_2, \dots, Y_T | \lambda)$ is computed. The evaluation problem is solved by forward or backward probabilities (Stratonovich 1960).

Let $\alpha_j(t)$ be the forward probability which is defined as the probability of being in state j at time t having seen the observation sequence $\{Y_1, Y_2, \dots, Y_t\}$ where $t \leq T$. Forward probabilities calculated for each step are greater than or equal to zero; namely, $\alpha_j(t) \geq 0$, for all $j \in \Omega$.

$$\alpha_j(t) = P(Y_1, Y_2, \dots, Y_t, X_t = j | \lambda) \quad (3.12)$$

$$= \sum_{i=1}^N P(Y_1, Y_2, \dots, Y_{t-1}, Y_t, X_{t-1} = i, X_t = j | \lambda)$$

$$\begin{aligned}
&= \sum_{i=1}^N P(Y_1, Y_2, \dots, Y_{t-1} = v_k, X_{t-1} = i | \lambda) \\
&\quad \cdot P(X_t = j, Y_t = v_l | Y_1, Y_2, \dots, Y_{t-1} = v_k, X_{t-1} = i) \\
&= \sum_{i=1}^N \alpha_i(t-1) P(X_t = j, Y_t = v_l | X_{t-1} = i, Y_{t-1} = v_k) \\
&= \sum_{i=1}^N \alpha_i(t-1) \frac{P(X_{t-1} = i, X_t = j, Y_{t-1} = v_k, Y_t = v_l)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&\quad \frac{P(X_{t-1} = i)P(X_t = j | X_{t-1} = i)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&\quad \cdot P(Y_{t-1} = v_k | X_t = j, X_{t-1} = i) \\
&= \sum_{i=1}^N \alpha_i(t-1) \frac{P(Y_t = v_l | X_t = j, X_{t-1} = i, Y_{t-1} = v_k)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&\quad \frac{P(X_{t-1} = i)p_{ij}P(Y_{t-1} = v_k | X_{t-1} = i)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&= \sum_{i=1}^N \alpha_i(t-1) \frac{P(Y_t = v_l | X_t = j, X_{t-1} = i, Y_{t-1} = v_k)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&\quad \frac{p_{ij}P(X_{t-1} = i, Y_{t-1} = v_k)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&= \sum_{i=1}^N \alpha_i(t-1) \frac{P(Y_t = v_l | X_t = j, X_{t-1} = i, Y_{t-1} = v_k)}{P(X_{t-1} = i, Y_{t-1} = v_k)} \\
&= \sum_{i=1}^N \alpha_i(t-1) p_{ij} b_{ij}(v_k, v_l)
\end{aligned}$$

$$\Rightarrow \alpha_j(t) = \begin{cases} \pi_j \varphi_j(Y_0 = v_l) & , t = 0 \\ \sum_{i=1}^N \alpha_i(t-1) p_{ij} b_{ij}(v_k, v_l) & , t = 1, 2, \dots, T \end{cases} \quad (3.13)$$

where $i, j \in \Omega$, $v_k, v_l \in \mathcal{O}$, and $k, l = 1, 2, \dots, M$.

Based on the definition of the forward probability, the likelihood value of the model in terms of $\alpha_j(t)$ is

$$L(\lambda) = P(Y_1, Y_2, \dots, Y_T) \quad (3.14)$$

$$\begin{aligned} &= \sum_{j=1}^N P(Y_1, Y_2, \dots, Y_T, X_T = j) \\ &= \sum_{j=1}^N \alpha_j(T), \end{aligned}$$

where T is the length of the observation sequence.

Let $\beta_j(t)$ be the backward probability which is defined as the probability of observing $\{Y_{t+1}, Y_{t+2}, \dots, Y_T\}$, given that the current state and observation symbol. Backward probabilities calculated for each step are greater than or equal to zero; namely, $\beta_j(t) \geq 0$, for all $j \in \Omega$.

$$\beta_j(t) = P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_t, \lambda) \quad (3.15)$$

$$\begin{aligned} &= \sum_{i=1}^N P(Y_{t+2}, \dots, Y_T, Y_{t+1} = v_l, X_{t+1} = i | X_t = j, Y_t = v_k, \lambda) \\ &= \sum_{i=1}^N P(X_{t+1} = i, Y_{t+1} = v_l | X_t = j, Y_t = v_k) \\ &\quad \cdot P(Y_{t+2}, \dots, Y_T | X_{t+1} = i, Y_{t+1} = v_l, X_t = j, Y_t = v_k) \\ &= \sum_{i=1}^N P(X_{t+1} = i | X_t = j, Y_t = v_k) \\ &\quad \cdot P(Y_{t+1} = v_l | X_t = j, X_{t+1} = i, Y_t = v_k) \\ &\quad \cdot P(Y_{t+2}, \dots, Y_T | X_{t+1} = i, Y_{t+1} = v_l) \\ &= \sum_{i=1}^N P(X_{t+1} = i | X_t = j) b_{ji}(v_k, v_l) \beta_i(t+1) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N p_{ji} b_{ji}(v_k, v_l) \beta_i(t+1) \\
\Rightarrow \beta_j(t) &= \begin{cases} \sum_{i=1}^N p_{ji} b_{ji}(v_k, v_l) \beta_i(t+1) & , t = 1, 2, \dots, T-1 \\ 1 & , t = T \end{cases}, \quad (3.16)
\end{aligned}$$

where $i, j \in \Omega$, $v_k, v_l \in \mathcal{O}$, and $k, l = 1, 2, \dots, M$.

The computational space and time complexities for MEHMM are the same as HMM. Since forward and backward probabilities are calculated for each value of t , ($t \in T$) and also for each state i , ($i \in \Omega$) the space complexity of MEHMM is $\mathcal{O}(NT)$. In addition, each forward or backward probability consists of the sum of the number of terms as many as the number of hidden states. So, the time complexity of the algorithm for the evaluation problem is $\mathcal{O}(N^2T)$.

The posterior marginal probabilities of hidden states $\{X_1, X_2, \dots, X_T\}$ at any time, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$, is calculated by the same formula as in HMM. That is,

$$\begin{aligned}
P(X_t = j | Y_1, Y_2, \dots, Y_T) &\underset{\tilde{X}_t}{\propto} P(X_t = j, Y_1, Y_2, \dots, Y_T) \\
&= \beta_j(t) \alpha_j(t).
\end{aligned} \quad (3.17)$$

3.2.2 Learning Problem

The aim of the learning procedure is to find the most likely set of state transition and observation emission probabilities, given a set of output sequences and general structure of MEHMM (number of hidden states). Forward and backward probabilities are defined as the same in classical HMM for MEHMM, however the computation is different. Hence, the calculation of the new estimates in maximization part of Baum-Welch algorithm makes the difference. It is known that Baum-Welch algorithm (Baum

et al., 1970) is used to estimate the complete parameter set $\lambda = (P, B, \pi, \phi)$ of MEHMM for the solution of learning problem.

First of all, initial values of the complete parameter set must be defined. For the Baum-Welch algorithm, some probabilistic definitions must be provided.

The general idea for calculating the new estimates of the state transition probabilities p_{ij} is given below,

$$p_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions out of state } i}.$$

The general idea for calculating the new estimates of the observation emission probabilities $b_{ij}(v_k, v_l)$ is given below,

$$b_{ij}(v_k, v_l) = \frac{\text{expected number of transitions from state } i \text{ to } j \text{ with outputs } v_k \text{ and } v_l, \text{ respectively}}{\text{expected number of transitions out of state } i \text{ with output } v_k \text{ to state } j}.$$

The general idea for calculating the new estimates of the initial state probabilities π_j is given below,

$$\pi_j = \text{expected frequency in state } j \text{ at time } t = 1.$$

Let the random variable $\xi_{ij}(t)$ be the probability of being in state i at time t and in state j at time $t + 1$, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$. Let $L(\lambda)$ be the likelihood of the model which is calculated as given in (3.14). $\xi_{ij}(t)$ can be expressed in terms of forward and backward probabilities.

$$\begin{aligned} \xi_{ij}(t) &= P(X_t = i, X_{t+1} = j | Y_1, Y_2, \dots, Y_T) \\ &= \frac{P(X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \end{aligned} \tag{3.18}$$

$$= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1}) \cdot P(X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1})}{L(\lambda)}$$

$$= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1}) \cdot P(X_{t+1} = j, Y_{t+1} = v_l | X_t = i, Y_1, Y_2, \dots, Y_t = v_k) \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i)}{L(\lambda)}$$

$$= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j, Y_{t+1}) \cdot P(X_{t+1} = j, Y_{t+1} = v_l | X_t = i, Y_t = v_k) \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i)}{L(\lambda)}$$

$$= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j, Y_{t+1}) \cdot P(X_{t+1} = j | X_t = i, Y_t = v_k) \cdot P(Y_{t+1} = v_l | X_t = i, X_{t+1} = j, Y_t = v_k) \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i)}{L(\lambda)}$$

$$= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j, Y_{t+1}) \cdot P(X_{t+1} = j | X_t = i) \cdot P(Y_{t+1} = v_l | X_t = i, X_{t+1} = j, Y_t = v_k) \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i)}{L(\lambda)}$$

$$= \frac{\alpha_i(t) p_{ij} b_{ij}(v_k, v_l) \beta_j(t+1)}{L(\lambda)}$$

Note that the probabilities $\xi_{ij}(t)$ for each step are greater than or equal to zero and sum of the probabilities $\xi_{ij}(t)$ for all states at time t equals to one. That is,

$$\xi_{ij}(t) \geq 0 \text{ for all } i, j \in \Omega; \quad (3.19)$$

$$\sum_{i=1}^N \sum_{j=1}^N \xi_{ij}(t) = 1, \quad (3.20)$$

where $t = 1, 2, \dots, T-1$.

Let the random variable $\gamma_j(t)$ be defined as the probability of being in state j at time t , given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$.

$$\begin{aligned}
\gamma_j(t) &= P(X_t = j | Y_1, Y_2, \dots, Y_T) \\
&= \frac{P(X_t = j, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \\
&= \frac{P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_1, Y_2, \dots, Y_t)}{P(X_t = j, Y_1, Y_2, \dots, Y_t)} \\
&= \frac{L(\lambda)}{L(\lambda)} \\
&= \frac{P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_t)}{P(X_t = j, Y_1, Y_2, \dots, Y_t)} \\
&= \frac{L(\lambda)}{L(\lambda)} \\
&= \frac{\alpha_j(t)\beta_j(t)}{L(\lambda)}
\end{aligned} \tag{3.21}$$

Note that sum of the probabilities $\gamma_j(t)$ for all $j \in \Omega$ at time t is equal to one and each is greater than or equal to zero. That is,

$$\gamma_j(t) \geq 0 \text{ for all } j \in \Omega; \tag{3.22}$$

$$\sum_{j=1}^N \gamma_j(t) = 1, \tag{3.23}$$

where $t = 1, 2, \dots, T$.

Now, let us consider the expected values defined in terms of probabilities $\xi_{ij}(t)$ and $\gamma_j(t)$. We have two new expected values in addition to the expected values which are defined in (2.26), (2.27), and (2.31).

Let $E(N_{i \rightarrow j}^{v_k \rightarrow v_l})$ be the expected number of times observation v_k occurs in state i and observation v_l occurs in state j after a transition. It can be calculated as follows,

$$E(N_{i \rightarrow j}^{v_k \rightarrow v_l}) = \sum_{\substack{t=1 \\ Y_{t+1}=v_l \\ Y_t=v_k}}^{T-1} \xi_{ij}(t). \quad (3.24)$$

Let $E(N_{i \rightarrow j}^{v_k \rightarrow})$ be the expected number of times that observation v_k occurs in state i and a transition is occurred to state j . It can be obtained as follows,

$$E(N_{i \rightarrow j}^{v_k \rightarrow}) = \sum_{\substack{t=1 \\ Y_t=v_k}}^{T-1} \xi_{ij}(t). \quad (3.25)$$

In the second step, called the maximization-step, updated values of the parameters are obtained by using the expected values. The estimated value of the transition probabilities for the n th iteration is

$$p_{ij}^{(n)} = \frac{E(N_{i \rightarrow j})}{E(N_{i \rightarrow})} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}. \quad (3.26)$$

The estimated value of the observation emission distribution for the n th iteration is

$$b_{ij}^{(n)}(v_k, v_l) = \frac{E(N_{i \rightarrow j}^{v_k \rightarrow v_l})}{E(N_{i \rightarrow j}^{v_k \rightarrow})} = \frac{\sum_{\substack{t=1 \\ Y_{t+1}=v_l \\ Y_t=v_k}}^{T-1} \xi_{ij}(t)}{\sum_{\substack{t=1 \\ Y_t=v_k}}^{T-1} \xi_{ij}(t)}. \quad (3.27)$$

And, the estimated value of the initial state distribution for the n th iteration is

$$\pi_j^{(n)} = E(N_j(1)) = \gamma_j(1). \quad (3.28)$$

For the proposed model, the general idea of initial observation distribution $\varphi_j(v_k)$ is

$$\varphi_j(v_k) = \frac{\text{sum of probabilities of a transition to state } j \text{ and observing } v_k}{\text{sum of probabilities of a transition to state } j \text{ and observing any output}}.$$

To calculate the probabilities $\varphi_j(v_k)$, $\mathbf{B}$ matrix is used. Actually, this general idea is the limiting distribution of observation symbols, and $\varphi_j(v_k)$ gives the probability distribution of initial observation emissions in the long term. So, the maximization step of the initial observation distribution depends on the matrix $\mathbf{B}$. That is,

$$\varphi_j^{(n)}(v_k) = \frac{\sum_{i=1}^N \sum_{k=1}^M b_{ij}^{(n)}(v_k, v_l)}{\sum_{i=1}^N \sum_{l=1}^M \sum_{k=1}^M b_{ij}^{(n)}(v_k, v_l)}. \quad (3.29)$$

The iterations are continued until the absolute difference between the parameter estimates obtained two successive iteration; or the likelihood values obtained at $(n-1)$ th and n th iteration falls below a predefined value ϵ .

3.2.3 Decoding Problem

The aim of the decoding procedure is to estimate the most probable state transition sequence, given the observation sequence. In other words, given the observation sequence $\{Y_0, Y_1, \dots, Y_T\}$, and the model parameters $\lambda = (P, B, \pi, \phi)$, how do we choose a corresponding state sequence path which is optimal in some meaningful sense (i.e., best “explains” the observations)? The most common and effective way to solve this problem is the Viterbi algorithm (Viterbi, 1967).

Let S be hidden state sequence and S^* be the “best” possible hidden state sequence. Then the goal is to maximize $P(S^*|Y_0, Y_1, \dots, Y_T)$ by choosing the “best” S . That is,

$$S^* = \underset{S}{\operatorname{argmax}} [P(S|Y_0, Y_1, \dots, Y_T)]. \quad (3.30)$$

The following equation can be achieved by using the conditional probability and the chain rule

$$\begin{aligned} P(S|Y_0, Y_1, \dots, Y_T) &= \frac{P(S, Y_0, Y_1, \dots, Y_T)}{P(Y_0, Y_1, \dots, Y_T)} \\ &= \frac{P(S)P(Y_0, Y_1, \dots, Y_T|S)}{L(\lambda)} \end{aligned} \quad (3.31)$$

$P(Y_0, Y_1, \dots, Y_T)$ is the likelihood $L(\lambda)$ of the model and stay constant over time. So $P(S|Y_0, Y_1, \dots, Y_T)$ is proportional to the joint probability distribution $P(S, Y_0, Y_1, \dots, Y_T)$. That is,

$$P(S|Y_0, Y_1, \dots, Y_T) \propto_S P(S) \cdot P(Y_0, Y_1, \dots, Y_T|S). \quad (3.32)$$

So, the “best” possible hidden state sequence (S^*) can be estimated as follows,

$$S^* = \underset{S}{argmax} [P(S)P(Y_0, Y_1, \dots, Y_T|S)]. \quad (3.33)$$

The joint probability of the hidden state sequence $P(S)$ can be calculated as follows, note that X_0 is the initial state of the Markov chain.

$$\begin{aligned} P(S) &= P(X_0, X_1, \dots, X_T) \\ &= P(X_0)P(X_1|X_0)P(X_2|X_0, X_1) \dots \\ &\quad \cdot P(X_T|X_1, X_2, \dots, X_{T-1}) \\ &= P(X_0)P(X_1|X_0)P(X_2|X_1) \dots P(X_T|X_{T-1}) \end{aligned} \quad (3.34)$$

The conditional probability of the observation sequence $\{Y_0, Y_1, \dots, Y_T\}$ given the hidden state sequence $\{X_0, X_1, \dots, X_T\}$ can be calculated as follows,

$$P(Y_0, Y_1, \dots, Y_T|S) = \frac{P(Y_0, Y_1, \dots, Y_T, S)}{P(S)} \quad (3.35)$$

$$\begin{aligned}
& \frac{P(S)P(Y_0|S)P(Y_1|S, Y_0) \dots}{P(S)} \\
& = \frac{P(Y_T|S, Y_0, Y_1, \dots, Y_{T-1})}{P(S)} \\
& = P(Y_0|X_0)P(Y_1|X_0, X_1, Y_0) \dots P(Y_T|X_{T-1}, X_T, Y_{T-1}).
\end{aligned}$$

Therefore, the term in equation (3.33) forms the following,

$$\begin{aligned}
& P(S)P(Y_0, Y_1, \dots, Y_T|S) \tag{3.36} \\
& = P(X_0)P(X_1|X_0)P(X_2|X_1) \dots P(X_T|X_{T-1}) \\
& \quad \cdot P(Y_0|X_0)P(Y_1|X_0, X_1, Y_0) \dots P(Y_T|X_{T-1}, X_T, Y_{T-1}) \\
& = P(X_0 = r)P(Y_0 = v_m|X_0 = r) \\
& \quad \cdot \prod_{t=1}^T P(X_t = j|X_{t-1} = i) P(Y_t = v_l|X_{t-1} = i, X_t = j, Y_{t-1} = v_k) \\
& = \pi_r \varphi_r(v_m) \prod_{t=2}^T p_{ij} b_{ij}(v_k, v_l),
\end{aligned}$$

where $i, j, r \in \Omega$, $v_k, v_l, v_m \in \mathcal{O}$, and $k, l, m = 1, 2, \dots, M$.

