Enhancements in Markovian Dynamics

Reza Ali Akbar Soltan

Dissertation, submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Mechanical Engineering

Mehdi Ahmadian, Chair
Saied Taheri
T. Simin Hall
Steve C. Southward
Farshid M. Asl
Joseph A. Ball

March 27, 2012
Blacksburg, Virginia

Keywords:
Nonlinear Stochastic Model, Hidden Markov Model,
Maximum Likelihood Estimation, Expectation Maximization,
Duration Dependent Hidden Markov

Enhancements in Markovian Dynamics

Reza Ali Akbar Soltan

ABSTRACT

Many common statistical techniques for modeling multidimensional dynamic datasets can be seen as variants of one (or multiple) underlying linear/nonlinear model(s). These statistical techniques fall into two broad categories of *supervised* and *unsupervised* learning. The emphasis of this dissertation is on unsupervised learning under multiple generative models. For linear models, this has been achieved by collective observations and derivations made by previous authors during the last few decades. Factor analysis, polynomial chaos expansion, principal component analysis, gaussian mixture clustering, vector quantization, and Kalman filter models can all be unified as some variations of unsupervised learning under a single basic linear generative model. Hidden Markov modeling (HMM), however, is categorized as an unsupervised learning under multiple linear/nonlinear generative models. This dissertation is primarily focused on hidden Markov models (HMMs).

On the first half of this dissertation we study enhancements on the theory of hidden Markov modeling. These include three branches: 1) a robust as well as a closed-form parameter estimation solution to the expectation maximization (EM) process of HMMs for the case of elliptically symmetrical densities; 2) a two-step HMM, with a combined state sequence via an extended Viterbi algorithm for smoother state estimation; and 3) a duration-dependent HMM, for estimating the expected residency frequency on each state. Then, the second half of the dissertation studies three novel applications of these methods: 1) the applications of Markov switching models on the Bifurcation Theory in nonlinear dynamics; 2) a Game Theory application of HMM, based on fundamental theory of card counting and an example on the game of Baccarat; and 3) Trust modeling and the estimation of trustworthiness metrics in cyber security systems via Markov switching models.

As a result of the duration dependent HMM, we achieved a better estimation for the expected duration of stay on each regime. Then by robust and closed form solution to the EM algorithm we achieved robustness against outliers in the training data set as well as higher computational efficiency in the maximization step of the EM algorithm. By means of the two-step HMM we achieved smoother probability estimation with higher likelihood than the standard HMM.

Acknowledgements

First and foremost, I would like to give my very special thanks to my advisor Dr. Mehdi Ahmadian. Mehdi's view of system dynamics and controls theory and his mathematical discipline taught me to see nonlinear control and nonlinear dynamics from a new perspective that enhanced my understanding of what constitutes an acceptable solution for certain fundamental problems in dynamical systems. I can never thank him enough for giving me the freedom to choose the fields of my interest and helping me to pursue my goals. I have always valued his advice greatly, and I always will.

Secondly, I would like to give my very special thanks to my dear friend from several years ago, my manager when we work together and my PhD committee member at school; Dr. Farshid Maghami Asl. Since nearly seven years ago, Farshid has been one of my greatest mentors and his views, advises, suggestions and feedbacks have shaped and evolved my path of life. I have always valued his advice greatly, and I will never forget all his help and support.

The past several years at Virginia Tech were a wonderful opportunity for me to get engaged in interesting discussions with several VT professors and scientists. My other committee members Dr. Saied Taheri, Dr. Steve Southward, Dr. Joe Ball and Dr. Simin Hall are a few of

them. I am very grateful for this invaluable opportunity to work with them and I appreciate their advice on many important occasions by heart.

I owe my sincere gratitude to Dr. Ali Ghaffari, my previous advisor at K. N. Toosi University of Technology, who has greatly helped me during my very important situations and tough times. I will never forget Dr. Ghaffari's support, help and his great spirit. To me Dr. Ghaffari has been and will always be more than my advisor. I admire him and I treasure him like my own father, and I can never thank him enough for trusting me and bringing me back on the track.

I am incredibly grateful to my family: my beautiful and lovely mother Maryam, my hero father Ghasem, my beautiful sister Shaghayegh, my handsome brother Ramin and my kind and great brother in law Sadegh. We also have a new addition to our family: Shaili my lovely, beautiful, adorable niece. They always gave me their unconditional love and support during all these years. While I never had the opportunity to see my family during my five years of Masters and PhD studies in the United States, my heart has always been with them, specially my Mom who never stopped believing in me and has done everything in her power and beyond to help me achieve my goals during the whole course of my entire life.

Finally, my very special thanks go to the kindest person I have ever known, my lovely grandmother Fatemeh and also my aunts and uncles Masoumeh, Manijeh, Mansoureh, Hossein and Majid. Their beautiful smiles, emotional supports, and strong encouragements were the driving force of my efforts all these years. Without my family's help and enthusiasm, this work certainly would have never been possible.

Content

ACKNOWLEDGEMENTS ..... **IV**

1 INTRODUCTION **1**

1.1 OBJECTIVES 2

1.2 APPROACH 3

1.3 OUTLINE 4

1.4 CONTRIBUTIONS 4

2 BACKGROUND **6**

2.1 MEAN 7

2.2 VARIANCE AND COVARIANCE 8

2.3 PROBABILITY DENSITY FUNCTION 9

2.4 TRANSFORMING NOISE SPACE WITHOUT LOSS OF GENERALITY 11

2.5 PROBABILITY COMPUTATION 12

2.6 INFERENCE VS. SYSTEM IDENTIFICATION 14

2.6.1 Inference: Filtering and Smoothing *15*

2.6.2 System Identification: Expectation-Maximization (EM) *17*

3 HIDDEN MARKOV MODELS **21**

3.1 CONTINUOUS-TIME HIDDEN MARKOV MODEL 23

3.2 THE STRUCTURE OF MVN CONTINUOUS-TIME HMM 26

3.3 EXTENDED BAUM-WELCH ALGORITHM 31

3.3.1 Robust Parameter Estimation *34*

3.4 A CLOSED-FORM SOLUTION TO EM ALGORITHM 35

3.4.1 Features of the closed-form re-estimations *36*

3.4.1.1 Multivariate Gaussian 40

3.4.1.2 Multivariate Cauchy 41

4 TWO-STEP HIDDEN MARKOV MODEL **42**

4.1 FIRST STEP: STANDARD HMM 43

4.2 SECOND STEP: HMM WITH COMBINED STATE 44

4.2.1 Steady vs. Transient Probabilities *45*

4.3 A 2-STEP HMM EXPERIMENT 47

4.3.1 A Comparison *50*

5 DURATION DEPENDENT HMM **52**

5.1 BACKGROUND OF THE DURATION DEPENDENT HMM 52
5.2 A NOVEL DERIVATION OF DURATION-DEPENDENCY 56
5.3 EXAMPLE 66
5.3.1 Remarks *67*
5.3.2 Comparison *67*
6 HMM IN BIFURCATION THEORY 69
6.1 LOCAL BIFURCATION 69
7 APPLICATION OF HMM IN MEAN-REVERTING PROCESSES 73
7.1 THE GAME OF BACCARAT 74
7.1.1 How to use HMM to play? *77*
7.1.2 Optimal Leverage Factor *78*
7.2 TRUSTWORTHINESS IN CYBER SECURITY 80
7.2.1 Trust Model *82*
8 CONCLUSION 86
8.1 RECOMMENDATION FOR FUTURE STUDIES 88
REFERENCES 89