The most likely hidden state is estimated for each step of the observation sequence by using the formula (3.36). The “best” possible hidden state sequence is the most likely state sequence that should emit the observation sequence.

3.3 Experimental Study

The performance of MEHMM is examined by using two different types of observation sequence in the experimental study. One of them is an observation sequence which is generated from MEHMM with some arbitrary initial complete parameter sets. The second type of data sequence is generated from a classical HMM with conditionally independent and identically distributed observations. Generated

data sequences are considered as the observation sequences and 1000 data sequences are generated for both cases. Sample size is determined as $n=500$; that is there are 500 observations in each observation sequence. Comparison is made for different number of states, as well as for different number of observation symbols. The types of observation sequences used in the experimental study are given in Table 3.2.

Table 3.2 Types of observation sequences

Model Selected	Number of Hidden States (N)	Number of Observations (M)
HMM	2	2
	2	3
	3	2
	3	3
MEHMM	2	2
	2	3
	3	2
	3	3

All observation sequences are modeled by using classical HMM and also proposed MEHMM. Uniform distribution is used for the initial values of the parameters except the observation emission distributions and the parameter estimates are obtained by the Baum-Welch algorithm. Algorithm stops when the absolute difference between the current and the previous estimates is less than $\epsilon = 10^{-4}$, for all parameters. Parameter estimates are obtained by using both modeling approach for each observation data set. Moreover, likelihood values ($\hat{L}$) for each model are calculated, as well.

The most common way of measuring the performance of a model relative to others is calculating the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) values. The AIC and BIC values are calculated with the formulas given below,

$$AIC = -2 \ln(\hat{L}) + 2p, \quad (3.37)$$

$$BIC = -2 \ln(\hat{L}) + p \ln(T), \quad (3.38)$$

where p is the number of freely estimated parameters and T represents the length of observation sequence. AIC and BIC estimate values are calculated for all models and also for each of 1000 data sets. Smaller AIC and BIC values obtained by MEHMM considered as “success” for observation sequences which are generated from MEHMM. On the other hand, smaller AIC and BIC values obtained by classical HMM considered as “success” for observation sequences which are generated from classical HMM. The number of successes are obtained for each cases for 1000 observation sequences.

The performance of MEHMM can now be evaluated after defining and generating the observation sequences. There are four models to be tested, which differ according to the number of states and observations. AIC, BIC and number of success values are calculated for each model to measure the performance. The numerical results are gathered on Table 3.3 and Table 3.4, respectively.

Table 3.3 Results obtained by modeling the observation sequences which are generated based on MEHMM assumptions

Model		# of freely estimated parameters		Mean and std.dev. of AIC		Mean and std.dev. of BIC		# of success for MEHMM		Mean relative increment of MEHMM (%)	
N	M	HMM	MEHMM	HMM	MEHMM	HMM	MEHMM	AIC	BIC	AIC	BIC
2	2	5	7	723.3422 (6.1329)	690.2711 (5.4682)	747.8809 (6.1329)	724.6253 (5.4682)	1000	994	4.5719	3.1095
2	3	7	11	1098.36 (12.4671)	1062.87 (11.4515)	1132.71 (12.4671)	1116.85 (11.4515)	1000	957	3.2311	1.4001
3	2	11	14	722.2773 (7.5462)	688.5634 (6.8711)	776.2626 (7.5462)	757.2719 (6.8711)	991	949	4.6677	2.4464
3	3	14	20	1101.23 (7.9132)	1055.19 (5.9482)	1169.93 (7.9132)	1153.34 (5.9482)	952	927	4.1807	1.4180

The observation sequences, which are generated based on MEHMM assumptions, were modeled with both HMM and MEHMM. The results are given in Table 3.3. The

number of states (N) may be two or three; and in each state the number of symbols (M) that can be observed may take the values of two or three. The number of freely estimated parameters changes according to the number of states and observation symbols, for HMM and MEHMM. The number of freely estimated parameters of MEHMM are higher than that of HMM because the proposed model has one more additional parameter. The third and fourth columns give the mean AIC and BIC values and their standard deviations obtained by using HMM and MHMM, respectively. It is seen that the mean AIC, mean BIC and their standard deviations obtained by using MEHMM are smaller compared to the values that are obtained by using classical HMM, although the number of free parameters are high. In the fifth column of the table, the number of successes in terms of AIC and BIC compared to HMM is given. In the Table 3.3, since the results for observation sequences generated from MEHMM are given, obtaining smaller AIC and BIC values for proposed MEHMM has been defined as ‘success’. Hence, according to the AIC values, success is achieved in all 1000 samples for 2-state models. On the other hand, when the number of hidden state is 3, the number of success is 991 for 2-observation-model and 952 for 3-observation-model. According to the BIC values, MEHMM again outperforms rather than HMM. The number of success range from 927 to 994. Finally, the last column gives the mean relative increment of AIC and BIC values obtained by using MEHMM relative to HMM. A positive mean relative increment means that the performance of proposed model is better since AIC values are lower compared to HMM. AIC values decreased by at least 3.2311% (2-state, 3-obs.) and at most 4.6677% (3-state, 2 obs.) compared to HMM. Besides, BIC values decreased by at least 1.4001% (2-state, 3-obs.) and the most 3.1095% (2-state, 2 obs.) compared to HMM. The best results by percentage and the best performance improvements are obtained for 3-state-2-observation model according to AIC and 2-state-2-observation according to BIC compared to classical HMM.

Table 3.4 Results obtained by modeling the observation sequences which are generated based on HMM assumptions

Model		# of freely estimated parameters		Mean and std.dev. of AIC		Mean and std.dev. of BIC		# of success for MEHMM		Mean relative increment of MEHMM (%)	
N	M	HMM	MEHMM	HMM	MEHMM	HMM	MEHMM	AIC	BIC	AIC	BIC
2	2	5	7	685.2376 (4.1871)	699.3381 (4.9947)	709.7763 (4.1871)	733.6923 (4.9947)	2	0	-2.0577	-3.3695
2	3	7	11	1054.85 (5.4548)	1063.23 (6.1449)	1089.20 (5.4548)	1117.21 (6.1449)	92	4	-0.7944	-2.5716
3	2	11	14	697.6769 (6.8893)	709.1468 (8.7459)	751.6622 (6.8893)	777.8553 (8.7459)	23	6	-1.6440	-3.4846
3	3	14	20	1049.58 (4.7132)	1055.88 (4.9519)	1118.28 (4.7132)	1154.03 (4.9519)	78	0	-0.6002	-3.1968

The observation sequences, which are generated based on HMM assumptions, were modeled with both HMM and MEHMM. The results are given in Table 3.4. From Table 3.4, it can be seen that the mean AIC, mean BIC and their standard deviations obtained by using HMM are less than those obtained by using MEHMM for all type of models. This time, ‘success’ is the number of observation sequence in which AIC and BIC values for MEHMM is less than classical HMM. In fact, Table 3.4 shows that the number of success is at most 92 for the AIC values for the 2-state-3-observation model and at most 6 for the BIC values for the 3-state-2-observation model. MEHMM is never a better model according to BIC values for 2-state-2-observation and 3-state-3-observation models. The success percentage does not exceed 10% for all types of models for MEHMM. It is seen from the last column that the mean relative increment (%) of AIC and BIC values obtained by MEHMM relative to HMM is negative for all models. This means that the AIC and BIC values of MEHMM are higher than that of obtained by HMM and the performance of the proposed model is lower than HMM for observation sequences generated from HMM, as expected. Mean relative increment values obtained from MEHMM increased by at least 0.6002% (3-state, 3-obs.) and at most 2.0577% (2-state, 2-obs.) according to AIC, also increased by at least 2.5716% (2-state, 3-obs.) and at most 3.4846% (3-state, 2-obs.) according to BIC, compared to the values obtained by HMM.

To reach the conclusion that one model outperforms than another, it may not be enough just to look at the number of successes. In order to interpret the difference between AIC and BIC values for HMM and MEHMM, a method, based on the weights of expected information loss, is used. Detailed information for this method can be achieved from Burnham & Anderson (2002). It is a quick strength-of-evidence interpreting technique between models. Let Δ_i be the difference of AIC or BIC values of HMM and MEHMM for the two types of 1000 observation sequences. That is,

$$\Delta_i^{AIC} = AIC_i - AIC_{min}, \quad (3.39)$$

$$\Delta_i^{BIC} = BIC_i - BIC_{min}, \quad (3.40)$$

where AIC_{min} and BIC_{min} are the smallest value among HMM and MEHMM to be compared. If the difference $\Delta_i \leq 2$, substantial evidence is supported to say that both HMM and MEHMM fit the observation sequence at the same quality. If the difference is between $4 \leq \Delta_i \leq 7$, less evidence is supported. If the difference $\Delta_i \geq 10$, there is substantial evidence to say that one is better than the other. Then, by the help of Δ_i , Akaike and Bayesian weights (w_i^{AIC} and w_i^{BIC}) are calculated by using respective formulas given by Burnham & Anderson (2002). That is,

$$w_i^{AIC} = \frac{e^{(-\frac{1}{2}\Delta_i^{AIC})}}{\sum_{r=1}^R e^{(-\frac{1}{2}\Delta_r^{AIC})}}, \quad (3.41)$$

$$w_i^{BIC} = \frac{e^{(-\frac{1}{2}\Delta_i^{BIC})}}{\sum_{r=1}^R e^{(-\frac{1}{2}\Delta_r^{BIC})}}, \quad (3.42)$$

where, R is the number of models to be compared. The sum of Akaike and Bayesian weights (weight of evidence) are equal to 1 for all models compared. These weights are used to interpret how much a model represents data relative to the other. The mean

values $\bar{\Delta}$ and $\bar{w}$ for HMM and MEHMM which are calculated from the 1000 observation sequences are given in Table 3.5 and Table 3.6, respectively.

Table 3.5 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of MEHMM

Number of Parameter		Observation sequences generated based on the assumptions of MEHMM					
N	M	$\bar{\Delta}^{AIC}$	$\bar{\Delta}^{BIC}$	$\bar{w}_{MEHMM}^{AIC}$	$\bar{w}_{HMM}^{AIC}$	$\bar{w}_{MEHMM}^{BIC}$	$\bar{w}_{HMM}^{BIC}$
2	2	33.0711	23.2556	0.9999	0.0001	0.9999	0.0001
2	3	35.4900	15.8600	≈ 1	≈ 0	0.9996	0.0004
3	2	33.7139	18.9907	0.9999	0.0001	0.9999	0.0001
3	3	46.0400	16.5900	≈ 1	≈ 0	0.9997	0.0003

It is clearly seen from Table 3.5 that, for the group of observation sequences generated from MEHMM, the mean increase in HMM's AIC values vary between 33.0711 and 46.0400. The mean increase in HMM's BIC values varies between 15.8600 and 23.2556. According to both AIC and BIC values, we have exactly no evidence to support that data fits HMM. $\bar{\Delta}^{AIC}$ and $\bar{\Delta}^{BIC}$ values are much greater than 10 for all models. Hence, we can conclude that modeling the observation sequences generated from MEHMM gives poor results for HMM. In addition, it is observed that the mean Akaike weights for MEHMM vary between 0.9999 and ≈ 1 and the mean Bayesian weights for MEHMM vary between 0.9996 and 0.9999.

Table 3.6 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of HMM

Number of Parameter		Observation sequences generated based on the assumptions of HMM					
N	M	$\bar{\Delta}^{AIC}$	$\bar{\Delta}^{BIC}$	$\bar{w}_{MEHMM}^{AIC}$	$\bar{w}_{HMM}^{AIC}$	$\bar{w}_{MEHMM}^{BIC}$	$\bar{w}_{HMM}^{BIC}$
2	2	14.1005	23.9160	0.0009	0.9991	0.0001	0.9999
2	3	8.3800	28.0100	0.0150	0.9850	0.0001	0.9999
3	2	11.4699	26.1931	0.0033	0.9967	0.0001	0.9999
3	3	6.3000	35.7500	0.0411	0.9589	≈ 0	≈ 1

Table 3.6 shows that for the group of observation sequences generated from classical HMM, the mean increase in MEHMM's AIC values vary between 6.3000 and 14.1005. The mean increase in MEHMM's BIC values vary between 23.9160 and 35.7500. We have less evidence to support 3-state-3-observation MEHMM and almost no evidence to support 2-state-3-observation MEHMM relative to HMM according to mean AIC differences. And also, there is exactly no evidence to support for 2-state-2-observation MEHMM and 3-state-2-observation MEHMM relative to HMM according to mean AIC differences. There is no evidence to support all MEHMMs relative to HMM according to mean BIC differences. Hence, we can conclude that modeling the observation sequences generated from HMM gives poor results for MEHMM. It is observed that the mean Akaike weights for MEHMM vary between 0.0009 and 0.0411 and the mean Bayesian weights for MEHMM vary between 0,0001 and ≈ 1 . These weights can be expressed as the expected number of the best representative models and given in Table 3.7. Values in parentheses are the number of successes for MEHMM obtained in the previous analysis.

Table 3.7 Comparison of the number of successes and the expected number of more likely model for MEHMM

Number of Parameter		Observation sequences generated based on the assumptions of MEHMM		Observation sequences generated based on the assumptions of HMM	
N	M	AIC	BIC	AIC	BIC
2	2	999 (1000)	≈ 999 (994)	≈ 1 (2)	≈ 0 (0)
2	3	1000 (1000)	≈ 999 (957)	≈ 15 (92)	≈ 0 (4)
3	2	999 (991)	≈ 999 (949)	≈ 3 (23)	≈ 0 (6)
3	3	1000 (952)	≈ 999 (927)	≈ 41 (78)	≈ 0 (0)

The number of success achieved is only based on the lower AIC and BIC values, decrements in AIC and BIC values are not considered. However, by computing the Akaike and Bayesian weights, significance of the decrement in AIC and BIC values can also be taken into account. We have even stronger support of substantial evidence to say that, MEHMM models the observation sequences generated from MEHMM much better than HMM, and HMM models the observation sequences generated from

HMM much better than MEHMM. The reason is that, for MEHMM generate observation sequences, expected number of best representative models are greater than the number of successes and much closer to 1000. For HMM generate observation sequences, all the expected number of the best representative models are smaller than the number of successes and much closer to zero. To sum up, when the differences of AIC and BIC are taken into consideration, MEHMM and HMM show much higher performance on observation sequences conforming to their assumptions.

From these comparisons, we may draw a conclusion that; the proposed model gives better results and much outperforms than classical HMM when there exists a first order Markov dependence between the observation symbols. However, the proposed model is insufficient for conditionally independent observations compared to HMM.

3.4 A Case Study on Stock Price Trend Probability Forecasting

One of the areas where HMM is applied is the modeling and forecasting of stock prices. Some of the studies in the literature (de Angelis & Paas, 2013; Hassan & Nath, 2005; Nguyen, 2018; Zhang, Li, Wang, Fang, & Yu, 2019) assume that the stock prices corresponding to the observations are continuous random variables and follow Gaussian distribution. A discrete-time HMM in which each day is taken as the time unit and stock prices are grouped is presented by Kavitha, Udhayakumar, & Nagarajan (2013).

In this study, a real-life stock price closing value sequence is modeled by using both HMM and MEHMM where time units are taken as days and observations are categorized as increasing or decreasing. A comparison is provided. The data set is obtained from the official website of the stock exchange market of the National Association of Securities Dealers Automated Quotations (www.nasdaq.com). Data set includes 1195 closing values of the stock prices of Microsoft Corporation for each day from 01/01/2016 to 30/09/2020. Figure 3.2 shows the time-series plot of stock price closing values of the Microsoft Corporation between the chosen dates.

From Figure 3.2, it is seen that there is an increasing trend over the years. In order to eliminate the effect of the trend and to ensure a first order Markov dependence between closing values, instead of using the closing values directly, the first difference is taken and then the relative change values are calculated for each day. Hence, the length of the observation sequence is now 1194. Let CV_t be the closing value of the stock price of day t . The relative change values can be calculated for all days with the following formula.

$$\text{Relative change (\%)} = \frac{CV_t - CV_{t-1}}{CV_{t-1}} \times 100 \quad (3.43)$$

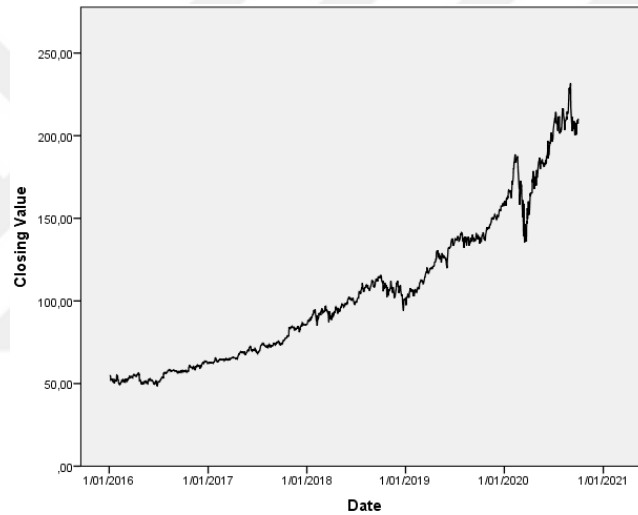


Figure 3.2 Time-series plot of stock price closing values of Microsoft Corporation

Figure 3.3 shows the time-series plot of the first-difference values of closing values. The first-difference values are actually the numerator of the equation (3.43). It is seen from Figure 3.3 that there is no trend in the first-difference data, but there exists a heteroscedasticity problem since the variability of the first difference values increases over time.

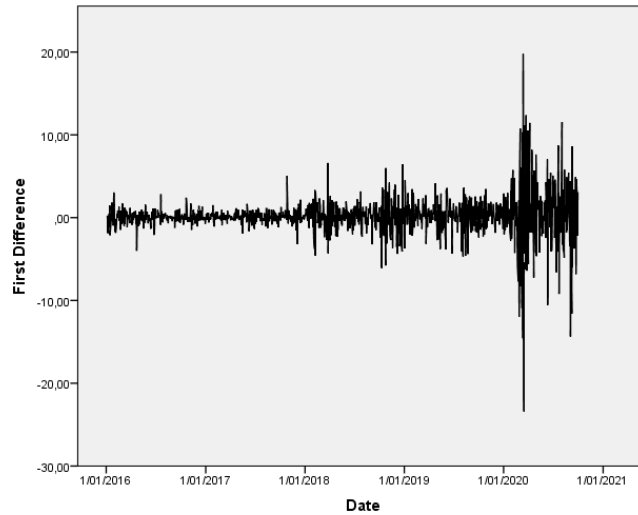


Figure 3.3 Time-series plot of the first-difference values

Relative changes are deal with the change rate of the first difference values rather than the amount of change. Therefore, the heteroscedasticity problem can be eliminated by considering the relative change values. The relative change values for all days are shown in Figure 3.4.

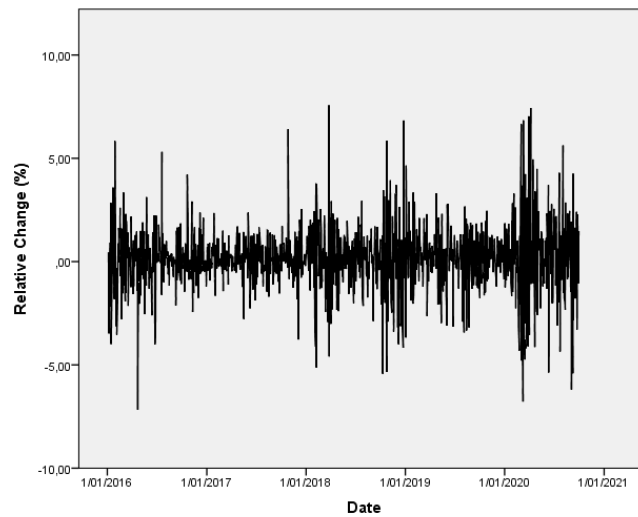


Figure 3.4 Time-series plot of the relative change (%) values

It is seen from the Figure 3.4 that after the relative changes are calculated, the variability of the relative change values does not change over time and randomly distributed. The aim is to model of these change rates with HMM and MEHMM and

to reveal if there exists a first order Markov dependence between successive relative changes while the system makes transitions between unobservable (hidden) states.

The hidden states are considered as market conditions on any given day. Let *state 1* and *state 2* be bad and good market conditions for the 2-state model, respectively. Besides, let *state 1*, *state 2*, and *state 3* be bad, moderate, and good market conditions for the 3-state model, respectively. It is assumed that market conditions are not observable to the researcher and directly affect the closing value of the stock prices. *Obs. 1* corresponds to the decrease rate of the closing value compared to the previous day and *obs. 2* corresponds to the increase rate of the closing value compared to the previous day for the 2-observation model. Negative relative changes are defined as *obs. 1* and positive relative changes are defined as *obs. 2* for 2-observation model.

3.4.1 Model Selection

HMM assumes that the trend in the closing price of a given day conditionally depends only on current market conditions. However, today's trend in the closing value may also likely be dependent on the previous day's trend in the closing value. For this purpose, the observation sequence with the length of $T = 1194$ is modeled by using both HMM and MEHMM in order to reveal the relationship between trend in the closing values on consecutive days. In order to model with HMM and MEHMM, four different models varying according to the number of parameters are constructed; which are 2-state-2-observation HMM, 2-state-2-observation MEHMM, 3-state-2-observation HMM and 3-state-2-observation MEHMM. The models which are compared are given in the Table 3.8.

Table 3.8 Types of observation sequences

Dependency Structure	Number of Hidden States (N)	Number of Observations (M)
HMM	2	2
	3	2
MEHMM	2	2
	3	2

By changing the initial values each time and selecting them arbitrarily, model parameters of HMM and MEHMM are estimated many times. The algorithm stops for all parameters of the models when the absolute differences between the current and the previous estimate values are less than $\epsilon = 10^{-4}$. The Baum-Welch algorithm converges to the optimum values of parameter estimates in 47 iterations for the 2-state-2-observation HMM, in 63 iterations for the 2-state-2-observation MEHMM, in 34 iterations for the 3-state-2-observation HMM, in 41 iterations for the 3-state-2-observation MEHMM. Table 3.9 shows the log-likelihood, AIC, BIC, weight values and ranks of the models.

Table 3.9 Comparison of the models

Model Type	# of States	# of Obs.	P	$\ln L$	AIC (Rank)	w_t^{AIC}	BIC (Rank)	w_t^{BIC}
HMM	2	2	5	-810.9313	1637.8626 (4)	≈ 0	1663.2879 (3)	≈ 0
MEHMM	2	2	7	-796.2477	1606.4954 (2)	0.0002	1642.0908 (1)	0.9999
HMM	3	2	11	-794.3561	1610.7122 (3)	0.0001	1666.6479 (4)	≈ 0
MEHMM	3	2	14	-781.0198	1590.0396 (1)	0.9997	1661.2305 (2)	0.0001

From AIC and BIC values of four models, given in Table 3.9, it is clear that MEHMM gives better results than HMM for both 2-state-2-observation and 3-state-3-observation models. Relative decrease in AIC value is 1.9525%, in BIC value is 1.2908% for the 2-state-2-observation MEHMM. Besides, relative decrease in AIC value is 1.3002% and in BIC is 0.3261% for the 3-state-2-observation MEHMM. The smallest AIC value is obtained for 3-state-2-observation MEHMM and the smallest

BIC values is obtained for the 2-state-2-observation MEHMM. This means that MEHMM characterizes the observation sequence better for both 2-state and 3-state models. Furthermore, it can be concluded that the relative change of the closing value in the current day has a first-order Markov dependence on the previous days' relative change. In addition, MEHMMs take the best two places (rank 1 and rank 2) according to the AIC and BIC values. While calculating the BIC value, the amount of penalty resulting from the number of free parameters is high compared to the AIC value. Besides, the weight value (w_i^{BIC}) is more dominant (0.9999), when analyzes are made with the 2-state-2-observation MEHMM. According to the Akaike weights, the 3-state-2-observation MEHMM is obviously noticeable as the best representative model. According to Akaike weights, 9997 times the best representative model will more likely be the 3-state-2-observation MEHMM out of 10000 modeling repeats. And also, according to the Bayesian weights, the 2-state-2-observation MEHMM is obviously noticeable as the best representative model. According to Bayesian weights 9999 times the best representative model will more likely be the 2-state-2-observation MEHMM out of 10000 modeling repeats.

The final estimates of the complete parameter set of the 2-state-2-observation MEHMM is given as follows,

$$\hat{B} = \begin{matrix} & & \text{State 1} & & \text{State 2} & & \\ & & \text{Obs. 1} & \text{Obs. 2} & \text{Obs. 1} & \text{Obs. 2} & \\ \text{State 1} & \text{Obs. 1} & [0.7695 & 0.2305] & \text{Obs. 1} & [0.4010 & 0.5990] \\ & \text{Obs. 2} & [0.8510 & 0.1490] & \text{Obs. 2} & [0.4220 & 0.5780] \\ \text{State 2} & \text{Obs. 1} & [0.1900 & 0.8100] & \text{Obs. 1} & [0.2882 & 0.7118] \\ & \text{Obs. 2} & [0.4956 & 0.5044] & \text{Obs. 2} & [0.1368 & 0.8632] \end{matrix},$$

$$\hat{P} = \begin{matrix} & \text{State 1} & \text{State 2} \\ \text{State 1} & [0.4515 & 0.5485] \\ \text{State 2} & [0.4370 & 0.5630] \end{matrix},$$

$$\hat{\pi} = \begin{matrix} & \text{State 1} & \text{State 2} \\ \begin{bmatrix} 0 & 1 \end{bmatrix}, \end{matrix}$$

$$\hat{\phi} = \begin{matrix} & \text{Obs. 1} & \text{Obs. 2} \\ \begin{matrix} \text{State 1} \\ \text{State 2} \end{matrix} \begin{bmatrix} 0.5085 & 0.4915 \\ 0.1172 & 0.8828 \end{bmatrix}. \end{matrix}$$

According to matrix $\hat{P}$, the hidden state transition probability matrix, if the market conditions are bad for a given day, the market conditions will be bad on the next day with the probability of 0.4515 and market conditions will be good on the next day with the probability of 0.5485. On the other hand, if the market conditions are good for a given day, the market conditions will be bad on the next day with the probability of 0.4370 and market conditions will be good on the next day with the probability of 0.5630. According to the matrix $\hat{B}$, the observation emission distribution, given that the unobservable market conditions are both bad for consecutive days, it is likely to occur that the relative change will be negative in the next day. As seen in the matrix, the probability of a transition from *obs. 1* to *obs. 1* is 0.7695 and the probability of a transition from *obs. 1* to *obs. 2* is 0.8510. The opposite situation occurs when the unobservable market conditions are good for two consecutive days. If the relative change at the end of a given day is either positive or negative, an increase in the relative change occurs at the end of the next day. If given that the unobservable market conditions are not same for consecutive days, there is no particularly noticeable transition probability distribution and the observation emission distribution matrices are closer to uniform. Apart from these, when we look at the initial distributions, the market conditions at the beginning of the process are definitely good with probability 1. If the process starts with bad market conditions, *obs. 1* is emitted with a probability of 0.5085 and *obs. 2* is emitted with a probability of 0.4915. Besides, if the process starts with good market conditions, *obs. 1* is emitted with a probability of 0.1172 and *obs. 2* is very likely be emitted with a probability of 0.8828.

The hidden state sequence with a length of $T = 1194$ is estimated using the final values of parameter estimates by the modified Viterbi algorithm. *State 1* (bad market

condition) appears 552 times and *state 2* (good market condition) appears 642 times in the hidden sequence. In addition, a negative relative change (*obs. 1*) is emitted 535 times and a positive relative change (*obs. 2*) is emitted 659 times at the end of a day. Table 3.10 shows how many times *obs. 1* and *obs. 2* occur in the hidden states.

Table 3.10 Matching of hidden decoded states and observation sequence

	Number of obs.	State 1	State 2
Obs. 1	535	529	6
Obs. 2	659	23	636

It can be seen from Table 3.10 that, bad market conditions appeared 535 times, and for 529 of them, the end-of-day closing value decreased compared to the previous day. On the other hand, good market conditions appear 659 times, and almost for each of them (636 times), the end-of-day closing value increased compared to the previous day. In the presence of bad market conditions, the increase in relative change occurred only 23 times. As a result, the pattern of observation sequence and state sequence is consistent as expected. When the market conditions are good, the stock price tends to increase compared to the previous day, whereas the market conditions are bad, the stock prices tend to decrease compared to the previous day.

In order to show how good the obtained parameter estimates fit the observation sequence, an observation sequence with a length of $T = 1194$ is generated using the last estimation values both for 2-state-2-observation HMM and MEHMM, and the transition probability matrix between observations is obtained and given by the matrices $\mathbf{O}_{gen}^{HMM}$ and $\mathbf{O}_{gen}^{MEHMM}$. Besides, the observation transition probabilities of the actual observation sequence which is between 01/01/2016 and 30/09/2020, is also constructed. That is,

$$\mathbf{O}_{gen}^{HMM} = \begin{matrix} & \begin{matrix} \text{Obs. 1} & \text{Obs. 2} \end{matrix} \\ \begin{matrix} \text{Obs. 1} \\ \text{Obs. 2} \end{matrix} & \begin{bmatrix} 0.4513 & 0.5487 \\ 0.4297 & 0.5703 \end{bmatrix} \end{matrix},$$

$$\mathbf{O}_{gen}^{MEHMM} = \begin{array}{cc} & \begin{array}{cc} \text{Obs. 1} & \text{Obs. 2} \end{array} \\ \begin{array}{c} \text{Obs. 1} \\ \text{Obs. 2} \end{array} & \begin{bmatrix} 0.4190 & 0.5810 \\ 0.5256 & 0.4744 \end{bmatrix} \end{array},$$

$$\mathbf{O} = \begin{array}{cc} & \begin{array}{cc} \text{Obs. 1} & \text{Obs. 2} \end{array} \\ \begin{array}{c} \text{Obs. 1} \\ \text{Obs. 2} \end{array} & \begin{bmatrix} 0.3925 & 0.6075 \\ 0.4939 & 0.5061 \end{bmatrix} \end{array}.$$

A goodness of fit test for testing the null hypothesis of equality of the generated and the actual observation transition probability matrices is performed (Bhat & Miller, 2002). The null hypothesis is $H_0: O = O_{gen}$ and the test statistics is

$$\sum_{i=1}^M \sum_{j=1}^M \frac{T_i [O_{gen}(i, j) - O(i, j)]^2}{O(i, j)} \sim \chi_{M(M-1)-d}^2, \quad (3.44)$$

where d is the number of zero elements in $\mathbf{O}$ and T_i is the number of the i th observation in the generated observation sequence.

The test statistic for the 2-state-2-observation MEHMM is $\chi_{test}^{2MEHMM} \cong 2.0794$, the test statistic for the 2-state-2-observation HMM is $\chi_{test}^{2HMM} \cong 9.0759$, and the critical value is $\chi_{0.05,2}^2 = 5.9915$. Calculated test statistic does not exceed the critical value and the null hypothesis cannot be rejected for the 2-state-2-observation MEHMM. The difference between the observed and the estimated values are not significant for 2-state-2-observation MEHMM. The result is strengthening our intuition that the estimated model parameters are not local maximum and good convergence is provided. On the other hand, calculated test statistic exceeds the critical value and the null hypothesis is rejected for the 2-state-2-observation HMM. It is concluded that, by using 2-state-2-observation HMM, there is a significant difference between the observation transition matrix of actual data and the observation transition matrix of generated data. Hence, modeling with a 2-state-2-observation MEHMM fits the observation sequence better relative to the classical 2-state-2-observation HMM for the actual data of 1194 relative change values of interest.

3.4.2 Bootstrap Confidence Intervals for the 2-State-2-Obs. MEHMM Parameters

In this section, confidence intervals (CI) are calculated for the 2-state-2-observation MEHMM parameters with bootstrap method (Efron & Tibshirani, 1993). Visser, Raijmakers, & Molenaar (2000) well describes the bootstrap technique for constructing confidence intervals for HMM parameters. The procedure is described below:

1. A new observation sequence of the same size as the actual data is generated using the final estimates of the fitted model.
2. The generated data is modeled by using MEHMM and model parameters are estimated.
3. These steps are repeated for n times, where n is the number of bootstrap samples. Let $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_n$ be the estimates of the complete parameter sets of n generated data sets. Since the final estimates of MEHMM are maximum likelihood estimates, each parameter estimates of $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_n$ has an empirical distribution around the true parameters. The standard error, $\hat{\sigma}_{\lambda}$, for each parameter can be obtained by calculating the standard deviation of the empirical distribution. Then the approximate 95% confidence intervals are obtained with the formula $\hat{\lambda} \pm (1.96) \hat{\sigma}_{\lambda}$.

95% CIs are constructed for each parameter by generating observation sequences with a length of 500, 1000 and 10000. The final estimate values of 2-state-2-observation MEHMM are used for generating the observation sequences. Each observation has a length of $T = 1194$ which is the same as with the actual observation sequence. The estimates and standard deviations of related estimate for the parameters $\mathbf{B}$, and $\mathbf{P}$ and $\boldsymbol{\phi}$ are given in Table 3.11, Table 3.12 and Table 3.13, respectively. There is no variability in the estimates of initial state distribution, and the process starts from *state 2* with a probability of 1, for all generated observation sequences. So, confidence intervals can not be constructed for the parameter π .