LIST OF FIGURES

Figure 2-1. The block diagram of the linear dynamic system model 11
Figure 3-1: A two state Markov switching model..... 23
Figure 3-2: A 3D Time series of a Markov switching model with two states..... 25
Figure 4-1: A sample 2-state HMM with its univariate time series..... 48
Figure 4-2: (a) The true 2-state sequence and the first step estimation of Markov probabilities, (b) The second step observation sequence, calculated as the difference of the first step Markov probabilities..... 48
Figure 4-3: The second step estimated 4-states for the second step observation sequence..... 49
Figure 4-4 Top. (a) The first step versus second step state probability estimations. 49
Figure 5-1: Kernel of a Sigmoid function..... 56
Figure 5-2: The truncated distribution of the duration of stay at each regime for a 2-state HMM..... 63
Figure 5-3: (a) A sample univariate time series, as well as the Markov states and their estimations. (b) Duration Sigmoid functions are shown to illustrate the theory..... 66
Figure 5-4: State probabilities versus joint State and Duration probabilities 68
Figure 6-1: The trajectory and its nonlinear model with four equilibriums 70
Figure 6-2: Locus of 4 equilibriums in a financial time series 71
Figure 6-3: Locus of the stable equilibrium and its basin of attraction..... 71
Figure 7-1: (a) The time series of $y(t)$, the moving average of length 1000, and (b) $w(t)$. For illustration purposes $y(t)$ and $w(t)$ are also shown in the time interval of [0 5200]..... 75
Figure 7-2: (a) The output of the regime switching model on the input signal $w(t)$. The top figure shows the signal $w(t)$ and its regime switching between 2 states, one with a local mean of +3 and one with the local mean of -3. (b) Markov probabilities..... 77
Figure 7-3: The zoomed-in version of the last figure, at the time span of [1500,2500], for better clarity of the concept..... 78
Figure 7-4: (a) The input time series Y(t) and (b) its regime switching..... 84

LIST OF TABLES

Table 5-1: A possible realization of 2-state process states and corresponding durations 56
Table 5-2: Elements of the Duration Dependent Model..... 57

1 Introduction

During the last decade there has been an increasing attention to stochastic continuous-time-discrete-state dynamical systems, modeled by hidden Markov models (HMM). The proposed method is based on combing mixture models and Markov switching processes to model the temporal structure of time series. This combination historically has been capable of point forecasting as well as density forecasting [1]. Meanwhile, we bear in mind that common statistical techniques for modeling multidimensional static datasets and multidimensional time series also include (but not limited to) factor analysis (FA), principle component analysis (PCA), Gaussian mixture clustering (GMC), vector quantization, independent component analysis (ICA), and Kalman filter models (also known as linear dynamical model) [2].

It is noticeable that some of these studies, done by mathematicians have been published in math journals, which typically are not read by engineers. Also some of these techniques are brought into a very specific branch of science and stayed there for decades, while the same underlying ideas in other branches of science suffers from the lack of that knowledge. An example of that is Kalman filter, which has been extensively used by control engineers, since 1960 when Kalman presented the theory [3]. Then, Linear Quadratic Regulators (LQR), when corrupted with Gaussian noise, enjoyed the existence of Kalman filter and a new branch of

control design, called Linear Quadratic Gaussian (LQG) controllers, was created. On the other hand, HMM has been extensively used in speech recognition research communities for decades [4-6].

Lets take a quick look at the recent literature for these related works. Hinton *et al*. in [7] used Maximum Likelihood Estimation (MLE) techniques via PCA and FA for recognizing handwritten digits using mixture of linear models. Furthermore, within the framework of [7] it was noted that a mixture of local linear models is an effective way to capture the underlying styles of handwritten digits. HMM, as we mentioned before, when the state process is modeled as a Markov chain, has been successfully applied to speech recognition [4, 5, 8-12], blind equalization of data transmission systems [13], image segmentation [14], and other applications [15].

1.1 Objectives

This dissertation aims for two main objectives: 1) theoretical enhancements in Markovian dynamics and hidden Markov modeling and 2) applications of these enhancements in science and engineering.

Therefore the first part studies in depth the theoretical enhancements. These advances include three main segments:

A robust, as well as a closed-form derivation of emission parameters estimation in multivariate continuous-time observation HMMs.

A new two-step derivation of first-order HMMs with combined state sequence for smoother probability estimation.

A novel duration modeling of HMM with sigmoid functions and the derivation of Markov transition probabilities, conditional on joint densities of states and durations.

Then the second part of this dissertation aims for applications of these advances in Markovian systems. The applications could also be seen as examples to illustrate the theory. These applications are:

Continuous time multivariate hidden Markov models in Bifurcation theory.

Applications in Game theory and the card game of Baccarat

The applications in cyber security and trust modeling

1.2 Approach

To accomplish the objectives of this dissertation we challenge each of the problems with novel approaches.

For robust emission parameter estimation, we break the Expectation-Maximization (EM) algorithm to its most basic components. Then in the maximization process, which historically uses a constraint optimization method e.g., Lagrange multipliers, we use a robust optimization method to damp out the outlier in the training data set.

For the closed-form solution to the EM algorithm, we take the likelihood surface, which is very complex in general, and break it down to convex components, and then model it with mixture of elliptically symmetric distributions where the closed-form solution does exist for them.

In the two-step first-order HMM, we run a standard HMM once and then take the difference of its probabilities over time as the observation sequence for the second step. Then we extract the transient and steady probabilities to achieve smoother probability estimation for the hidden state sequence.

In the duration dependent HMM, we model durations of each state explicitly with a sigmoid function and then we derive an extended forward-backward method for the joint densities of states and durations.

For the applications, however, we use a standard HMM with our closed-form solution to its EM algorithm and apply it to the card game of Baccarat. For that we use the fundamental theory of card counting to create a mean-reverting signal and then estimate and forecast the outcome of the game based on its past.

In Bifurcation theory we also used our standard HMM with closed-form solution and compared the Markov switching points in time with bifurcation points in the system's space.