Table 3.11 Estimate values and estimated standard deviations of the parameter B

Parameter	$b_{11}(1, 1)$	$b_{11}(1, 2)$	$b_{11}(2, 1)$	$b_{11}(2, 2)$
Estimate	0.7695	0.2305	0.8510	0.1490
M	$\hat{\sigma}_B$			
500	0.0744	0.0744	0.0718	0.0718
1000	0.0521	0.0521	0.0637	0.0637
10000	0.0486	0.0486	0.0428	0.0428
Parameter	$b_{12}(1, 1)$	$b_{12}(1, 2)$	$b_{12}(2, 1)$	$b_{12}(2, 2)$
Estimate	0.4010	0.5990	0.4220	0.5780
M	$\hat{\sigma}_B$			
500	0.1129	0.1129	0.1365	0.1365
1000	0.1067	0.1067	0.1072	0.1072
10000	0.0833	0.0833	0.0701	0.0701
Parameter	$b_{21}(1, 1)$	$b_{21}(1, 2)$	$b_{21}(2, 1)$	$b_{21}(2, 2)$
Estimate	0.1900	0.8100	0.4956	0.5044
M	$\hat{\sigma}_B$			
500	0.0883	0.0883	0.1468	0.1468
1000	0.0724	0.0724	0.1289	0.1289
10000	0.0512	0.0512	0.0916	0.0916
Parameter	$b_{22}(1, 1)$	$b_{22}(1, 2)$	$b_{22}(2, 1)$	$b_{22}(2, 2)$
Estimate	0.2882	0.7118	0.1368	0.8632
M	$\hat{\sigma}_B$			
500	0.1045	0.1045	0.0885	0.0885
1000	0.0829	0.0829	0.0658	0.0658
10000	0.0593	0.0593	0.0425	0.0425

Table 3.12 Estimate values and estimated standard deviations of the parameter P

Parameter	p_{11}	p_{12}	p_{21}	p_{22}
Estimate	0.4515	0.5485	0.4370	0.5630
M	$\hat{\sigma}_\phi$			
500	0.1416	0.1416	0.1339	0.1339
1000	0.1269	0.1269	0.1012	0.1012
10000	0.0884	0.0884	0.0752	0.0752

Table 3.13 Estimate values and estimated standard deviations of the parameter ϕ

Parameter	$\phi_1(1)$	$\phi_1(2)$	$\phi_2(1)$	$\phi_2(2)$
Estimate	0.5085	0.4915	0.1172	0.8828
M	$\hat{\sigma}_\phi$			
500	0.1168	0.1168	0.1019	0.1019
1000	0.0943	0.0943	0.0826	0.0826
10000	0.0681	0.0681	0.0582	0.0582

As the number of generated observation sequences increases, the 95% CI for the related parameter gets narrow, since the standard deviation decreases. In addition, as the values of the parameter estimates approach to 0 or 1, the standard deviation value of the related estimate decreases.

3.4.3 Results

For the 2-state-2-observation MEHMM, the estimated observation transition probabilities are obtained by the marginalization and normalization of the $\hat{B}$ matrix according to the hidden states and given below.

$$\begin{array}{cc} & \begin{array}{cc} \text{Obs. 1} & \text{Obs. 2} \end{array} \\ \begin{array}{c} \text{Obs. 1} \\ \text{Obs. 2} \end{array} & \begin{bmatrix} 0.4121 & 0.5879 \\ 0.4763 & 0.5237 \end{bmatrix} \end{array}$$

From the estimated observation transition matrix, if the relative change of the stock price for day $t-1$ is negative, the relative change of the stock price will be again negative with a probability of 0.4121 and positive with a probability of 0.5879 for day t . Besides, if the relative change of the stock price for day $t-1$ is positive, the relative change of the stock price will be negative with a probability of 0.4763 and again positive with a probability of 0.5237 for day t . It is more likely to be a positive relative change today, regardless of the previous day's relative change. But, a decrease in relative change for two days in a row is less likely. If there is a decrease one day, the next day it is more likely to increase than a increase for two consecutive days. The given matrix is a transition matrix of an ergodic and irreducible Markov chain, and the steady-state probabilities for each state are estimated and obtained as follows:

$$\Pi = \begin{array}{cc} & \text{Obs. 1} & \text{Obs. 2} \\ \text{Obs. 1} & 0.4476 & 0.5524 \end{array}.$$

The steady-state probabilities indicate that in the long term, the relative change of the stock price at the end of a given day is negative with a probability of 0.4476 and positive with a probability of 0.5524.

Let m_{ii} be the expected number of transitions before the system returns state i , given that the current state is i . m_{ii} is called the *mean first return time*. For an ergodic irreducible Markov chain, mean first return times can be calculated from $m_{ii} = 1/\Pi_i$ and obtained as follows:

$$\begin{array}{cc} m_{11} & m_{22} \\ 2.2341 & 1.8102 \end{array}.$$

If the relative change of the stock price is negative at the end of a given day, it repeats for the first time after a mean of 2.2341 days. On the other hand, if the relative change of the stock price is positive at the end of a given day, it repeats for the first time after a mean of 1.8102 days. It can be seen from this result that the closing value of the stock price is more likely to increase compared to the previous day and the mean return time is less than two days.

CHAPTER 4

BINARY-DEPENDENT HIDDEN MARKOV MODEL

The main contribution of the introduced Binary-dependent hidden Markov model (BDHMM) is to develop an alternative model to decrease the loss of information that may arise from the assumption of conditionally i.i.d. observations and hidden states which are dependent only on the previous hidden state. BDHMM is capable of capturing a possible first-order Markov dependence between the last observed binary couple of random variables (X_t, Y_t) and the previously observed binary couple (X_{t-1}, Y_{t-1}) . Since more information is used by BDHMM, it may provide better representation for some of the appropriate real-life problems compared with HMM and also MEHMM. It is assumed that the conditional distribution of the current state and observation depends on both the previous state and previous observation. In other words, a binary dependence is proposed between observations and hidden states, different from Berchtold (1999) and Li (2005). The most recent hidden state and emitted observation are Markov dependent on the previous hidden state and observation. The most important point in the proposed model is that, unlike the other extended or modified HMMs, transitions between hidden states do not occur independently of observation symbols. Hidden states and observations make interrelated transitions depending on a joint probability distribution.

The contribution of this section is three-fold.

- A binary-dependent HMM, which uses more information from the past compared to HMM and MEHMM, is proposed.
- Baum-Welch and Viterbi algorithms are modified according to the binary dependency assumption.
- The proposed model allows us to reveal the first-order dependence between the current and the previous binary information of the system if there is.

The parameter estimation procedure that is used for estimating the most likely hidden state sequence, and the structure of BDHMM are similar to classical HMM. However, the complete parameter set of the proposed model differs from HMM and MEHMM hence, the algorithms used for estimation need to be modified according to the Markov dependence assumption between the current and previous binary information of the process.

If there is a real Markov dependence between successive steps of the process, modeling with classical HMM would not be appropriate way and BDHMM will be an alternative way of modeling. Moreover, one of the most important point of BDHMM is that; the proposed model allows us to identify if there is a first order Markov dependence between the last and the previous binary informations. For an observation sequence of interest, if BDHMM outperforms than HMM based on the AIC and BIC values, it is concluded that there is a first order Markov dependence between the successive binary steps which is ignored in HMM.

4.1 Model Definition

Let X_t be the random variable that represents the hidden state of the Markov chain at time t , with discrete state space $\Omega = \{1, 2, \dots, N\}$. For a given state i ($i \in \Omega$), an observation symbol is emitted. Let v_k be the k th observation symbol and there are M possible observation symbols for each state. The system makes a transition and emits one of the observation symbols in the set of $\mathcal{O} = \{v_1, v_2, \dots, v_M\}$. Let Y_t be the random variable that represents the observation symbol in a state at time t , where $Y_t \in \mathcal{O}$. Transitions between states are determined by a Markov chain; according to an N -dimensional transition probability matrix, as in classical HMM. The conditional probability distribution of a current state depends only on the most recent state because of the Markovian property; given below.

$$P(X_t | X_{t-1}, X_{t-2}, \dots, X_1, Y_{t-1}, Y_{t-2}, \dots, Y_1) = P(X_t | X_{t-1}), \quad (4.1)$$

where $t = 1, 2, \dots, T$.

Since the current system information $\{X_t, Y_t\}$ is assumed to have a first-order Markov dependence with the previous system information $\{X_{t-1}, Y_{t-1}\}$, observation transition distribution depends on both the current state and the previous system information. That is,

$$P(X_t, Y_t | X_{t-1}, \dots, X_1, Y_{t-1}, Y_{t-2}, \dots, Y_1) = P(X_t, Y_t | X_{t-1}, Y_{t-1}), \quad (4.2)$$

where $t = 1, 2, \dots, T$.

And the conditional probability of the observation sequence $\{Y_t, Y_{t+1}, \dots, Y_T\}$ given the past, depends only on the previous system information $\{X_{t-1}, Y_{t-1}\}$,

$$P(Y_t, Y_{t+1}, \dots, Y_T | X_1, X_2, \dots, X_{t-1}, Y_1, Y_2, \dots, Y_{t-1}) = P(Y_t, Y_{t+1}, \dots, Y_T | X_{t-1}, Y_{t-1}), \quad (4.3)$$

where $t = 1, 2, \dots, T$.

The transition structure of BDHMM between the hidden states and the observations are shown in Figure 4.1.

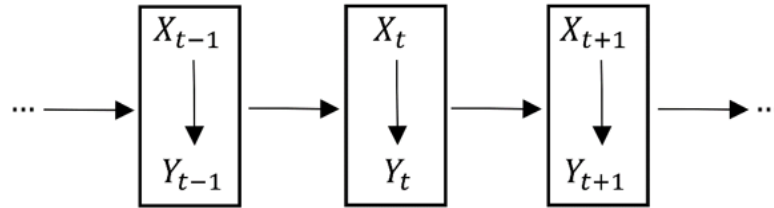


Figure 4.1 Graphical illustration of BDHMMs

In order to characterize BDHMM, following elements are needed to describe completely. The complete parameter set of BDHMM is $\lambda = (\Psi, \pi, \phi)$;

The state-observation probability distribution. The system releases an observation after each transition according to the observation probability distribution. The

probability of producing t -th system information $\{X_t, Y_t\}$, may change according to the previous system information $\{X_{t-1}, Y_{t-1}\}$. Hence, the state-observation probability matrix is a $(N \times M) \times (N \times M)$ dimensional square matrix and denoted by Ψ . Matrix Ψ gives the probability of emitting an observation v_l at state j given the previous system information. That is,

$$\psi_{ij}(v_k, v_l) = P(X_t = j, Y_t = v_l | X_{t-1} = i, Y_{t-1} = v_k), \quad (4.4)$$

where $i, j \in \Omega$, $v_k, v_l \in \mathcal{O}$, $k, l = 1, 2, \dots, M$, and $t = 1, 2, \dots, T$.

Let the sub-matrix Ψ_{v_k, v_l} defined as below:

$$\Psi_{v_k, v_l} = \begin{bmatrix} \psi_{11}(v_k, v_l) & \psi_{12}(v_k, v_l) & \dots & \psi_{1N}(v_k, v_l) \\ \vdots & \ddots & & \vdots \\ \psi_{N1}(v_k, v_l) & \psi_{N2}(v_k, v_l) & \dots & \psi_{NN}(v_k, v_l) \end{bmatrix}.$$

Then the matrix Ψ takes the form of

$$\Psi = \begin{bmatrix} \Psi_{v_1, v_1} & \Psi_{v_1, v_2} & \dots & \Psi_{v_1, v_M} \\ \vdots & \ddots & & \vdots \\ \Psi_{v_M, v_1} & \Psi_{v_M, v_2} & \dots & \Psi_{v_M, v_M} \end{bmatrix}.$$

All the elements of $\Psi = \{\psi_{ij}(v_k, v_l)\}$ are greater than or equal to zero. In addition, the sum of the probabilities for each row is equal to one. That is,

$$\psi_{ij}(v_k, v_l) \geq 0, \quad i, j \in \Omega, v_k, v_l \in \mathcal{O}; \quad (4.5)$$

$$\sum_{l=1}^M \sum_{j=1}^N \psi_{ij}(v_k, v_l) = 1 \quad \text{for all } i \in \Omega, v_k \in \mathcal{O}. \quad (4.6)$$

The initial state distribution. It is an N -dimensional row vector which is denoted by π_j for state j , and gives the initial probabilities for each state. Then

$$\pi_j = P(X_0 = j), \quad (4.7)$$

where $\pi_j \geq 0$, and $\sum_{j=1}^N \pi_j = 1$ for $j = 1, \dots, N$.

The initial observation distribution. It is related to the initial emission probabilities of observations given the initial state distribution. There are N hidden states, and M possible observation can be produced in each state. Hence, let $\boldsymbol{\phi}$ be an $M \times N$ dimensional matrix that gives the probabilities of emitting observations for the initial state. That is,

$$\varphi_j(v_k) = P(Y_0 = v_k | X_0 = j), \quad (4.8)$$

where $j \in \Omega$, $v_k \in \mathcal{O}$, and $k = 1, 2, \dots, M$. The matrix representation of the initial observation distribution is as follows,

$$\boldsymbol{\phi} = \begin{bmatrix} \varphi_1(v_1) & \varphi_1(v_2) & \cdots & \varphi_1(v_M) \\ \varphi_2(v_1) & \varphi_2(v_2) & \cdots & \varphi_2(v_M) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_N(v_1) & \varphi_N(v_2) & \cdots & \varphi_N(v_M) \end{bmatrix}.$$

Note that, all elements of the $\boldsymbol{\phi}$ are greater than or equal to zero, and the sum of each row is equal to one. That is,

$$\varphi_j(v_k) \geq 0, \quad j \in \Omega, v_k \in \mathcal{O}; \quad (4.9)$$

$$\sum_{k=1}^M \varphi_j(v_k) = 1 \text{ for all } j \in \Omega, v_k \in \mathcal{O}. \quad (4.10)$$

The differences between complete parameter sets for classical HMM and proposed BDHMM are given in Table 4.1.

Table 4.1 Complete parameter sets of HMM and BDHMM

Parameter	HMM	BDHMM
P	$p_{ij} = P(X_{t+1} = j X_t = i)$	—
B	$b_j(k) = P(Y_t = v_k X_t = j)$	—
π	$\pi_i = P(X_0 = i)$	$\pi_i = P(X_0 = i)$
ϕ	—	$\phi_j(v_k) = P\{Y_0 = v_k X_0 = j\}$
Ψ	—	$\psi_{ij}(v_k, v_l) = P\{X_t = j, Y_t = v_l X_{t-1} = i, Y_{t-1} = v_k\}$

Different from classical HMM, according to the assumptions of BDHMM, the information obtained from only X_{t-1} is insufficient to obtain the conditional distribution of the $\{X_t, Y_t\}$. The reason is that, the hidden state and its related observation is conditionally dependent on also previous observation. The probability of an observation value at time t also varies according to the observation value at time $(t - 1)$. If a Markov dependence really exists between $\{X_{t-1}, Y_{t-1}\}$ and $\{X_t, Y_t\}$ for a given observation sequence and hidden state sequence, then a researcher who uses classical HMM can reveal only the impact of a hidden state on the related observation $P(Y_t | X_t)$. Hence, the relationship between binary variables, that is the impact of the previous observation on the current observation are ignored. As shown in Table 4.1, an observation symbol Y_{t-1} is emitted right after the related hidden state X_{t-1} . When an observation Y_{t-1} is emitted, it is known that the related hidden state X_{t-1} is made a transition. This means that before making a transition to the hidden state X_t , $\{X_{t-1}, Y_{t-1}\}$ is occurred and now X_{t-1} is hidden information and also Y_{t-1} is observable information. The occurrence of the information $\{X_{t-1}, Y_{t-1}\}$, and the next step of the system is modeled through the state-observation probability distribution (Ψ).

4.2 Fundamental Problems of BDHMM

As it is known from classical HMM, the model has three fundamental problems which are evaluation, learning and decoding problems. But, according to the

assumptions of proposed BDHMM, the algorithms used to solve these problems need to be modified.

4.2.1 Evaluation Problem

The aim of the evaluation is to compute the probability that an observation sequence is produced by given BDHMM. In other words, the likelihood of the observation sequence $P(Y_1, Y_2, \dots, Y_T | \lambda)$ is computed. The evaluation problem is solved by forward or backward probabilities (Stratonovich 1960). Forward and backward probabilities are defined as the same in MEHMM for the modified procedure, however, the computation is different. Since BDHMM contains a new type of parameter, the Baum-Welch algorithm needs to be modified. The calculation of the new estimates in the maximization part makes the difference.

Let $\alpha_j(t)$ be the forward probability which is defined as the probability of being in state j at time t having seen the observation sequence $\{Y_1, Y_2, \dots, Y_t\}$ where $t \leq T$. Forward probabilities calculated for each step are greater than or equal to zero; that is $\alpha_j(t) \geq 0$, for all $j \in \Omega$.

$$\alpha_j(t) = P(Y_1, Y_2, \dots, Y_t, X_t = j | \lambda) \quad (4.11)$$

$$\begin{aligned} &= \sum_{i=1}^N P(Y_1, Y_2, \dots, Y_{t-1}, Y_t, X_{t-1} = i, X_t = j | \lambda) \\ &= \sum_{i=1}^N P(Y_1, Y_2, \dots, Y_{t-1} = v_k, X_{t-1} = i | \lambda) \\ &\quad \cdot P(X_t = j, Y_t = v_l | Y_1, Y_2, \dots, Y_{t-1} = v_k, X_{t-1} = i) \\ &= \sum_{i=1}^N \alpha_i(t-1) P(X_t = j, Y_t = v_l | X_{t-1} = i, Y_{t-1} = v_k) \\ &= \sum_{i=1}^N \alpha_i(t-1) \psi_{ij}(v_k, v_l) \end{aligned}$$

$$\Rightarrow \alpha_j(t) = \begin{cases} \pi_j \varphi_j(Y_0 = v_l) & , t = 1 \\ \sum_{i=1}^N \alpha_i(t-1) \psi_{ij}(v_k, v_l) & , t = 2, \dots, T \end{cases} \quad (4.12)$$

where $i, j \in \Omega$, $v_k, v_l \in \mathcal{O}$, and $k, l = 1, 2, \dots, M$.

Based on the definition of the forward probability, the likelihood value of the model in terms of $\alpha_j(t)$ is,

$$\begin{aligned} L(\lambda) &= P(Y_1, Y_2, \dots, Y_T) \\ &= \sum_{j=1}^N P(Y_1, Y_2, \dots, Y_T, X_T = j) \\ &= \sum_{j=1}^N \alpha_j(T), \end{aligned} \quad (4.13)$$

where T is the length of the observation sequence.

Let $\beta_j(t)$ be the *backward probability* which is also redefined as the probability of observing $Y_{t+1}, Y_{t+2}, \dots, Y_T$, given that the current state and observation symbol. Backward probabilities calculated for each step are greater than or equal to zero; that is $\beta_j(t) \geq 0$, for all $j \in \Omega$.

$$\beta_j(t) = P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_t, \lambda) \quad (4.14)$$

$$\begin{aligned} &= \sum_{i=1}^N P(Y_{t+2}, \dots, Y_T, Y_{t+1} = v_l, X_{t+1} = i | X_t = j, Y_t = v_k, \lambda) \\ &= \sum_{i=1}^N P(X_{t+1} = i, Y_{t+1} = v_l | X_t = j, Y_t = v_k) \\ &\quad \cdot P(Y_{t+2}, \dots, Y_T | X_{t+1} = i, Y_{t+1} = v_l, X_t = j, Y_t = v_k) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N \psi_{ji}(v_k, v_l) P(Y_{t+2}, \dots, Y_T | X_{t+1} = i, Y_{t+1} = v_l) \\
&= \sum_{i=1}^N \psi_{ji}(v_k, v_l) \beta_i(t+1) \\
\Rightarrow \beta_j(t) &= \begin{cases} \sum_{i=1}^N \psi_{ji}(v_k, v_l) \beta_i(t+1) & , t = 1, 2, \dots, T-1 \\ 1 & , t = T \end{cases}, \quad (4.15)
\end{aligned}$$

where $i, j \in \Omega$, $v_k, v_l \in \mathcal{O}$, and $k = 1, 2, \dots, M$.

The computational space and time complexities for forward and backward probabilities are the same as HMM for the proposed model. The space complexity of BDHMM is $\mathcal{O}(NT)$, because the number of N forward or backward probabilities are calculated for a length of $t = 1, 2, \dots, T$. In addition, the number of terms that each forward or backward probability includes as many as the number of hidden states. So, the time complexity of the algorithm for the evaluation problem is $\mathcal{O}(N^2T)$.

The posterior marginal probabilities of hidden states $\{X_1, X_2, \dots, X_T\}$ at any time, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$, is calculated by the same formula as in HMM. That is,

$$\begin{aligned}
P(X_t = j | Y_1, Y_2, \dots, Y_T) &\underset{\tilde{X}_t}{\propto} P(X_t = j, Y_1, Y_2, \dots, Y_T) \\
&= \beta_j(t) \alpha_j(t).
\end{aligned} \quad (4.16)$$

4.2.2 Learning Problem

The aim of the learning procedure is to find the most likely set of state transition and observation emission probabilities, given a set of output sequences and general structure of BDHMM (number of hidden states). Forward and backward probabilities

are defined as the same in classical HMM for BDHMM, however the computation is different. Hence, the calculation of the updated estimates in maximization step of Baum-Welch algorithm makes the difference. It is known that Baum-Welch algorithm (Baum et al., 1970) is used to estimates the complete parameter set $\lambda = (P, B, \pi, \phi)$ of BDHMM for the solution of learning problem.

Initial values of the complete parameter set must be defined. For the Baum-Welch algorithm some definitions related to BDHMM will be given.

The general idea for calculating the new estimates of the state-observation probability distribution $\psi_{ij}(v_k, v_l)$ is given below,

$$\psi_{ij}(Y_{t-1} = v_k, Y_t = v_l) = \frac{\text{expected number of transitions from state } i \text{ to } j \\ \text{with outputs } v_k \text{ and } v_l, \text{ respectively}}{\text{expected \# of transitions out of state } i \text{ with output } v_k}.$$

The general idea for calculating the new estimates of the initial state probabilities π_j is given below,

π_j = expected frequency in state j at time ($t = 1$).

Let the random variable $\xi_{ij}(t)$ be the probability of being in state i at time t and in state j at time $t+1$, given the observation sequence $\{Y_1, Y_2, \dots, Y_T\}$. And, let $L(\lambda)$ be the likelihood of the model which is calculated as given in (4.13).

$$\begin{aligned} \xi_{ij}(t) &= P(X_t = i, X_{t+1} = j | Y_1 Y_2 \dots Y_T) \\ &= \frac{P(X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \\ &= \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1}) \cdot P(X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1})}{L(\lambda)} \end{aligned} \tag{4.17}$$

$$\begin{aligned}
& \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_t = i, X_{t+1} = j, Y_1, Y_2, \dots, Y_{t+1})}{L(\lambda)} \\
& \cdot P(X_{t+1} = j, Y_{t+1} = v_l | X_t = i, Y_1, Y_2, \dots, Y_t = v_k) \\
& \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i) \\
& = \frac{P(Y_{t+2}, Y_{t+3}, \dots, Y_T | X_{t+1} = j, Y_{t+1})}{L(\lambda)} \\
& \cdot P(X_{t+1} = j, Y_{t+1} = v_l | X_t = i, Y_t = v_k) \\
& \cdot P(Y_1, Y_2, \dots, Y_t, X_t = i) \\
& = \frac{\alpha_i(t) \psi_{ij}(v_k, v_l) \beta_j(t+1)}{L(\lambda)}
\end{aligned}$$

Note that the probabilities $\xi_{ij}(t)$ for each step are greater than or equal to zero and sum of the probabilities $\xi_{ij}(t)$ for all states at time t is equal to one. That is,

$$\xi_{ij}(t) \geq 0, \quad i, j \in \Omega; \quad (4.18)$$

$$\sum_{i=1}^N \sum_{j=1}^N \xi_{ij}(t) = 1, \quad (4.19)$$

where, $t = 1, 2, \dots, T - 1$.

Let the random variable $\gamma_j(t)$ defined as the probability of being in state j at time t , given the observation sequence $\{Y_1 Y_2 \dots Y_T\}$.

$$\gamma_j(t) = P(X_t = j | Y_1, Y_2, \dots, Y_T) \quad (4.20)$$

$$\begin{aligned}
& = \frac{P(X_t = j, Y_1, Y_2, \dots, Y_T)}{P(Y_1, Y_2, \dots, Y_T)} \\
& = \frac{P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_1, Y_2, \dots, Y_t)}{L(\lambda)} \\
& \cdot P(X_t = j, Y_1, Y_2, \dots, Y_t) \\
& = \frac{P(Y_{t+1}, Y_{t+2}, \dots, Y_T | X_t = j, Y_t)}{L(\lambda)} \\
& \cdot P(X_t = j, Y_1, Y_2, \dots, Y_t)
\end{aligned}$$

$$= \frac{\alpha_j(t)\beta_j(t)}{L(\lambda)}$$

Note that sum of the probabilities $\gamma_j(t)$ for all states at time t is equal to one and they are greater than or equal to zero. That is,

$$\gamma_j(t) \geq 0, \quad j \in \Omega; \quad (4.21)$$

$$\sum_{j=1}^N \gamma_j(t) = 1, \quad (4.22)$$

where $j \in \Omega$, $t = 1, 2, \dots, T$.

Now, let us write the expected values defined in the general probability ideas in terms of probabilities $\xi_{ij}(t)$ and $\gamma_j(t)$.

Let $E(N_{i \rightarrow j}^{v_k \rightarrow v_l})$ be the expected number of times observation v_k occurs in state i and observation v_l occurs in state j after a transition. It can be calculated as follows,

$$E(N_{i \rightarrow j}^{v_k \rightarrow v_l}) = \sum_{\substack{t=1 \\ Y_{t+1}=v_l \\ Y_t=v_k}}^{T-1} \xi_{ij}(t). \quad (4.23)$$

Let $E(N_{i \rightarrow}^{v_k \rightarrow})$ be the expected number of transitions out of state i and emitting observation v_k . It can be calculated as follows,

$$E(N_{i \rightarrow}^{v_k \rightarrow}) = \sum_{\substack{t=1 \\ Y_{t-1}=v_k}}^T \gamma_i(t). \quad (4.24)$$

And finally, let $E(N_j(1))$ be the expected frequency in state j at time $t = 1$. It can be calculated as follows,

$$E(N_j(1)) = \gamma_j(1). \quad (4.25)$$

In the second step, called the maximization-step, the updated estimated values of the parameters are obtained by using the expected values. The estimated value of the state-observation probability distribution for the n th iteration is

$$\psi_{ij}^{(n)}(Y_{t-1} = v_k, Y_t = v_l) = \frac{E(N_{i \rightarrow j}^{v_k \rightarrow v_l})}{E(N_{i \rightarrow}^{v_k \rightarrow})} = \frac{\sum_{\substack{t=1 \\ Y_{t+1}=v_l \\ Y_t=v_k}}^{T-1} \xi_{ij}(t)}{\sum_{\substack{t=1 \\ Y_{t-1}=v_k}}^T \gamma_i(t)}. \quad (4.26)$$

The estimated value of the initial state distribution for the n th iteration is

$$\pi_j^{(n)} = E(N_j(1)) = \gamma_j(1). \quad (4.27)$$

For the proposed BDHMM, the general idea of initial observation distribution is

$$\varphi_j(v_k) = \frac{\text{limiting distribution of emitting } v_k \text{ at state } j}{\text{limiting distribution of emitting } v_k \text{ at any state}}.$$

To calculate the probabilities $\varphi_j(v_k)$, Ψ matrix is used. $\varphi_j(v_k)$ gives the probability distribution of initial observation emissions in the long term. Let $\mathbf{W}$ be an (NxM) dimensional row vector which gives the limiting probabilities of Ψ . The entry w_{j,v_k} represents the probability of being at state j and emitting v_k in the long term. That is,

$$\mathbf{W} = [w_{1,v_1} \dots w_{N,v_1} \dots w_{1,v_2} \dots w_{N,v_2} \dots \dots w_{1,v_M} \dots w_{N,v_M}]. \quad (4.28)$$

So, the maximization step of the initial observation distribution for the n th iteration is calculated by

$$\varphi_j^{(n)}(v_k) = \frac{w_{j,v_k}}{\sum_{i=1}^N w_{i,v_k}}. \quad (4.29)$$

The iterations are continued until the absolute difference between the parameter estimates obtained two successive iteration; or the likelihood values obtained at $(n-1)$ th and n th iteration falls below a predefined value ϵ .

4.2.3 Decoding Problem

The aim of the decoding procedure is to estimate the most probable state transition sequence, given the observation sequence. In other words, given the observation sequence $\{Y_0, Y_1, \dots, Y_T\}$, and the model parameters $\lambda = (\Psi, \pi, \phi)$, how do we choose a corresponding state sequence path which is optimal in some meaningful sense (i.e., best “explains” the observations)? The most common and effective way to solve this problem is the Viterbi algorithm (Viterbi, 1967).

Let S be hidden state sequence and S^* be the “best” possible hidden state sequence. Then the goal is to maximize $P(S^*|Y_0, Y_1, \dots, Y_T)$ by choosing the “best” S . That is,

$$S^* = \underset{S}{argmax} [P(S|Y_0, Y_1, \dots, Y_T)]. \quad (4.30)$$

The following equation can be achieved by using the conditional probability and the chain rule.

$$\begin{aligned} P(S|Y_0, Y_1, \dots, Y_T) &= \frac{P(S, Y_0, Y_1, \dots, Y_T)}{P(Y_0, Y_1, \dots, Y_T)} \\ &= \frac{P(S, Y_0, Y_1, \dots, Y_T)}{L(\lambda)} \end{aligned} \quad (4.31)$$

$P(Y_0, Y_1, \dots, Y_T)$ is the likelihood $L(\lambda)$ of the model and stay constant over time, so $P(S|Y_0, Y_1, \dots, Y_T)$ is proportional to the joint probability distribution $P(S, Y_0, Y_1, \dots, Y_T)$. That is,

$$P(S|Y_0, Y_1, \dots, Y_T) \propto_S P(S, Y_0, Y_1, \dots, Y_T). \quad (4.32)$$

So, the “best” possible hidden state sequence (S^*) can be estimated as follows,

$$S^* = \underset{S}{argmax} [P(S, Y_0, Y_1, \dots, Y_T)]. \quad (4.33)$$

The joint probability of the state sequence (S) and observation sequence $\{Y_0, Y_1, \dots, Y_T\}$ can be calculated as follows,

$$\begin{aligned} P(S, Y_0, Y_1, \dots, Y_T) &= P(X_0, X_1, \dots, X_T, Y_0, Y_1, \dots, Y_T) \\ &= P(X_0)P(Y_0|X_0)P(X_1, Y_1|X_0, Y_0) \dots \\ &\quad \cdot P(X_T, Y_T|X_{T-1}, \dots, X_0, Y_{T-1}, \dots, Y_0) \\ &= P(X_0)P(Y_0|X_0)P(X_1, Y_1|X_0, Y_0) \dots \\ &\quad \cdot P(X_T, Y_T|X_{T-1}, Y_{T-1}) \\ &= P(X_0 = r)P(Y_0 = v_m|X_0 = r) \\ &\quad \cdot \prod_{t=1}^T P(X_t = j, Y_t = v_l|X_{t-1} = i, Y_{t-1} = v_k) \\ &= \pi_r \varphi_r(v_m) \prod_{i=1}^T \psi_{i,j}(v_k, v_l), \end{aligned} \quad (4.34)$$

where $i, j, r \in \Omega$, $v_k, v_l, v_m \in \mathcal{O}$, and $k, l, m = 1, 2, \dots, M$.

The estimate of the hidden state with the highest likelihood is calculated at the end of each step based on the observation symbol which is emitted from the related hidden

state. A pattern must exist between the estimated values of hidden state sequence and the observation sequence, and they must be also consistent for each step as the model fit with the observation sequence increases.

4.3 Experimental Study

Within the scope of experimental work, we examine whether proposed BDHMM will produce better results when there is a binary dependency between successive steps of the process. For this purpose, the first group of observation sequences is generated by using an arbitrary initial state-observation distribution (Ψ). Next, these observation sequences are modeled with HMM and BDHMM, parameters are estimated and the results are compared. It is expected that BDHMM provides better estimations for this framework. Furthermore, the second group includes observation sequences that are generated from classical HMM with some arbitrary initial parameters where the observations are conditionally independent. Each of the conditionally independent observation sequences is also modeled with both HMM and BDHMM. Then the results are compared for the second group using the parameter estimates. HMM is expected to give better results for this second context. Both groups of observation sequences include four different models such as 2-state-2-observation, 2-state-3-observation, 3-state-2-observation, 3-state-3-observation. And also, both data set include 1000 observation sequences each with 500 observations. The types of observation sequences used in the experimental study are given in Table 4.2.

Table 4.2 Types of observation sequences

Dependency Structure	Number of Hidden States (N)	Number of Observations (M)
HMM	2	2
	2	3
	3	2
	3	3
BDHMM	2	2
	2	3
	3	2
	3	3

The state-observation distribution and the observation emission distribution are selected arbitrarily according to the overall behavior of the observation sequences and also parameters for each observation sequence are estimated by the Baum-Welch algorithm for both HMM and BDHMM. The algorithm stops for all parameters of the models when the absolute differences between the current and the previous estimates are less than $\epsilon = 10^{-4}$. Moreover, likelihood values ($\hat{L}$) for each model are calculated. The performance of the models is evaluated and model selection is done according to the AIC the BIC values. AIC and BIC values indicate the amount of information lost during modeling, hence small AIC and BIC values imply the best model fit.

As a result of modeling the observation sequences generated based on BDHMM assumptions by using both HMM and BDHMM, AIC and BIC values are calculated. The mean and standard deviation values of AIC and BIC values are given in the first and the second part of Table 4.3, respectively. Four types of models are tested with different numbers of states (N) and observation symbols (M).