And finally for estimating the trustworthiness we used a standard HMM as well as the duration dependent HMM to estimate the metric of trust based on an observation sequence.

1.3 Outline

Based on what we explained in the last two sections, the rest of this dissertation is organized as follows:

Chapter 2 gives a basic review and background of linear/linearized models.

Chapter 3 explains our derivation of continuous time multivariate HMM, in the context of system dynamics.

Chapter 4 introduces our derivation of two-step HMMs.

Chapter 5 introduces our derivation of the duration dependent HMM.

Chapter 6 introduces the applications of continuous time HMM in bifurcation theory.

Chapter 7 brings the application of HMM into "Mean Reverting" processes, with two examples of casino game of Baccarat and the metric of trustworthiness in cyber security

and finally Chapter 8 concludes this dissertation.

1.4 Contributions

By the end of these studies we claim that we solved some fundamental problems of HMM that the Markovian dynamics have been involved with for decades.

The first contribution of this work is to achieve computational efficiency on the order of 50 times faster than the current methods. Time consuming calculations and computational costs of HMM has been one of the biggest obstacles on the way of using it for real time applications. The closed-form solution to the local maxima of the likelihood surface of the emission model partially improves the existing answers to the problem of computational complexity.

The second contribution of this work is to have smoother and better state probability estimations. This is particularly important for out-of-sample application where HMM will typically have some delay to catch a potential regime switches.

The third contribution of this work is to effectively estimate the expected duration of stay on each regime with a set of sigmoid functions.

Applying these theories to different science and engineering problems are other contributions of this study.

2 Background

Markov switching models have applications in sudden regime shifting and basically modeling the time series by mixture of Gaussians where theoretically, a mixture of Gaussians is able to model any distribution [16]. Hidden Markov Modeling is a probabilistic technique for the study of possibly stochastic time series [17, 18]. Modeling with many of the probability distributions, the cost of implementation is linear with respect to the length of the data and models can be nested to reflect hierarchical sources of knowledge [18]. Although initially introduced in the late 1960s and early 1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years [4]. Tracing back the work of Markov [10] and Shannon [19, 20] was concerned with Markov chains. While the state sequence is observed in Markov chain [21], in hidden Markov model, the output properties impose a veil [22] between the state sequence and the observer of the time series. In the effort to lift the veil, a substantial body of theory was developed during the years of 1960s to 1990s. Leonard Baum, collaborating with Eagon and Petrie [23-25] dealt with finite probability spaces and addressed the problem of tractability of probability computation, iterative maximum likelihood estimation of model parameters from observed time series, recovery of hidden states, and the proof of consistency of the estimates [18].

The major development of the Hidden Markov theory (1970) was the maximization technique of Baum, Petrie, Soules and Weiss [26]. This was after two articles (1966 and 1967) by Baum *et al*, where they unveil some statistical inference for probabilistic functions of finite state Markov chains as well as proving an equality with applications to statistical estimation for probabilistic functions [24, 25]. Baum and his colleagues including Lloyd Welch at the Institute for Defense Analysis (IDA) in Princeton, NJ, made very important breakthroughs that led to a wide range of theoretical outgrowths in this branch of science. They included a number of generalizations of both the spectral and temporal components of the model, e.g. variable-duration hidden Markov models [27], hidden-filter hidden Markov models [17], and trainable finite state hidden grammars [28]. A special case of the results in [26] has been addressed by Dempster *et al,* as the Expectation Maximization (i.e. EM) algorithm [29, 30].

To be able to relate to this research and the terms that we will be using, knowledge of probability and distributions, mean, variance and covariance as well as probability and cumulative density functions (PDF and CDF) is needed. While we just touch based on these very basic preliminaries, we refer the interested reader to the work of Templeton [31] for a detailed background. Therefore this chapter gives the basic background necessary for the discussions in the following chapters.

2.1 Mean

In statistics, mean has two related meanings:

The arithmetic mean (and is distinguished from the geometric mean or harmonic mean).

The expected value of a random variable, which is also called the population mean.

There are other statistical measures that use samples that some people confuse with averages, including *median* and *mode*. Other simple statistical analyses use measures of spread, such as *range*, *interquartile range*, or *standard deviation*. For a real-valued random variable, the mean is the expectation of it. Note that not every probability distribution has a defined mean (or variance), e.g., Cauchy distribution as an example.

For a data set, the mean is the sum of the values divided by the number of values. This mean is a type of arithmetic mean. If the data set were based on a series of observations obtained

by sampling a statistical population, this mean is termed the "sample mean" to distinguish it from the "population mean". The mean is often quoted along with the standard deviation: the mean describes the central location of the data, and the standard deviation describes the spread. An alternative measure of dispersion is the mean deviation, equivalent to the average absolute deviation from the mean. It is less sensitive to outliers, but less mathematically tractable.

If a series of observations is sampled from a larger population (measuring the heights of a sample of adults drawn from the entire world population, for example), or from a probability distribution which gives the probabilities of each possible result, then the larger population or probability distribution can be used to construct a "population mean", which is also the expected value for a sample drawn from this population or probability distribution. For a finite population, this would simply be the arithmetic mean of the given property for every member of the population. For a probability distribution, this would be a sum or integral over every possible value weighted by the probability of that value. It is a universal convention to represent the population mean by the symbol μ. In the case of a discrete probability distribution, the mean of a discrete random variable x is given by taking the product of each possible value of x and its probability Pr(x), and then adding all these products together, giving $\mu = \Sigma(x.Pr(x))$ [32]. The sample mean may differ from the population mean, especially for small samples, but the law of large numbers dictates that the larger the size of the sample, the more likely it is that the sample mean will be close to the population mean [32].

2.2 Variance and Covariance

In probability theory and statistics, the variance is used as a measure of how far a set of numbers is spread out from each other. It is one of several descriptors of a probability distribution, describing how far the numbers lie from the mean (expected value). In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those based on moments are advantageous in terms of mathematical and computational simplicity.

The variance is a parameter describing in part either the actual probability distribution of an observed population of numbers, or the theoretical probability distribution of a not-fully-observed population of numbers. In the latter case a sample of data from such a distribution can be used to construct an estimate of its variance: in the simplest cases this estimate can be the sample variance, defined as:

$$\mathrm{Var}(x) = E[(x-\mu)^2],$$

with "E[.]" being the "expected value of". Covariance is a measure of how much two variables change together. Variance is a special case of the covariance when the two variables are identical:

$$\mathrm{Cov}(x,y) = E[(x-E[x])(y-E[y]].$$

2.3 Probability Density Function

In probability theory, a probability density function (pdf), or density of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point. The probability for the random variable to fall within a particular region is given by the integral of this variable's density over the region. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one. The terms probability density function, or simply probability function have also sometimes been used to denote the probability density function. However, special care should be taken around this usage since it is not standard among all statisticians. In other sources, "probability distribution function" may be used when the probability distribution is defined as a function over general sets of values, or it may refer to the cumulative distribution function, or it may be a probability mass function rather than the density. Further confusion of terminology exists because density function has also been used for what is here called the "probability mass function" [32].