Table 4.3 Results obtained by modeling the observation sequences which are generated based on BDHMM assumptions

Model		# of freely estimated parameters		Mean and std.dev. of AIC		Mean and std.dev. of BIC		# of success for BDHMM		Mean relative increment of BDHMM (%)	
N	M	HMM	BDHMM	HMM	BDHMM	HMM	BDHMM	AIC	BIC	AIC	BIC
2	2	5	6	701.6427 (5.4206)	695.8790 (4.0071)	773.7671 (5.4206)	766.8835 (4.0071)	967	911	0.8214	0.8896
2	3	7	9	1087.20 (10.3809)	1078.55 (8.7582)	1155.01 (10.3809)	1143.79 (8.7582)	924	905	0.7954	0.9714
3	2	11	13	701.5756 (8.0513)	686.1305 (4.0840)	824.2971 (8.0513)	807.8568 (4.0840)	979	938	2.2014	1.9945
3	3	14	16	1097.58 (9.6841)	1051.77 (4.6737)	1185.01 (9.6841)	1171.68 (4.6737)	1000	864	4.1737	1.1254

The number of freely estimated parameters changes according to the number of states and observation symbols, for HMM and BDHMM. The number of freely estimated parameters of BDHMM are higher than that of HMM. From Table 4.3 we can conclude that; when observation sequences having a binary dependency, the mean AIC and BIC values obtained by using BDHMM are lower than those of HMM. In addition, the standard deviations of both AIC and BIC values are obtained with BDHMM smaller than those of HMM. Hence, according to the AIC values, success is achieved in all 1000 samples for 3-state-3-observation model and the number of success for BDHMM is 979 for 3-state-2-observation model. When the number of hidden state is 2, the number of success is 967 for 2-state-2-observation-model and 924 for 2-state-3-observation-model. According to the BIC values, the number of success is 911 for 2-state-2-observation model, 905 for 2-state-3-observation model, 938 for 3-state-2-observation model and 864 for 3-state-3-observation model. From the last two columns of Table 4.3, the mean relative change of BDHMM for both AIC and BIC is positive for four model types and percentage increases between 0.7954 and 4.1737 are observed. The positive relative changes indicate that BDHMM presents better for this case. When the observation sequences in the first group are modeled with both HMM and BDHMM and their AIC and BIC values are compared, it is clearly seen that the number of success is relatively high. This means BDHMM represents

best for most of the observation sequences that have a binary dependency, as expected. Moreover, even the lowest number of success is 905, which means that at least a 90.5% success rate is achieved for each four model types.

By modeling the second group of observation sequences which are generated based on HMM assumptions using both HMM and BDHMM, the calculated mean and standard deviation values of AIC and BIC are presented in Table 4.4.

Table 4.4 Results obtained by modeling the observation sequences which are generated based on HMM assumptions

Model		# of freely estimated parameters		Mean and std.dev. of AIC		Mean and std.dev. of BIC		# of success for BDHMM		Mean relative increment of BDHMM (%)	
		HMM	BDHMM	HMM	BDHMM	N	M	HMM	BDHMM	HMM	BDHMM
2	2	5	6	684.6077 (8.0037)	690.7743 (8.8882)	749.1830 (8.0037)	754.5359 (8.8882)	78	105	-0.9007	-0.7144
2	3	7	9	1086.60 (8.9413)	1098.84 (10.7899)	1129.26 (8.9413)	1146.47 (10.7899)	45	72	-1.1261	-1.5241
3	2	11	13	701.9994 (3.9779)	711.8926 (7.8528)	801.4781 (3.9779)	809.8539 (7.8528)	68	111	-1.4092	-1.0450
3	3	14	16	1095.33 (4.8848)	1105.13 (9.5587)	1159.90 (4.8848)	1168.20 (9.5587)	122	163	-0.8945	-0.7151

From Table 4.4 it can be seen that the mean AIC and BIC values obtained by using HMM are lower than those of BDHMM, for each of the four model types. In addition, the deviations from true estimation values are lower in HMM modeling and the estimation values are more consistent for 1000 observation sequences. Hence, according to the AIC values, the number of success is just 78 for 2-state-2-observation model, 45 for 2-state-3-observation model, 68 for 3-state-2-observation model and 122 for 3-state-3-observation model. According to the BIC values, the number of success is only 105 for 2-state-2-observation model, 72 for 2-state-3-observation model, 111 for 3-state-2-observation model and 163 for 3-state-3-observation model. And also the mean relative change of BDHMM for both AIC and BIC is negative in four model types and percentage decreases between 0.7144 and 1.5241 are observed. When the

AIC and BIC values compared as a result of the modeling of the observation sequences in the second group, it is observed that the number of successes is relatively low for BDHMM for each of four model types. It means that HMM fits better with the observation sequences that have conditionally independent observations.

The Akaike and the Bayesian weights are also calculated for the generated observation sequences as in section 3.3. The sum of Akaike and Bayesian weights (weight of evidence) are equal to 1 for all models compared. These weights are used to interpret how much a model represents data relative to the other. The mean values $\bar{\Delta}$ and $\bar{w}$ for HMM and BDHMM which are calculated from the 1000 observation sequences are given in Table 4.5 and Table 4.6, respectively.

Table 4.5 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of BDHMM

Number of Parameter		Observation sequences generated based on the assumptions of BDHMM					
N	M	$\bar{\Delta}^{AIC}$	$\bar{\Delta}^{BIC}$	$\bar{w}_{BDHMM}^{AIC}$	$\bar{w}_{HMM}^{AIC}$	$\bar{w}_{BDHMM}^{BIC}$	$\bar{w}_{HMM}^{BIC}$
2	2	5.7637	6.8836	0.9469	0.0531	0.9689	0.0311
2	3	8.6475	11.2206	0.9869	0.0131	0.9963	0.0037
3	2	15.4451	16.4403	0.9996	0.0004	0.9997	0.0003
3	3	45.8099	13.3361	≈ 1	≈ 0	0.9987	0.0013

It is clearly seen from Table 4.5 that, for the group of first 1000 observation sequences which are generated by an arbitrary initial state-observation distribution (Ψ), the mean increase in HMM's AIC values vary between 5,7637 and 45,8099. The mean increase in HMM's BIC values varies between 6,8836 and 16,4403. According to both AIC and BIC values, we have less evidence to support that data fits the 2-state-2-observation HMM. Moreover, we have almost no evidence to support that data fits 2-state-3-observation HMM according to AIC and no evidence according to BIC values. Furthermore, there is no evidence that 3-state-2-observation and 3-state-3-observation HMMs represent data, according to both AIC and BIC values. Hence, we can conclude that modeling the observation sequences by HMM gives poor results relative to BDHMM.

Table 4.6 $\bar{\Delta}$ and $\bar{w}$ values for observation sequences generated based on the assumptions of HMM

Number of Parameter		Observation sequences generated based on the assumptions of HMM					
N	M	$\bar{\Delta}^{AIC}$	$\bar{\Delta}^{BIC}$	$\bar{w}_{BDHMM}^{AIC}$	$\bar{w}_{HMM}^{AIC}$	$\bar{w}_{BDHMM}^{BIC}$	$\bar{w}_{HMM}^{BIC}$
2	2	6.1666	5.3529	0.0438	0.9562	0.0643	0.9357
2	3	12.2371	17.2106	0.0022	0.9978	0.0002	0.9998
3	2	9.8932	8.3758	0.0071	0.9929	0.0149	0.9851
3	3	9.7982	8.2945	0.0074	0.9926	0.0156	0.9844

For the group of second 1000 observation sequences which are generated by classical HMM (conditionally independent observations), the mean increase in BDHMM's AIC values vary between 6,1666 and 12,2371. The mean increase in BDHMM's BIC values varies between 5,3529 and 17,2106. We have less evidence to support 2-state-2-observation BDHMM according to both AIC and BIC values. And also, there is no evidence to support for 2-state-3-observation BDHMM according to both AIC and BIC values. The same conclusion can be made for 3-state-2-observation and 3-state-3-observation BDHMMs according to both AIC and BIC values.

In addition, for the first group of observation sequences, it is observed that the mean Akaike weights for BDHMM vary between 0.9469 and ≈ 1 and the mean Bayesian weights for BDHMM vary between 0,9689 and 0,9997. For the second group of observation sequences, it is observed that the mean Akaike weights for BDHMM vary between 0,0022 and 0,0438, and the mean Bayesian weights for BDHMM vary between 0,0002 and 0,0643. These weights can be expressed as the expected number of the best representative models and given in Table 4.7. Values in parentheses are the number of successes for BDHMM obtained in the previous analysis.

Table 4.7 Comparison of the number of successes and the expected number of more likely model for BDHMM

Number of Parameter		Observation sequences generated based on the assumptions of BDHMM		Observation sequences generated based on the assumptions of HMM	
N	M	AIC	BIC	AIC	BIC
2	2	≈ 947 (967)	≈ 969 (911)	≈ 44 (78)	≈ 64 (105)
2	3	≈ 987 (924)	≈ 996 (905)	≈ 2 (45)	≈ 0 (72)
3	2	≈ 1000 (979)	≈ 1000 (938)	≈ 7 (68)	≈ 15 (111)
3	3	≈ 1000 (1000)	≈ 999 (864)	≈ 7 (122)	≈ 16 (163)

The number of success achieved is only based on the lower AIC and BIC values, decrements in AIC and BIC values are not considered. However, by computing the Akaike and Bayesian weights, significance of the decrement in AIC and BIC values are also taken into account. According to Table 4.7, we have even stronger support of substantial evidence to say that, BDHMM models the first group of BDHMM generate observation sequences much better than HMM, and HMM models the second group of HMM generate observation sequences much better than HMM. The reason is that, for the first group of observations and for almost all but one of the expected numbers of the best representative models, the AIC and BIC values are higher than the number of successes for BDHMM, and closer to 1000. For the second group of observations and for all the expected numbers of the best representative models, the AIC and BIC values are lower than the number of successes for BDHMM and closer to 0. To sum up, when the differences of AIC and BIC are taken into consideration, BDHMM performs much better for the first group of observations and HMM performs much better for the second group of observations.

From these comparisons, we may draw a conclusion that; the proposed BDHMM gives better results than the classical HMM when there exists a binary dependency between the consecutive steps of the process. However, the proposed model is insufficient compared to HMM for conditionally independent observations.

4.4 An Application for the Strongest Earthquake Occurrences

One of the common implementations of HMM is modeling and predicting earthquake occurrences. As a known fact, earthquake occurrence is associated with the structural changes in the crust of the Earth. Earthquake magnitudes or occurrence times are affected by various factors such as accumulated stress on the fault, or depth of the earthquake. Usually, the dynamics of the crust cannot be observable, hence hidden Markov models are appropriate to model historical earthquakes and make predictions.

Analyzing seismic data and developing statistical methods for earthquake occurrences have been a challenging subject matter for decades. The elastic rebound theory, by Reid (1910) is one of the classical models for earthquake mechanisms. It states that stress in a region accumulated due to the movement of tectonic plates, and is released when it exceeds some value of strength. Vere-Jones (1978) proposed a stochastic version of the elastic rebound theory called *stress release model*. For a broad review of seismicity-based earthquake forecasting techniques, readers may refer Tiampo & Shcherbakov (2012).

There exist many probabilistic models that characterize earthquakes. Since the Poisson process is a relatively simple model, and also it has fewer parameters to be estimated, it is the most common model. An extensive body of literature for earlier models are given by Anagnos & Kiremidjan (1988). Though it is used widely, the hazard rate of the Poisson model is constant, which means in a small increment of time, the earthquake probability is constant. On the other hand, elastic rebound theory states that neither the times nor the magnitudes of earthquakes may not be statistically independent. Markov and semi-Markov models allow describing the dependence in the sequence of the earthquakes. In Markov processes, exponential sojourn times are assumed; on the other hand, with semi-Markov models, it is possible to assume different sojourn time distributions such as Weibull or Gamma, to assure increasing hazard rates. Semi-Markov process has been used to model the earthquake occurrences and used for earthquake forecasting by Altinok & Kolcak (1999). Alvarez (2005) proposed a semi-Markov model where a mixture of exponential and Weibull

distribution is assumed for waiting times. Votsi, Limnios, Tsaklidis, & Papadimitriou (2012) provided a semi-Markov model in which seismotectonic features have been incorporated as a component. Sadeghian (2012) described a semi-Markov model to predict the probability of the time and place of occurrence of earthquakes. Another study that assumed sojourn times as Weibull distribution is Masala (2012). Masala (2012) proposed both homogeneous and non-homogeneous semi-Markov model and he allowed dependence in both space and time, in their model.

The first application of HMM in seismology is Granat & Donellan (2002). They claimed that HMM forms the basis of one type of unsupervised learning clustering method that is particularly applicable to geophysical systems because it includes in the model the time relationship between the states of the system. They used HMM for clustering purpose and proposed a novel method called *Deterministic Annealing Hidden Markov Model* (DAHMM), and applied this method to GPS seismicity data. According to the results, their method is able to separate the data into classes that correspond to the events observed. Chambers, Ebel, Kafka, & Baglivo (2003) stated that the Poisson process is insufficient for the prediction of earthquakes and applied two-state and three-state HMM to represent the earthquakes in New England, and provided estimates for inter-event times. Data set consists of 110 inter-event times (in days) between 1975 and 2000 and earthquakes with magnitude $M \geq 2.7$. The hidden states are taken as rapid response and slow response to a quake. Ebel, Chambers, Kafka, & Baglivo (2007) proposed an earthquake clustering and an HMM for the earthquakes in California and gave both short term and long-term predictions. Inter-event times between earthquakes are assumed to be exponential. Orfanogiannaki, Karlis, & Papadopoulos (2010) introduced Poisson HMM (PHMM) to model the seismic variations and stated that PHMM can describe short-term earthquake occurrences in the time domain. Data set consists of the earthquakes larger than magnitude $M > 3.2$ around the Ionian Sea, Greece, between 1990 and 2006. Observations are the daily number of earthquakes and inter-event times. An appropriate model selection is done among different models and the number of hidden states is determined. Chambers, Baglivo, Ebel, & Kafka (2012) gave estimates for short-term earthquake occurrences using HMM. They also presented a model that

includes location information. Votsi, Limnios, Tsaklidis, & Papadimitriou (2013) examined various HMM combinations with a different number of states and different numbers of observations and used HMM to reveal the actual states. Besides expected occurrence times until the next earthquake are estimated. Yip, Ng, & Yau (2018) proposed an HMM to predict both the time and the magnitude of an earthquake simultaneously. They assume that when the stress level exceeds a threshold, an earthquake occurs and there emits some energy. Hence, they examined the underground stress level in two states such as low level of earthquake frequency and high level of earthquake frequency. The model includes earthquake magnitudes as observations and transition probability matrix varying with time. Observations are the earthquake amplitudes, inter-event times and earthquake frequencies.

There are some studies that model the earthquake occurrences by hidden semi-Markov process. A comprehensive implementation of hidden semi-Markov models for the seismic hazard assessment was presented by Pertsinidou, Tsaklidis, Papadimitriou, & Limnios (2017). The earthquakes which have a magnitude $M \geq 2.7$ around the North and South Aegean Sea were taken into consideration. Hidden states correspond to the stress field and, five different possible observations, including location information, can occur at the end of each transition. Models with different dimensions were developed where the sojourn times between earthquakes are Poisson, Logarithmic, and Negative Binomial. A detailed information for the researchers, who are interested in seismology from the statistical point of view, about discrete HMM and hidden Markov renewal model is presented by Votsi, Limnios, Papadimitriou, & Tsaklidis (2019).

To summarize, for many existing studies that model the earthquake occurrence by HMM, the states of the hidden Markov chain are usually defined as the unobservable state of the crust. And also, the observation symbol which conditionally depends on the current hidden state is usually defined as earthquake frequency, earthquake magnitude or elapsed time between two successive earthquakes. What all these existing studies have in common is that the observed random variables are conditionally independent and identically distributed. That is in the physical meaning

if we take the observation random variable as the magnitudes, the magnitude of an earthquake is assumed independent from the previous earthquake magnitude. However, according to the seismicity theory, there is a dependence between the earthquake magnitudes. It is very likely that the magnitude of an earthquake and its related stress level, which corresponds to the hidden states, are affected by those of the previous earthquake. In other words, the idea that the current observation and hidden state may depend on the previous observation and hidden state is the motivation of this section and the aim is to model the earthquake occurrences by using more information from the past and make more accurate predictions. If a Markov dependence exists between the magnitude of an earthquake and its related stress field, and the magnitude of the previous earthquake and its related stress field, it will be possible to reveal this dependency structure using BDHMM rather than HMM, since HMM does not take the binary dependency into consideration.

Since the behavior of the fault system and the accumulated stress on the crust cannot be directly observable, HMMs are suitable for modeling earthquake occurrences. Stress level corresponds to the hidden states, whereas the magnitudes correspond to the observations. When HMM is used for modeling, it is assumed that successive earthquake magnitudes are conditionally independent, and the system makes transitions, independent from the magnitudes according to a Markov chain, in which the states are stress level.

According to the stress release model (Vere-Jones, 1978), the stress level in a region increases between earthquakes and is reduced stochastically as a result of an earthquake. According to this theory, in a certain region tectonic stress accumulates along time on a fault system, and some part of the accumulated stress releases with an earthquake with magnitude M . Furthermore, since the energy releases with an earthquake at time $(t - 1)$, the amount of energy released is an important factor that determines the amount of energy that will be released in the next occurrence. Hence, the magnitude of the next earthquake depends not only on the fault block in that region but also on the previous earthquake magnitude.

Therefore, we assume that the binary information of the current stress level and the related earthquake magnitude, conditionally depends on the binary information of the previous stress level and the related earthquake magnitude. If there is indeed such a binary dependence, BDHMM will be more appropriate in modeling strong earthquake occurrences in the region of interest, relative to HMM. Hence, the proposed model is investigated on a real-life problem about the occurrences of the strongest earthquakes.

Earthquake magnitude data which is used by Votsi et al. (2013) is used in this section. It is available in the earthquake catalog of the Geophysics Department of the Aristotle University of Thessaloniki (<http://geophysics.geo.auth.gr/ss/>). The data set consists of 116 strong earthquakes with magnitudes $M > 6.5$ between the time interval of 1845-2010. The detailed information about the geological features of the study region and the complete data of earthquake magnitudes are given by Votsi et al. (2013).

Votsi et al. (2013) present two approaches to model the earthquake data. In their second approach, the earthquakes that occurred between 1845-2010 are considered and years are taken as a time unit. This approach is based on recording the strongest earthquake that occurred each year. The earthquake magnitudes are classified into three groups as a result of k-means clustering. If an earthquake with $M \geq 6.5$ has not occurred during any year between 1845-2010, it is recorded an observation symbol of type 1. If the magnitude of the strongest earthquake observed in any given year is between $M \in [6.5, 6.9]$, it is recorded as an observation symbol of type 2, and if stronger than $M \geq 7.0$, it is recorded as an observation symbol of type 3. In this way, an observation sequence with a length of 165 is obtained from 1845 to 2010.

It is possible to reveal whether there is a Markov dependence between the combination of the strongest earthquake and the stress level in that region in the current year and this combination in the previous year by using proposed BDHMM.

4.4.1 Model Selection

2-state-3 observation and 3-state-3-observation models are constructed both with classical HMM and BDHMM. Let *state 1* be defined as the *low-stress level* and *state 2* be defined as the *high-stress level* for the two-state models. And, let *state 1* be defined as the *low-stress level*, *state 2* as the *average-stress level*, and *state 3* as the *high-stress level* for the three-state models. Earthquake magnitudes which correspond to the observations are classified into three groups. If no earthquake is observed with magnitude $M \geq 6.5$ in any given year, it is recorded as *obs. 1*; if an earthquake is observed with magnitude between $M \in [6.5, 6.9]$, in a given year, it is recorded as *obs. 2*; and finally, the earthquakes with $M \geq 7.0$ are recorded as *obs. 3*. The models which are compared are given in the Table 4.8.

Table 4.8 Types of observation sequences

Dependency Structure	Number of Hidden States (N)	Number of Observations (M)
HMM	2	3
	2	3
MEHMM	3	3
	3	3

By changing the initial values each time and selecting arbitrarily, model parameters of HMM and BDHMM are estimated many times by the help of Baum-Welch and modified Baum-Welch algorithm, respectively. The algorithm stops for all parameters of the models when the absolute differences between the current and the previous estimate values are less than $\epsilon = 10^{-4}$. The Baum-Welch algorithm converges to the optimum values of parameter estimates in 28 iterations for the 2-state-3-observation and in 233 iterations for the 3-state-3-observation HMMs. The modified Baum-Welch algorithm converges to the optimum values of parameter estimates in 154 iterations for the 2-state-3-observation and in 182 iterations for the 3-state-3-observation BDHMMs. Finally, the estimations of complete parameter sets that maximize the likelihood value for each model are adopted. The log-likelihood values of the

aforementioned four models are compared for selection of the most representative model for the earthquake occurrences. Besides, AIC and BIC values are obtained to decide the most likely model. A comparison for the models by Log-likelihood, AIC, and BIC values and weights are presented in Table 4.9. P is the number of freely estimated elements of parameters of the corresponding model.

Table 4.9 Comparison of the models

Model Type	# of States	# of Obs.	P	$\ln L$	AIC (Rank)	w_i^{AIC}	BIC (Rank)	w_i^{BIC}
HMM	2	3	7	-157.4325	328.8650 (3)	≈ 0	350.6066 (2)	0.0001
BDHMM	2	3	9	-143.0939	304.1878 (1)	0.9886	332.1413 (1)	0.9999
HMM	3	3	14	-154.7332	337.4664 (4)	≈ 0	380.9496 (4)	≈ 0
BDHMM	3	3	16	-140.5534	313.1068 (2)	0.0114	362.8019 (3)	≈ 0

From AIC and BIC values of four models, given in Table 4.9, it is clear that BDHMM gives better results than HMM for both 2-state-3-observation and 3-state-3-observation models. The mean relative increase in AIC value is 7.5%, in BIC value is 5.3% for the 2-state-3-observation model. Besides, the mean relative increase in AIC value is 7.2% and in BIC is 4.8% for the 3-state-3-observation model. The smallest AIC and BIC values are obtained for the 2-state-3-observation BDHMM. This means that BDHMM with two hidden states characterizes the observation sequence, better. Furthermore, it can be concluded that the earthquake with maximum magnitude in the current year and its related stress level has a first-order Markov dependence on those of the previous year. In addition, BDHMMs take the best two places according to the AIC values. This means that even the 2-state-3-observation HMM, which can be considered better HMM, does not fit better than the 3-state-3-observation BDHMM, which is the worse BDHMM. On the other hand, the second-best model is the 2-state-3-observation HMM according to the BIC values. According to the Akaike and Bayesian weights, the 2-state-3-observation BDHMM is obviously noticeable as the best representative model. According to Akaike weights 9886 times out of 10000, and according to Bayesian weights 9999 times out of 10000, the best representative model will more likely be BDHMM with 2-state-3-observation.

Let O represent “observation”, and S represent “state”. Then the joint event that emits an observation symbol j in state i is defined as " $S_i O_j$ ". The final estimates of the complete parameter set of 2-state-3-observation BDHMM is given as follows,

$$\hat{\Psi} = \begin{matrix} & \begin{matrix} S_1 O_1 & S_2 O_1 & S_1 O_2 & S_2 O_2 & S_1 O_3 & S_2 O_3 \end{matrix} \\ \begin{matrix} S_1 O_1 \\ S_2 O_1 \\ S_1 O_2 \\ S_2 O_2 \\ S_1 O_3 \\ S_2 O_3 \end{matrix} & \left[\begin{array}{cc|cc|cc} \hline 0.5838 & 0.0134 & 0.1942 & 0.0348 & 0.1738 & 0 \\ \hline 0.0997 & 0.5122 & 0.2872 & 0.1009 & 0 & 0 \\ \hline 0.6350 & 0 & 0 & 0 & 0 & 0.3650 \\ \hline 0 & 0.0109 & 0.6883 & 0.3005 & 0 & 0.0003 \\ \hline 0 & 0.5000 & 0 & 0.1000 & 0 & 0.4000 \\ \hline 0 & 0.4348 & 0.2609 & 0 & 0 & 0.3043 \\ \hline \end{array} \right] \end{matrix},$$

$$\hat{\pi} = \begin{matrix} & \begin{matrix} \text{State 1} & \text{State 2} \end{matrix} \\ \left[\begin{array}{cc} 0 & 1 \end{array} \right], \end{matrix}$$

$$\hat{\phi} = \begin{matrix} & \begin{matrix} \text{State 1} & \text{State 2} \end{matrix} \\ \begin{matrix} \text{Obs. 1} \\ \text{Obs. 2} \\ \text{Obs. 3} \end{matrix} & \left[\begin{array}{cc} 0.6396 & 0.3604 \\ 0.7845 & 0.2155 \\ 0.3045 & 0.6955 \end{array} \right]. \end{matrix}$$

According to matrix $\hat{\Psi}$, if there is no earthquake recorded with magnitude $M \geq 6.5$ in the previous year, the system will remain probably in the same stress level in the current year. We come to this end because the transition probability $S_1 O_1$ to $S_1 O_1$ is estimated as 0.5838 and the transition probability from $S_2 O_1$ to $S_2 O_1$ is estimated as 0.5122. It should be noted that in matrix $\hat{\Psi}$, the row sums are equal to 1 and accordingly these probabilities are the highest probabilities relative to the other probabilities in the row which they belong.

If the strongest earthquake of the previous year is $M \in [6.5, 6.9]$, the behavior of the process will change depending on the previous hidden state. If the system is in the low-stress level, the strongest earthquake in the following year is certainly not between

6.5 and 6.9. The system remains at the same stress level and emits $M < 6.5$ with a probability of 0.6350, or $M \geq 7.0$ is emitted by making a transition to a high-stress level with a probability of 0.3650 for the current year. If the stress level is high in the previous year, it is very likely that the strongest earthquake in the current year occurs between 6.5 and 6.9. This is because the transition probability from S_2O_2 to S_1O_2 is estimated as 0.6883 and from S_2O_2 to S_2O_2 is estimated as 0.3005.

If the previous year's strongest earthquake is $M \geq 7.0$, which is the most severe class, and the corresponding stress level is low, the system will definitely make a transition to the high-stress level before the strongest earthquake occurs in the current year. In addition, on the condition that the previous stress level is high; the process will emit S_2O_3 again with a probability of 0.3043, the process will remain in the same stress level but the strongest earthquake observed will $M < 6.5$ with a probability of 0.4348. This may indicate that there may be aftershocks. With the same condition, with the probability of 0.2609, the system will make a transition to a low-stress level and the strongest earthquake will $M \in [6.5, 6.9]$.

The hidden state sequence is estimated by modified Viterbi algorithm and presented in Figure 4.2. According to Figure 4.2, it is revealed that *state 1* appears more often than *state 2* in the estimated state sequence. *State 1* is the most likely state to emit *obs. 1* which is the most common symbol of the observation sequence. Furthermore, the observed earthquake magnitudes scattered along the time and presented in Figure 4.3. It can also be seen from Figure 4.3 that, observation symbols *obs. 2* and *obs. 3* appear less frequently than *obs. 1*. Considering the relationship between stress levels and magnitudes, it is expected that *obs. 1* will be emitted in *state 1* and, *obs. 2* and *obs. 3* which are stronger magnitude classes will be emitted in *state 2*. The consistency between the actual observations and the estimated states can be seen when two figures examined together.

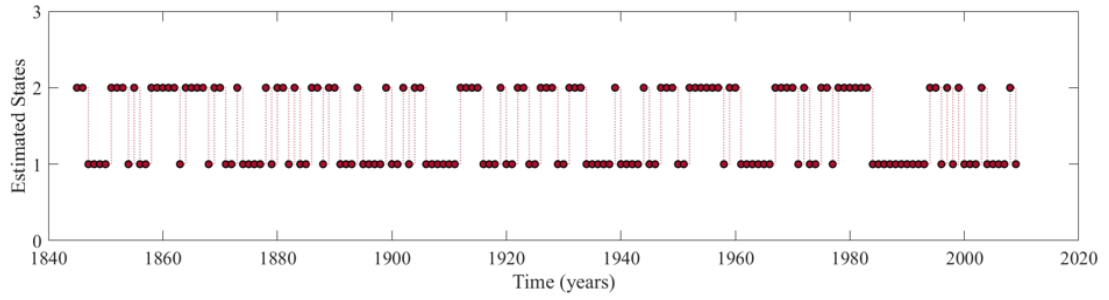


Figure 4.2 Distribution of the decoded hidden state sequence between 1845 and 2010

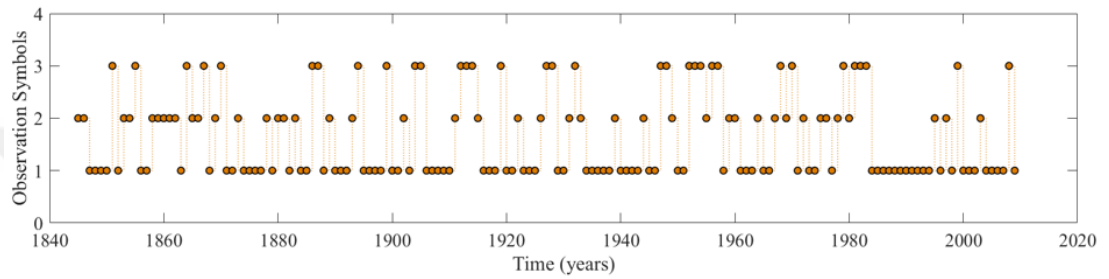


Figure 4.3 Distribution of the actual observation sequence between 1845 and 2010

The hidden state sequence with a length of $T = 165$ is estimated using the final values of parameter estimates by the modified Viterbi algorithm. *State 1* (low stress level) appears 88 times and *state 2* (high stress level) appears 77 times in the hidden sequence. In addition, *obs. 1* is emitted 89 times, *obs. 2* is emitted 43 times and *obs. 3* is emitted 33 times between 1845 and 2010. Table 4.10 shows how many times *obs. 1* and *obs. 2* occur in which hidden state.

Table 4.10 Matching of hidden decoded states and observation sequence

	Number of obs.	State 1	State 2
Obs. 1	89	85	4
Obs. 2	43	3	40
Obs. 3	33	0	33

The actual observations are classified and the number of observations in each class, as well as the decoded state in which these observations emitted, are shown in Table 4.10. For example, *obs. 1* is emitted 89 times and 85 of them are emitted in

state 1. The pattern of the decoded states obtained by BDHMM's final estimates matches with the pattern of the observation sequence, and the matching is almost error-free.

It is possible to obtain the transition probabilities between the observation symbols from the actual data set and summarize them in a matrix. The number of transitions between observation symbols is divided by row sums to calculate the transition probabilities between observations. Hence, it is possible to compare the observation transition probabilities of actual data with the observation transition probabilities of the generated data. Two observation sequences with the length of 165 which is the same length with the actual data is generated using the final parameter estimate values both for 2-state-3-observation BDHMM and 2-state-3-observation HMM. The frequency matrix of observation transitions is constructed and the observation transition probabilities of generated data are obtained both for BDHMM and HMM. The generated and actual observation probability transition matrices are as follows:

$$\mathbf{O}_{gen}^{HMM} = \begin{matrix} & \text{Obs. 1} & \text{Obs. 2} & \text{Obs. 3} \\ \text{Obs. 1} & [0.8039 & 0.1246 & 0.0715] \\ \text{Obs. 2} & [0.5781 & 0.2174 & 0.2045] \\ \text{Obs. 3} & [0.7064 & 0.1605 & 0.1331] \end{matrix},$$

$$\mathbf{O}_{gen}^{BDHMM} = \begin{matrix} & \text{Obs. 1} & \text{Obs. 2} & \text{Obs. 3} \\ \text{Obs. 1} & [0.5682 & 0.3068 & 0.1250] \\ \text{Obs. 2} & [0.5349 & 0.2093 & 0.2558] \\ \text{Obs. 3} & [0.4848 & 0.2424 & 0.2728] \end{matrix},$$

$$\mathbf{O} = P(Y_{t+1} = j | Y_t = i) = \begin{matrix} & \text{Obs. 1} & \text{Obs. 2} & \text{Obs. 3} \\ \text{Obs. 1} & [0.6023 & 0.2841 & 0.1136] \\ \text{Obs. 2} & [0.4884 & 0.2325 & 0.2791] \\ \text{Obs. 3} & [0.4546 & 0.2121 & 0.3333] \end{matrix}.$$

The goodness of fit test for testing the null hypothesis of equality of the generated and the actual observation transition probability matrices is performed again (Bhat &

Miller, 2002). The test statistic for the 2-state-3-observation BDHMM is $\chi^2_{test}^{BDHMM} \cong 1.3578$, the test statistic for the 2-state-3-observation HMM is $\chi^2_{test}^{HMM} \cong 19.6792$, and the critical value is $\chi^2_{0.05,6} = 12.5916$. Calculated test statistic does not exceed the critical value and the null hypothesis cannot be rejected for the 2-state-3-observation BDHMM. The difference between the observed and the estimated values are not significant for 2-state-3-observation BDHMM. The result is strengthening our intuition that the estimated model parameters are not local maximum and good convergence is provided. On the other hand, calculated test statistic exceeds the critical value and the null hypothesis is rejected for the 2-state-3-observation HMM. It is concluded that, by using 2-state-3-observation HMM, there is a significant difference between the observation transition matrix of actual data and the observation transition matrix of generated data. Hence, modeling with BDHMM fits the observation sequence better relative to classical HMM for the 165 strong earthquakes of interest.

4.4.2 Bootstrap Confidence Intervals for the 2-State-3-Obs. BDHMM Parameters

In this section, confidence intervals (CI) are calculated for the 2-state-3-observation BDHMM parameters with bootstrap method (Efron & Tibshirani, 1993). The detailed information about the method was given in Section 3.4.2.

95% CIs are constructed for each parameter by generating observation sequences with a length of 500, 1000 and 10000. The final estimate values of 2-state-3-observation BDHMM are used for generating the observation sequences. Each observation has a length of 165 which is the same as with the actual observation sequence. The estimates and standard deviations of related estimate for the parameters Ψ and ϕ are given in Table 4.11 and Table 4.12, respectively. There is no variability in the estimates of initial state distribution, and the process starts from *state 2* with a probability of 1, for all generated observation sequences.