The second half of this dissertation, however, studies the application of HMM in the bifurcation theory in the context of nonlinear dynamics. Then it introduces two novel

applications of HMM. The first one is related to forecasting the trend of a time series based on a relevant signal. This application was studied in the context of Game Theory based on fundamental theory of card counting, and was implemented on a game named: Baccarat. The second application, on the other hand, was to estimate and forecast a metric of trustworthiness in sensors and networks in the context of cyber security.

Within the context of this dissertation, in the first half, we explore two advancements in the computational efficiency and robustness of HMM, as well as two breakthroughs in the theory of HMM; The first theoretical enhancement is an extended Viterbi algorithm with a two-step HMM, where the step one is a first-order and the step two is a second-order HMM with combined state sequences, in order to achieve smoother state probability estimations. The second theoretical advancement of this dissertation is a duration-dependent derivation of HMMs in order to better estimate the expected time frame residency of each state with a set of sigmoid functions. This part of the research was motivated by excitatory and inhibitory interactions of neuron synopses in the central nervous system of the human body, modeled by Wilson and Cowan in [33].

All derivations of HMM that we study in this dissertation are continuous-time, discrete-state models. However, for the sake of completeness and a comparison, the basic models that we explain in this chapter are out of the context of HMM, related to linear dynamical systems with Gaussian noise. In such models we assume that the state of the process can be presented at any time by an N-vector of state variables: $\mathbf{x}$, which often cannot be observed directly. However, the system also generates an observable M-vector $\mathbf{y}$, which we do have access to, at each time step.

The state $\mathbf{x}$ is assumed to evolve by a first order Markovian dynamics and each output vector $\mathbf{y}$ is generated from the current state by a linear observation process. Both, the hidden state evolution and the observation sequence are corrupted by Gaussian noise and disturbances. In this regards the generative model can be written as [1, 2]:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{w}_\bullet & \mathbf{w}_\bullet &\sim \mathcal{N}(0,\mathbf{Q}) \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{v}_\bullet & \mathbf{v}_\bullet &\sim \mathcal{N}(0,\mathbf{R}) \end{aligned} \qquad \textbf{(2-1)}$$

where $\mathbf{A}$ is an $N \times N$ state transition matrix and $\mathbf{C}$ is a $M \times N$ observation emission matrix. Note that all vectors are column vectors.

To study the model we will use the following notation. To denote the transpose of a vector or matrix, we use the notation of a superscript "T", (e.g. $\mathbf{x}^T$). The determinant of the matrix is denoted by the norm sign (e.g. $|\mathbf{A}|$), and the matrix inversion is denoted by a superscript of "-1" (e.g. $\mathbf{x}^{-1}$). The symbol "~" means "distributed according to". Also a multivariate normal (gaussian) distribution with mean μ and covariance matrix Σ is written as $N(\mu, \Sigma)$. The same gaussian evaluated at point z is denoted as $N(\mu, \Sigma) \mid z$. N-vector $\mathbf{w}$ and M-vector $\mathbf{v}$ are random variables representing the state evolution noise and observation disturbance. These noises and disturbances are assumed to be independent of each other and independent of the values of $\mathbf{x}$ and $\mathbf{y}$. Furthermore both of these noise and disturbance sources are temporally white (uncorrelated from time to time), and spatially gaussian with zero mean and covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ respectively. Note that we denote them by $\mathbf{w}_\bullet$ and $\mathbf{v}_\bullet$ to emphasize that they do not have any knowledge about the time index of the evolution. Figure 2-1 shows the block diagram of the linear dynamic system of Eq. (2-1).

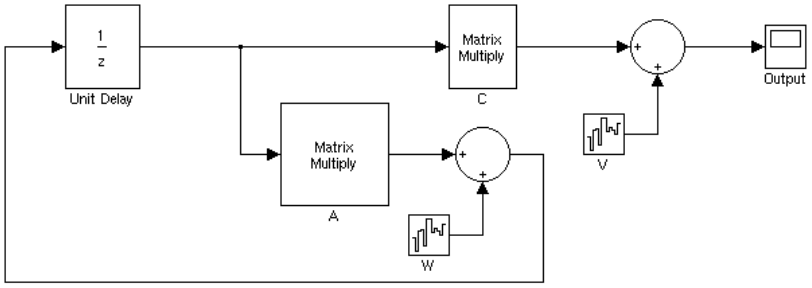

Figure 2-1. The block diagram of the linear dynamic system model

2.4 Transforming noise space without loss of generality

The assumption of zero-mean noise does not restrict the general linear model. In this section we explain why that is the case. This is conditional to the fact that we are allowed to change the structure of the linear system and the measurement device. In other words, if we were allowed to change $\mathbf{A}$ and $\mathbf{C}$ in Eq. (2-1) we could always add a $N+1^{st}$ dimension to the state vector, which is fixed at unity. Then adding an extra column to the right side of $\mathbf{A}$, holding the noise mean and an extra row of zero to the bottom of $\mathbf{A}$ (except unity at the bottom right corner) takes care of the non-zero mean for $\mathbf{w}_\bullet$. Similarly adding an extra column to C takes care of the

non-zero mean for $\mathbf{v}_\bullet$. Therefore, for example, both systems of Eq. (2-2) and Eq. (2-3), presented below are essentially the same.

$$\begin{bmatrix} x_{t+1}^1 \\ x_{t+1}^2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_t^1 \\ x_t^2 \end{bmatrix} + \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, Q_{2x2} \right) \tag{2-2}$$

$$\begin{bmatrix} x_{t+1}^1 \\ x_{t+1}^2 \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \mu_1 \\ a_{21} & a_{22} & \mu_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_t^1 \\ x_t^2 \\ 1 \end{bmatrix} + \begin{bmatrix} \mathcal{N}(0_{2x1}, Q_{2x2}) \\ 0 \end{bmatrix} \tag{2-3}$$

where in Eq. (2-3) the zero-mean gaussian noise is added only to the first two original states. Note that since the state evolution noise is gaussian and its dynamics is linear, $\mathbf{x}_t$ will be a first-order Gauss-Markov random process. Therefore the noise processes are essential elements of the stochastic model.

2.5 Probability Computation

There are several reasons that the Gaussian linear models are very popular amongst mathematicians and engineers. One of the reasons is the law of large numbers. But perhaps the most important reason from the viewpoint of an engineer is their computational tractability. This comes from two fortunate analytical properties of gaussian processes: First, the sum of two independent gaussian distributions is also gaussian.

$$\mathcal{N}(\mu_1, \Sigma_1) \cup \mathcal{N}(\mu_2, \Sigma_2) = \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2) \tag{2-4}$$

Second, the output of a linear transformation where the input of it is gaussian is also gaussian. This means that through the assumption of having the initial condition distributed Gaussian,

$$\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \mathbf{Q}_1) \tag{2-5}$$

we are guaranteed that all the future states, as well as observations, are also distributed Gaussian. In fact we can write explicit formulas for the conditional expectations of the states and observations [2]:

$$\begin{aligned} P(\mathbf{x}_{t+1} \mid \mathbf{x}_t) &= \mathcal{N}(\mathbf{A}\mathbf{x}_t, \mathbf{Q}) \mid \mathbf{x}_{t+1} \\ P(\mathbf{y}_t \mid \mathbf{x}_t) &= \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R}) \mid \mathbf{y}_t \end{aligned} \tag{2-6}$$

Moreover, because of the Markovian properties of the dynamic systems along with the gaussian assumption of noises and initial states, we can write the expression for the joint probability of a *sequence* of T states and observations:

$$P(\{\mathbf{x}_1, ..., \mathbf{x}_T\} \mid \{\mathbf{y}_1, ..., \mathbf{y}_T\}) = P(\mathbf{x}_1) \prod_{t=1}^{T-1} P(\mathbf{x}_{t+1} \mid \mathbf{x}_t) \prod_{t=1}^{T} P(\mathbf{y}_t \mid \mathbf{x}_t) \tag{2-7}$$

Again, this is caused by computational convenience scientists introduced the notion of *cost* in the conditional probability of Eq. (2-7) by taking the negative log of it. This can be represented as the sum of matrix quadratic forms:

$$\begin{aligned} &-2 \log P(\{\mathbf{x}_1, ..., \mathbf{x}_T\} \mid \{\mathbf{y}_1, ..., \mathbf{y}_T\}) = \\ &\sum_{t=1}^{T} [(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t)^T \mathbf{R}^{-1} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t) + \log |\mathbf{R}|] \\ &+ \sum_{t=1}^{T} [(\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)^T \mathbf{Q}^{-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t) + \log |\mathbf{Q}|] \\ &+ (\mathbf{x}_1 - \mu_1)^T \mathbf{Q}_1^{-1} (\mathbf{x}_1 - \mu_1) + \log |\mathbf{Q}_1| + T(M + N) \log 2\pi \end{aligned} \tag{2-8}$$

In the next chapter, where we talk about the derivation of HMM we will explain the details of *efficiently* calculating Eq. (2-7) and other associated probabilities.

2.6 Inference vs. System Identification

Historically, system identification (system ID) is the process of developing or improving a mathematical representation of a physical system using experimental data [34]. System identification community in engineering aims for providing effective and accurate analytical tools, which typically includes underlying methodologies, computational procedures as well as implementation of them. In engineering structures, there are three types of system identification: 1- model parameter identification 2- structure-model parameter identification and 3- control-model identification [34]. All three types of system identifications mentioned above are important areas in science and technology, where they have different principal objectives and histories.

One point that, in this section, we want to touch based on is the key differences between system ID and inference. The notions of "learning", "estimating", "filtering", and "smoothing" have been around for decades and we aim here to know their technical differences. Going back to our linear model of Eq. (2-1), with hidden states, lets consider different hypothetical scenarios: In some cases, we know exactly what the hidden states are supposed to be, and we just want to estimate them. For example, in a vision problem, the hidden states might be the location and orientation of an object, in which we want to estimate. In a tracking control problem, the hidden states maybe position and velocity and so on. In these cases we can often write down *a priori* observation or states evolution matrices based on our knowledge of the physics and environment of the problem. In these problems the emphasis is to accurately infer the unobserved (or even sometimes missing) information from the data we do have [2].

In other scenarios, we are aiming for coming up with explanations or causes for our data and have no explicit model, what so ever, for what these causes should be. Therefore, the observation sequence and state evolution process are mostly or even sometimes entirely unknown. The emphasis here, however, is to accurately learn about a few parameters that can model the observation sequence well enough (i.e. assign it a high likelihood) [2]. Speech modeling is a good example of such a situation [4]; say our goal is to find a feasible model that performs well for recognition tasks, but the particular values of hidden states in our models may not be meaningful or even important to us [2]. Another example of these problems are financial systems, where our goal is to find feasible models that predict the price levels, returns, or risks,

based on an observation sequence (financial time series), where the underlying hidden states (e.g. supply and demand, or support and resistance) might not necessarily correspond to the meaningful values, in which its not even important to us as long as the model can effectively estimate the price levels, returns, or risks.

These two goals, estimating parameters, typically manifest themselves in the solution of two distinct problems: inference and system identification. Lets explain each of them in details.

2.6.1 Inference: Filtering and Smoothing

The first point that we are trying to explain in this section is "smoothing". The corresponding question to answer in this manner (keeping the system of Eq. (2-1), and the initial conditions of Eq. (2-5) in mind) is: Given fixed model parameters $\{\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \mu_1, \mathbf{Q}_1\}$, what can be said about the "best" hidden state sequence, given an observation sequence? According to its application this question is typically made precise in literature in several ways. However, in all of different applications a very basic quantity that needs to be computed is the *total likelihood* of the observation sequence:

$$P(\{\mathbf{y}_1,\ldots,\mathbf{y}_T\}) = \int_{\text{all possible}\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}} P(\{\mathbf{x}_1,\ldots,\mathbf{x}_T\},\{\mathbf{y}_1,\ldots,\mathbf{y}_T\})d\{\mathbf{x}_1,\ldots,\mathbf{x}_T\} \quad \textbf{(2-9)}$$

This marginalization requires, of course, an efficient way of integrating or summing the joint probabilities over all possible paths through state-space. To illustrate the importance of having an efficient way of calculating this likelihood lets take a look at some numbers. If an ergotic system has N distinct states with a one dimensional output $\mathbf{y}$, and the length of observation sequence T, the calculation of the likelihood in Eq. (2-9) involves on the order of $2T \cdot N^T$ calculations, since at every $t = 1, 2, \ldots, T$ there are N possible states which can be reached (i.e. there are N^T possible state sequence), and for each such state sequence about $2T$ calculations are required (to be precise we need $(2T\text{-}1)N^T$ multiplications and $N^T\text{-}1$ additions) [4]. This calculation is computationally infeasible even with today's highly efficient and fast computers.

To clarify this, note that for a small number N and T; e.g. N=5 (number of hidden states) and T=100 (length of observation sequence), there are on the order of $2.(100).5^{100} \approx 10^{72}$ computations! Clearly a more efficient procedure is required to compute the total likelihood in Eq. (2-9). Fortunately such a procedure exists and is called forward-backward procedure [23, 24, 35]. The key to the forward-backward algorithm is to refurbish and recycle the calculation in closed forms. The details of the forward-backward procedure, in the context of HMM, will be explained in the next chapter under the title: "Baum-Welch algorithm". For now, let's assume that the total likelihood in Eq. (2-9) is available.