Table 4.11 Estimate values and estimated standard deviations of the parameter Ψ

Parameter	$\psi_{11}(1,1)$	$\psi_{12}(1,1)$	$\psi_{11}(1,2)$	$\psi_{12}(1,2)$	$\psi_{11}(1,3)$	$\psi_{12}(1,3)$
Estimate	0.5838	0.0134	0.1942	0.0348	0.1738	0
M	$\hat{\sigma}_{\psi}$					
500	0.1393	0.0733	0.0581	0.0263	0.0832	1.78×10^{-8}
1000	0.1165	0.0608	0.0540	0.0217	0.0742	1.73×10^{-8}
10000	0.0852	0.0448	0.0444	0.0206	0.0665	1.47×10^{-8}
Parameter	$\psi_{21}(1,1)$	$\psi_{22}(1,1)$	$\psi_{21}(1,2)$	$\psi_{22}(1,2)$	$\psi_{21}(1,3)$	$\psi_{22}(1,3)$
Estimate	0.0997	0.5122	0.2872	0.1009	0	0
M	$\hat{\sigma}_{\psi}$					
500	0.1536	0.0764	0.0832	0.0529	0	0.0011
1000	0.1384	0.0673	0.0671	0.0496	0	0.0010
10000	0.1168	0.0404	0.0560	0.0404	0	0.0008
Parameter	$\psi_{11}(2,1)$	$\psi_{12}(2,1)$	$\psi_{11}(2,2)$	$\psi_{12}(2,2)$	$\psi_{11}(2,3)$	$\psi_{12}(2,3)$
Estimate	0.6350	0	0	0	0	0.3650
M	$\hat{\sigma}_{\psi}$					
500	0.0865	0	0	0	0	0.0865
1000	0.0678	0	0	0	0	0.0678
10000	0.0538	0	0	0	0	0.0538
Parameter	$\psi_{21}(2,1)$	$\psi_{22}(2,1)$	$\psi_{21}(2,2)$	$\psi_{22}(2,2)$	$\psi_{21}(2,3)$	$\psi_{22}(2,3)$
Estimate	0	0.0109	0.6883	0.3005	0	0.0003
M	$\hat{\sigma}_{\psi}$					
500	0	0.1193	0.1342	0.1541	0	0.0106
1000	0	0.1030	0.1168	0.1401	0	0.0094
10000	0	0.0801	0.0913	0.1156	0	0.0086
Parameter	$\psi_{11}(3,1)$	$\psi_{12}(3,1)$	$\psi_{11}(3,2)$	$\psi_{12}(3,2)$	$\psi_{11}(3,3)$	$\psi_{12}(3,3)$
Estimate	0	0.5000	0	0.1000	0	0.4000
M	$\hat{\sigma}_{\psi}$					
500	0	0.0657	0	0.0735	0	0.1451
1000	0	0.0515	0	0.0551	0	0.1218
10000	0	0.0458	0	0.0371	0	0.0873
Parameter	$\psi_{21}(3,1)$	$\psi_{22}(3,1)$	$\psi_{21}(3,2)$	$\psi_{22}(3,2)$	$\psi_{21}(3,3)$	$\psi_{22}(3,3)$
Estimate	0	0.4348	0.2609	0	0	0.3043
M	$\hat{\sigma}_{\psi}$					
500	0	0.1050	0.0920	0	0	0.0917
1000	0	0.0927	0.0708	0	0	0.0829
10000	0	0.0586	0.0587	0	0	0.0569

Table 4.12 Estimate values and estimated standard deviations of the parameter ϕ

Parameter	$\phi_1(1)$	$\phi_2(1)$	$\phi_1(2)$	$\phi_2(2)$	$\phi_1(3)$	$\phi_2(3)$
Estimate	0.6396	0.3604	0.7845	0.2155	0.3045	0.6955
M	$\hat{\sigma}_\phi$					
500	0.1381	0.1381	0.1100	0.1100	0.0798	0.0798
1000	0.1217	0.1217	0.0953	0.0953	0.0771	0.0771
10000	0.0815	0.0815	0.0793	0.0793	0.0442	0.0442

As the number of generated observation sequences increases, the 95% CI for the related parameter gets narrow, since the standard deviation decreases. The final maximum likelihood estimates for some of the parameters were found to be zero by the modified Baum-Welch algorithm. For all observation sequences with the length of 500, 1000 and 10000, the estimates of the related parameters are also found zero. In this case, no variation exists and the CI cannot be constructed.

4.4.3 Results

For the 2-state-3-observation BDHMM, the estimated observation transition probabilities are obtained by the marginalization and normalization of the $\hat{\Psi}$ matrix according to the hidden states and given below.

	Obs. 1	Obs. 2	Obs. 3
Obs. 1	0.6046	0.3085	0.0869
Obs. 2	0.3229	0.4944	0.1827
Obs. 3	0.4674	0.1804	0.3522

From the estimated observation transition matrix, if the strongest earthquake of the previous year is smaller than 6.5 or in the class of [6.5, 6.9], the strongest earthquake in the following year is most likely to be in the same class. The strongest class of an earthquake ($M \geq 7.0$) is probably followed by an obs. 1 ($M < 6.5$) class of earthquake with probability 0.4674 or an earthquake in the same class is occurred next year with a probability of 0.3522. The given matrix is a transition matrix of an ergodic and irreducible Markov chain, and the steady-state probabilities for each state of $\hat{\Psi}$ matrix can be obtained as follows:

$$\begin{array}{cccccc}
s_1 o_1 & s_2 o_1 & s_1 o_2 & s_2 o_2 & s_1 o_3 & s_2 o_3 \\
[0.3500 & 0.1972 & 0.1984 & 0.0545 & 0.0608 & 0.1391].
\end{array}$$

In addition, the steady-state probability vector Π for magnitude classes can be obtained as follows based on the marginalization of hidden states:

$$\begin{array}{ccc}
o_1 & o_2 & o_3 \\
\Pi = [0.5472 & 0.2529 & 0.1999].
\end{array}$$

The steady-state probabilities indicate that in the long term, the strongest earthquake occurs in a given year has a magnitude smaller than 6.5 with the probability of 0.5472, a magnitude in the class of [6.5, 6.9] with the probability of 0.2529, and a magnitude greater or equal to 7.0 with the probability of 0.1999.

Let m_{ii} be the expected number of transitions before the system returns state i , given that the current state is i . m_{ii} is called the *mean first return time*. For an ergodic irreducible Markov chain, mean first return times can be calculated from $m_{ii} = 1/\Pi_i$ and obtained as follows:

$$\begin{array}{ccc}
m_{11} & m_{22} & m_{33} \\
[1.8275 & 3.9541 & 5.0025].
\end{array}$$

If the magnitude of the strongest earthquake occurs is smaller than 6.5, it repeats for the first time after a mean of 1.83 years. If the magnitude of the strongest earthquake occurs is in the class of [6.5, 6.9], it repeats for the first time after 3.95 years on the average. Furthermore, if the magnitude of the strongest earthquake occurred is greater or equal to 7.0, it repeats for the first time after 5.00 years on the average.

CHAPTER 5

CONCLUSION

The main purpose of this thesis is to provide two extensions of hidden Markov model (HMM), which are called MEHMM and BDHMM. For this purpose, firstly, classical HMM is explained in detail. Observations are assumed to be conditionally depend on only the current hidden state and to be i.i.d. among themselves in classical HMMs. This assumption is not valid for the proposed models. The first model, MEHMM, assumes that there is a first-order Markov dependence between observations. The hidden state transitions occur according to a Markov chain and state transitions are independent of the observation symbols. On the other hand, the second model, BDHMM, assumes that each observation symbol and hidden state are first-order Markov dependent on the previous observation symbol and hidden state. The transition from the previous observation symbol to the current observation symbol is considered coordinated with the state transitions; hence the state-observation distribution exists in the proposed model different from classical HMM. The Baum-Welch algorithm and the Viterbi algorithm are modified as well, according to the assumptions and the parameters of MEHMM and BDHMM.

Experimental studies are performed to evaluate the performance of both MEHMM and BDHMM compared to classical HMM. Different observation sequences which are generated from HMM, MEHMM, or BDHMM are generated for 1000 times. They are modeled by using HMM and also the proposed models. The most likely model for the related observation sequence is selected by the help of the AIC and the BIC values. As a result, HMM offers better results for the observation sequences which have conditionally independent observation symbols, compared to MEHMM and BDHMM, whereas MEHMM provides better results for the observation sequences generated based on MEHMM assumptions and BDHMM provides better results for the binary dependent observation sequences. This is the expected result when a Markov-emission between observations or a binary dependency exists on the observation sequence, MEHMM and BDHMM outperforms compared to classical HMM.

The number of freely estimated parameters of MEHMM and BDHMM is more than those of classical HMM since the dimensions of some parameter matrices are higher for the proposed models. Although the number of freely estimated parameters is more than HMM, the proposed models are better when the observation sequences hold MEHMM and BDHMM assumptions. Thus, MEHMM and BDHMM contains more information compared to classical HMM. Besides, MEHMM and BDHMM can be considered as alternative models to reveal the first-order Markov dependence between observations and the binary first-order Markov dependence between the observation symbols and hidden states, respectively. If the AIC and the BIC values for MEHMM and BDHMM are smaller than the classical HMM, it means less information is lost by modeling by the proposed models. Then it can be concluded that HMM is inappropriate for the related observation sequence and one of the proposed models can be an alternative way for modeling.

Finally, two different case studies namely on the strong earthquake occurrences and the closing-value-trend of stock prices are presented. MEHMM is compared with the classical HMM to model the daily trend of the stock closing prices of the Microsoft Corporation. The observations are assumed to be the relative change of the stock closing value and the hidden states are assumed to be the market conditions. After modeling the stock price data by using two models for different number of hidden states and observations, MEHMM outperforms for the probabilistic prediction of daily trend of the closing value compared to classical HMM. It is concluded that the relative change of the stock closing price according to the previous day has a Markov dependence with the previous relative change. For an application of BDHMM, strong earthquakes that occur in the Aegean Sea region in a certain time period are modeled. The observations are assumed to be the classes including the strongest earthquake that occurred in a year and the hidden states are assumed to be the stress level of the ground. The aim was to determine whether there is a Markov dependence between the strongest earthquakes and related stress levels between consecutive years. After modeling the real-life earthquake data by using two models for different number of hidden states and observations, BDHMM outperforms for the modeling of the yearly strongest earthquakes for a given time period compared to classical HMM. It is concluded that

BDHMM provides a better model and that the next strongest earthquake and the related stress level depend on the current years' strongest earthquake and its stress level for the study area of interest.

To sum up briefly, the fact that HMM can be insufficient for some real-life problems due to its assumptions led to the emergence of new types of modified or extended HMMs, where two of them are proposed in this thesis. These models assume only a first-order Markov dependence between the observations or a binary first-order Markov dependence.

Further research directions. The construction of extended HMMs with higher-order Markov dependence structure can be considered as a further study. The hidden semi-Markov model is also an important model for the real-life applications and adaptation of MEHMM and BDHMM to hidden semi-Markov model is one of the further studies to consider.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716-723.
- AlKhateeb, J. K., Ren, J., Jiang, J., & Al-Muhtaseb, H. (2011). Offline handwritten Arabic cursive text recognition using hidden Markov models and re-ranking. *Pattern Recognition Letters*, 32 (8), 1081-1088.
- Altinok, Y., & Kolcak, D. (1999). An application of the semi-Markov model for earthquake occurrences in North Anatolia, Turkey. *Journal of the Balkan Geophysical Society*, 2 (4), 90-99.
- Alvarez, E. E. (2005). Estimation in stationary Markov renewal processes, with application to earthquake forecasting in Turkey. *Methodology and Computing in Applied Probability*, 7 (1), 119-130.
- Anagnos, T., & Kiremidjian, A. S. (1988). A review of earthquake occurrence models for seismic hazard analysis. *Probabilistic Engineering Mechanics*, 3 (1), 3-11.
- Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Inequalities III: Proceedings of the third symposium on inequalities* (1st ed.) (1-8). Los Angeles: University of California, Academic Press.
- Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37 (6), 1554-1563.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41 (1), 164-171.

- Berchtold, A. (1999). The double chain markov model. *Communications in Statistics – Theory and Methods*, 28 (11), 2569-2589.
- Bhat, U. N., & Miller, G. K. (2002). *Elements of applied stochastic processes* (3rd ed.). New Jersey: Wiley.
- Bose, C. B., & Kuo, S. (1994). Connected and degraded text recognition using hidden Markov model. *Pattern Recognition*, 27 (10), 1345-1363.
- Bryan, J. D., & Levinson, S. E. (2015). Autoregressive hidden Markov model and the speech signal, *Procedia Computer Science*, 61, 328-333.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.
- Chambers, D. W., Baglivo, J. A., Ebel, J. E., & Kafka, A. L. (2012). Earthquake forecasting using hidden Markov models. *Pure and Applied Geophysics*, 169 (4), 625-639.
- Chambers, D. W., Ebel, J. E., Kafka, A.L., & Baglivo, J. A. (2003). *A hidden Markov approach to modeling interevent earthquake times*. Retrieved January, 2020 , from https://www.researchgate.net/publication/234441301_A_Hidden_Markov_Approach_to_Modeling_Interevent_Earthquake_Times
- Ching, W. K., Fung, E. S., & Ng, M. K. (2003). Higher-order hidden Markov models with applications to DNA sequences. *Lecture Notes in Computer Science*, 2690, 535-539.
- de Angelis, L., & Paas, L. J. (2013). A dynamic analysis of stock markets using a hidden Markov model. *Journal of Applied Statistics*, 40 (8), 1682-1700.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38.
- du Preez, J. A. (1998). Efficient training of high-order hidden Markov models using first-order representations. *Computer Speech & Language* 12 (1), 23-39.
- Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids* (1st ed.). Cambridge: Cambridge University Press.
- Ebel, J. E., Chambers, D. W., Kafka, A. L., & Baglivo, J. A. (2007). Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California. *Seismological Research Letters*, 78 (1), 57-65.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. United Kingdom: Chapman & Hall/CRC.
- Granat, R., & Donnellan, A. (2002). A hidden Markov model based tool for geophysical data exploration. *Pure and Applied Geophysics*, 159 (10), 2271-2283.
- Hassan, M. R., & Nath, B. (2005). Stock market forecasting using hidden Markov model: a new approach. *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications (ISDA '05), Warsaw*, 192-196.
- Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33 (3), 251-272.
- Kavitha, G., Udhayakumar, A., & Nagarajan, D. (2013). Stock market trend analysis using hidden Markov models. *International Journal of Computer Science and Information Security*, 11, 103-110.

- Li, Y. (2005). Hidden Markov models with states depending on observations. *Pattern Recognition Letters*, 26 (7), 977-984.
- Masala, G. (2012). Earthquakes occurrences estimation through a parametric semi-Markov approach. *Journal of Applied Statistics*, 39 (1), 81-96.
- Nguyen, N. (2018). Hidden Markov model for stock trading. *International Journal of Financial Studies*, 6 (2), 1-17.
- Orfanogiannaki, K., Karlis, D., & Papadopoulos, G. A. (2010). Identifying seismicity levels via Poisson hidden Markov models. *Pure and Applied Geophysics*, 167 (8-9), 919-931.
- Pertsinidou, C. E., Tsaklidis, G., Papadimitriou, E., & Limnios, N. (2017). Application of hidden semi-Markov models for the seismic hazard assessment of the North and South Aegean Sea, Greece. *Journal of Applied Statistics*, 44 (6), 1064-1085.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2), 257-286.
- Rasku, J., Juhola, M., Toppila, E., Pyykkö, I. (2007). Recognition of balance signals between healthy subjects and otoneurological patients with hidden Markov models, *Biomedical Signal Processing and Control*, 2 (1), 1-8.
- Reid, H. F. (1910). *The mechanics of the earthquake, the California earthquake of April 18, 1906. Report of the state investigation comission*, Vol. 2. Washington D.C. : Carnegie Institution of Washington (reprinted in 1969).
- Sadeghian, R. (2012). Forecasting time and place of earthquakes using a semi-Markov model: with case study in Tehran province. *Journal of Industrial Engineering International*, 8, 1-7.

- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461-464.
- Stratonovich, R. L. (1960). Conditional Markov processes. *Theory of Probability and its Applications*, 5 (2), 156-178.
- Tamposis, I. A., Theodoropoulou, M. C., Tsirigos, K. D., & Bagos, P. G. (2018). Extending hidden Markov models to allow conditioning on previous observations. *Journal of Bioinformatics and Computational Biology*, 16 (5), 1-21.
- Tiampo, K. F., & Shcherbakov, R. (2012). Seismicity-based earthquake forecasting techniques: Ten years of progress. *Tectonophysics*, 522-523, 89-121.
- Vere-Jones, D. (1978). Earthquake prediction: a statistician's view. *Journal of Physics of the Earth*, 26 (2), 129-146.
- Visser, I., Raijmakers, M. E. J., & Molenaar, P. C. M. (2000). Confidence intervals for hidden Markov model parameters. *British Journal of Mathematical and Statistical Psychology*, 53 (2), 317-327.
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 (2), 260-269.
- Votsi, I., Limnios, N., Papadimitriou, E., & Tsaklidis, G. (2019). *Earthquake statistical analysis through multi-state modeling*. New Jersey: Wiley.
- Votsi, I., Limnios, N., Tsaklidis, G., & Papadimitriou, E. (2012). Estimation of the expected Number of earthquake occurrences based on semi-Markov models. *Methodology and Computing in Applied Probability*, 14 (3), 685-703.

- Votsi, I., Limnios, N., Tsaklidis, G., & Papadimitriou, E. (2013). Hidden Markov models revealing the stress field underlying the earthquake generation. *Physica A: Statistical Mechanics and its Applications*, 392 (13), 2868-2885.
- Yip, C. F., Ng, W. L., & Yau, C. Y. (2018) A hidden Markov model for earthquake prediction. *Stochastic Environmental Research and Risk Assessment*, 32 (5), 1415-1434.
- Zhang, X., Li, Y., Wang, S., Fang, B., & Yu, P. S. (2019). Enhancing stock market prediction with extended coupled hidden Markov model over multi-sourced data. *Knowledge and Information Systems*, 61, 1071-1090.
- Zucchini, W., & Guttorp, P. (1991). A hidden Markov model for space-time precipitation. *Water Resources Research*, 27 (8), 1917-1923.