Once the integral of total likelihood is available, it is simple to compute the conditional distribution for any one proposed hidden state sequence given the observations by dividing the joint probability by the total likelihood:

$$P(\{\mathbf{x}_1,...,\mathbf{x}_T\}\,|\,\{\mathbf{y}_1,...,\mathbf{y}_T\}) = \frac{P(\{\mathbf{x}_1,...,\mathbf{x}_T\},\{\mathbf{y}_1,...,\mathbf{y}_T\})}{P(\{\mathbf{y}_1,...,\mathbf{y}_T\})} \qquad \textbf{(2-10)}$$

Often we are interested in the probability distribution of a particular hidden state at a particular time t. Going back to *filtering,* which was the first part of inference, we attempt to compute the conditional posterior probability,

$$P(\mathbf{x}_t\,|\,\{\mathbf{y}_1,...,\mathbf{y}_t\}) \qquad \textbf{(2-11)}$$

given all the observation *up to* and including time t.

In *smoothing*, the second part of inference, however we compute the conditional posterior probability, which is the distribution over $\mathbf{x}_t$,

$$P(\mathbf{x}_t\,|\,\{\mathbf{y}_1,...,\mathbf{y}_T\}) \qquad \textbf{(2-12)}$$

given the entire sequence of observation. We will revisit these calculations with much more details in later sections.

It is also possible to ask for the conditional expectation of hidden states given observations that extend only a few time steps in future (partial prediction) or on the other hand, a few time steps before the current time step (partial smoothing).

Filtering and smoothing have been extensively studied for continuous state dynamical systems in the signal processing community, starting from the pioneering work of Kalman and Rauch [3, 36-40], although this literature is often not known in the machine learning community [2]. For discrete-state models, however, much of the literature stems form the pioneering works of Baum and his colleagues on HMM, at the Institute for Defense Analysis (IDA) in Princeton, NJ [23-26, 35] and Viterbi (1967) [41] and others on speech recognition and optimal decoding. The book by Elliott *et al.* [42] contains a thorough mathematical treatment of filtering and smoothing for many general systems and models [2].

2.6.2 System Identification: Expectation-Maximization (EM)

The underlying idea of Expectation-Maximization (EM) is the same as Maximum Likelihood Estimations (MLE). In other words, in EM the idea is to choose the model parameters to maximize the total joint density of states and observation sequences, however, unlike MLE we don't know that joint density, simply since the states are hidden. Therefore, we maximize the current expectation of the joint density, given the observations and the current fit of parameters, and we iterate forward [1].

More generally, the second problem of interest with linear gaussian models as we mentioned before is the system identification problem that tries to answer the following question: given only an (or several) observation sequence $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, find the parameters $\{\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \mu_1, \mathbf{Q}_1\}$ that maximizes the total likelihood of observation sequence given in Eq. (2-9).

The learning problem for static models has been extensively studied in the neural network and Fuzzy logic community, and for dynamic models also it has been studied under HMMs and more general Bayesian belief networks. There is also a corresponding area of study in controls theory, known as system identification [34], which corresponds mostly (but not necessarily) to continuous state dynamical systems.

There are several approaches to system identification depends upon the area of study [43, 44] but in this work we focus on a system identification method based on Expectation-Maximization (EM) algorithm. The EM algorithm specifically for linear gaussian dynamical systems was derived by Shumway and Stoffer [45] in 1982 and then summarized [46] in 2000. It again reintroduced in the neural computation field by Ghahramani *et al.* [47, 48] and in the speech recognition community by Digalakis *et al.* [15]. Again we would like to mention that the book by Elliot *et al.* [42] describes this topic very well.

The basis of all learning content via EM algorithm was presented by two powerful articles by Baum, *et al.* [25] and Dempster *et al.* [49]. The objective of the algorithm is to maximize the total likelihood of observation sequence in the presence of hidden states. In this manner let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ denote the observation sequence and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, be the sequence of hidden variables as a function of the parameters of the model denoted by θ [2]. Maximizing the likelihood as a function of θ is equivalent to the maximizing the log-likelihood function:

$$\Im(\theta) = \log \Pr(\mathbf{Y} \mid \theta) = \log \int_{\mathbf{X}} \Pr(\mathbf{X}, \mathbf{Y} \mid \theta)\, d\mathbf{X} \tag{2-13}$$

Using *any* distribution Q over the hidden states, we can find the lower boundary on $\Im$:

$$\begin{aligned} \log \int_{\mathbf{X}} \Pr(\mathbf{X}, \mathbf{Y} \mid \theta)\, d\mathbf{X} &= \log \int_{\mathbf{X}} Q(\mathbf{X}) \frac{\Pr(\mathbf{X}, \mathbf{Y} \mid \theta)}{Q(\mathbf{X})}\, d\mathbf{X} \\ &\geq \int_{\mathbf{X}} Q(\mathbf{X}) \log \frac{\Pr(\mathbf{X}, \mathbf{Y} \mid \theta)}{Q(\mathbf{X})}\, d\mathbf{X} \\ &= \int_{\mathbf{X}} Q(\mathbf{X}) \log \Pr(\mathbf{X}, \mathbf{Y} \mid \theta)\, d\mathbf{X} - \int_{\mathbf{X}} Q(\mathbf{X}) \log Q(\mathbf{X})\, d\mathbf{X} \\ &= F(Q, \theta) \end{aligned} \tag{2-14}$$

Note that the middle inequality in Eq. (2-14) is known as Jensen's inequality and can be easily proved by concavity of the log function.

Let us define the global energy of the configuration (**X**, **Y**) to be:

$-\log \Pr(\mathbf{X},\mathbf{Y}|\theta)$

Also note that the lower bound $F(Q, \theta) \leq \Im(\theta)$ is the negative of a quantity known in statistical physics as the free energy: the expected energy under Q minus the entropy of Q [50]. The key to EM algorithm is to alternate between maximizing F with respect to Q and the parameters θ respectively holding the other fixed [2].

Starting from some initial parameter value θ_0 the EM algorithm can be divided into two steps **E**-step and **M**-step:

$$\begin{aligned} &\textbf{E}\text{-Step:} && Q_{k+1} \leftarrow \underset{Q}{\text{Argmax}}\, F(Q,\theta_k) \\ &\textbf{M}\text{-Step:} && \theta_{k+1} \leftarrow \underset{\theta}{\text{Argmax}}\, F(Q_{k+1},\theta) \end{aligned} \qquad \textbf{(2-15)}$$

Roweis and Ghahramani in [2] mentioned that the maximum in the **E**-step results when Q is exactly the conditional distribution of **X**, or in the other words:

$$Q_{k+1}(\mathbf{X}) = \Pr(\mathbf{X} \mid \mathbf{Y},\theta_k) \qquad \textbf{(2-16)}$$

at which point the lower bounds becomes and equality, where:

$$F(Q_{k+1},\theta_k) = \Im(\theta_k) \qquad \textbf{(2-17)}$$

The maximum in the **M**-step, however, is obtained by maximizing the first term in the third line of the Eq. (2-14). The reason is that in the free energy equation, the entropy of Q does not depend on θ. This yields to the (perhaps more familiar) equation for the **M**-step,

$$\textbf{M}\text{-Step:} \qquad \theta_{k+1} \leftarrow \underset{\theta}{\text{Argmax}} \int_{\mathbf{X}} \Pr(\mathbf{X} \mid \mathbf{Y},\theta_k) \log \Pr(\mathbf{X},\mathbf{Y} \mid \theta)\, d\mathbf{X} \qquad \textbf{(2-18)}$$

There is a term associated with the convergence of EM algorithm, called "hill climbing". This means that, after each iteration the likelihood increases or asymptotically stays the same. Intuitively this is because of the fact that at the beginning of each **M**-step, $F = \Im$ and since **E**-step does not change θ, we are guaranteed not to decrease the likelihood after each combined **EM**-step. This phenomena has been proven by Baum and his colleagues in [8, 35].

In the next chapter we explain the **EM** algorithm specifically in the context of HMM, and illustrate the HMM system ID in details.

3 Hidden Markov Models

We now return to the fully dynamic model introduced slightly differently than the linear model presented by Eq. (2-1). Our key observation is that the dynamics described by Eq. (2-1) in discrete states are exactly equivalent to the traditional discrete time, discrete state Markov chain dynamics using a state probability transition matrix and an observation emission matrix. Rabiner talks in depth about this model in [4]. It is easy to see how to relate state probability transition matrix to the matrices **A** and **Q** in Eq. (2-1). This is the standard setup for a dynamic, discrete-time, discrete-state modeling of Markov chains. Our approach, however, in this dissertation is to take continuous time series and model them with Markov switching models.

So, let's start with a quick literature review about the standard discrete-time HMM and then move on to the continuous-time HMM.

Hidden Markov Modeling is a probabilistic technique for the study of possibly stochastic time series [17, 18]. Modeling with many of the probability distributions, the cost of implementation is linear with respect to the length of the data and models can be nested to reflect hierarchical sources of knowledge [18]. Although initially introduced in the late-1960s and early-1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years [4]. Tracing back the work of Markov [10] and Shannon [19, 20] was concerned with Markov chains. While the state sequence is observed in

Markov chain [21], in hidden Markov model, the output properties impose a veil [22] between the state sequence and the observer of the time series. In the effort to lift the veil, a substantial body of theory was developed during the years of 1960s to 1990s. Leonard Baum, collaborating with Eagon and Petrie [23-25] dealt with finite probability spaces and addressed the problem of tractability of probability computation, iterative maximum likelihood estimation of model parameters from observed time series, recovery of hidden states, and the proof of consistency of the estimates [18].

The major development of the theory (1970) was the maximization technique of Baum, Petrie, Soules and Weiss [26]. This was after two articles (1966 and 1967) by Baum et al, where they unveil some statistical inference for probabilistic functions of finite state Markov chains as well as proving an equality with applications to statistical estimation for probabilistic functions [24, 25]. Baum and his colleagues including Lloyd Welch at the Institute for Defense Analysis (IDA), Princeton, NJ, made very important breakthroughs that lead to a wide range of theoretical outgrowths in this branch of science. They included a number of generalizations of both the spectral and temporal components of the model, e.g., variable-duration hidden Markov models [27], hidden-filter hidden Markov models [17], and trainable finite state hidden grammars [28]. A special case of the results in [26] has been addressed by Dempster et al, as the Expectation Maximization algorithm [29, 30].

Hidden Markov Models (HMMs) has been vastly used in automatic speech recognition (A.K.A. Natural Language Programming or NLP) [4, 5, 8-12, 51, 52]. Furthermore its applications were widespread from weather predictions to finance and modeling the stock market.

The basic theory of HMM was implemented for speech processing applications by Baker [8] at CMU, and by Jelinek and his colleagues at IBM [53, 54] in 1960s, 70s, and 80s. However widespread understanding and application of the theory of HMMs to speech processing has occurred during the 1980s [4].

As mentioned earlier, there are several reasons that engineers typically do not vastly use HMMs. First, the basic theory of hidden Markov models was published in mathematical journals, which were not generally read by engineers. The second reason was that the theory did not provide sufficient tutorial materials for most readers to understand the theory and to be able to apply it to their own research. Rabiner [4] had a breakthrough on this by publishing a tutorial on

hidden Markov models. Although the tutorial was written on the speech recognition applications, it is very easy to follow and has been used by engineers and scientists in different branches of science and technology [4].

3.1 Continuous-Time Hidden Markov Model

HMM tries to come up with the good understanding of how noisy data was generated, both spatially and temporally. It comes from a Markov chain corrupted with noise. This derivation of HMM assumes that the observation sequence has come from multiple sources, each with its own dynamics and corrupted with its own noise.

Determining which source at each time is generating the data follows a Markovian dynamics, as illustrated in Figure 3-1

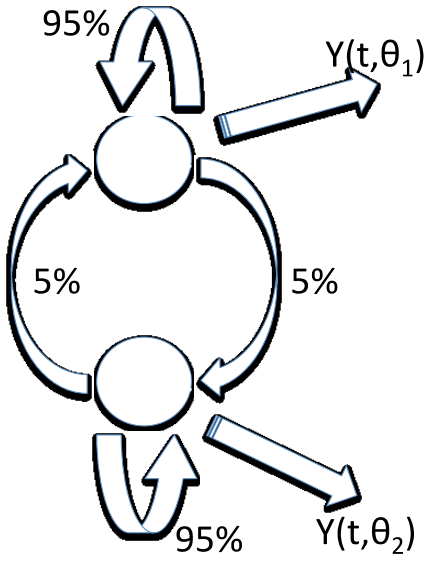

Figure 3-1: A two state Markov switching model

Therefore our Markov switching model can be written as:

$$\Pr(\mathbf{q}_{t+1}) = \mathbf{A}\Pr(\mathbf{q}_t)$$

$$\mathbf{y}_t = f(\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-k}, t, \Theta_j \mid q_t = S_j) + \mathbf{v}_t \quad : \mathbf{v}_t \sim \mathcal{N}(0, \Sigma_j),\ 1 \le j \le \mathcal{N}$$

$$\Rightarrow \mathbf{y}_t = \bar{\mathbf{y}}_t + \mathbf{v}_t \qquad \textbf{(3-1)}$$

where: $\bar{\mathbf{y}}_t = \mathrm{E}[\mathbf{y}_t] = \mathrm{E}[f(\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-k}, t, \Theta_j \mid q_t = S_j) + \mathbf{v}_t]$

$$= f(\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-k}, t, \Theta_j \mid q_t = S_j)$$

where $\mathbf{q}_t$ is the $N\times1$ state vector of $q_t = S_j,\ 1 \le j \le \mathcal{N}$ and Pr(.), when the argument is a vector, denotes the probability of all elements. Note that the observation sequence is an M-dimensional multivariate time series. Therefore Σ is an $M\times M$ covariance matrix that is different for each of the N state. Also Θ is the parameter set of the nonlinear emission model for each of the N state. Also looking at the Figure 3-1 and comparing it with Eq. (3-1), we can see that the sequence of q_t is the sequence of states which switches back and forth between the states 1 and 2 (in this case), with the denoted transition probabilities (e.g. 95% and 5%). Then being at each state, the multivariate observation at that time (i.e., $\mathbf{y}_t$), is generated by a k^{th}-order nonlinear dynamics with the parameter set Θ. At the end, a multivariate zero-mean Gaussian noise is added to the time series at each time.

This model is also known as "Regime Switching Model" or "Markov Switching Model". Note that in this dissertation, the "states" in the context of HMM are denoted by q_t, where in the context of system dynamics, the "states" are denoted by x_t. This is due to the fact that states in system dynamics typically correspond to a physical variable of the system, such as Position, velocity, or temperature. In HMM, however, the states are simply "regime," or sources that the observations are coming from. This is based on the continuous-time hidden Markov model that we are trying to study in this chapter.

Let's illustrate this concept a little bit further. Here a state transition probability matrix defines the Markov chain, while the output probability functions are defined as the observation probabilities or densities, or emission probabilities, i.e., $\Pr(\mathbf{y}_t)$. Figure 3-2 illustrates the concept of multivariate time series generated by the Markov switching model.

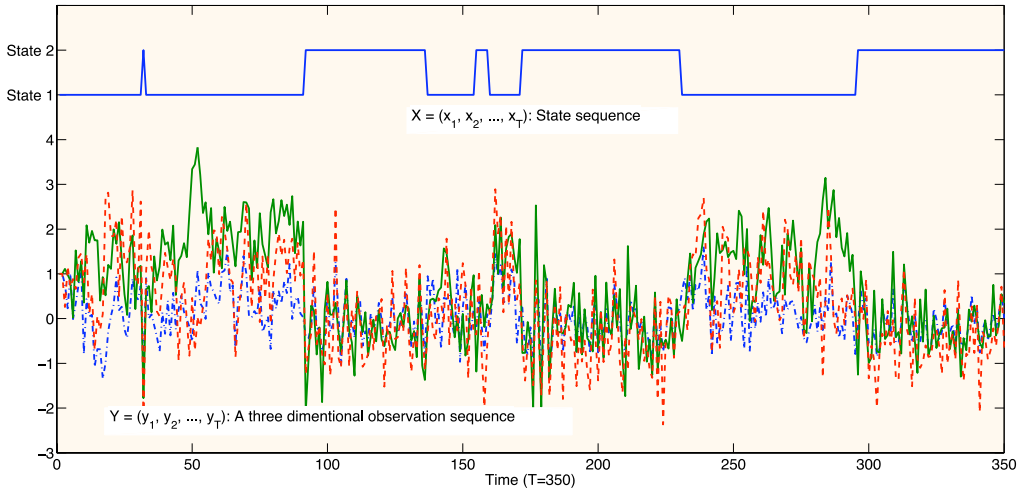

Figure 3-2: A 3D Time series of a Markov switching model with two states

Notice that according to the problem definition of our HMM, the only information we have from this system is the *M*-dimensional time series $\mathbf{Y}=(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$.

What we need to estimate as the output of our HMM system ID is

1- The parameter set Θ,

2- Covariance of the multivariate noise, for each state Σ,

3- The elements of transition Markov probabilities matrix **A**.

Also notice that since the nonlinear function "*f* " is unknown, the degree of nonlinearity of "*f* " as well as "*k* "; the order of "*f* " is unknown. Also the number of states (i.e. the size of **A)** is unknown.

To clarify the matter, let's mention that before starting to estimate the parameters, there are some parts of the structure that have to be assumed as *a priori*. The first is the number of states, i.e., size of **A**, however, sometime we can infer about this information from the physics of the problem. The second is an emission model for the nonlinear structure of "*f* " for each state. While this can be basically any function, a typical assumption is a linear multivariate Autoregressive model with Gaussian noise (MVN), in which with the sufficient number of states is able to model the observations fairly well. Shi in [1] used a Fuzzy based neural network as the

emission model and explained it for the univariate case. The last *a priori* information is the order of the generative emission model: " *k.* "

Now let's get down to the details of this derivation.

3.2 The Structure of MVN Continuous-Time HMM

Consider a system which maybe described at any time as being in one of the N distinct states; $S_1, S_2, ..., S_N$. Along the time, the system undergoes a change of state, according to a set of probabilities associated with the state. We denote the time index associated with the state changes as $t = 1, 2, \ldots$, and we denote the actual state at time t as q_t. We closely follow the notations of the tutorial by Rabiner [4], and a paper as well as the PhD dissertation by Shi [1]. A full probabilistic description of the above system would, in general, require specification of the current state, as well as all the previous states. However, the Markovian properties of the chain sequence allows us to draw the probabilistic distribution of the system, only based on the current state.

For this case of a discrete, first order, Markov chain, this truncated probabilistic transition can be written as:

$$\begin{aligned} &\Pr[q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...] \\ &= \Pr[q_t = S_j \mid q_{t-1} = S_i] \end{aligned} \qquad \textbf{(3-2)}$$

Furthermore, we assume that the right hand side of the Eq. (3-2) is independent of time. These assumptions lead us to the state transition probability matrix $\mathbf{A}=[a_{ij}]$, of the form

$$a_{ij} = \Pr[q_t = S_j \mid q_{t-1} = S_i], \qquad 1 \le i, j \le N \qquad \textbf{(3-3)}$$

where the matrix **A** has been already introduced in Eq. (3-1).

Since they obey the standard stochastic constraints the state transition probability elements have the properties (i.e. constraints):

$$a_{ij} \geq 0$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \textbf{(3-4)}$$

Up to this point, in this section, the stochastic process that we have introduced is called an observable Markov model, where each state corresponds to physical events. Also note that, this transition probability matrix has the same underlying concept of **A** matrix in Eq. (2-1), therefore to keep the notations consistent we named it also **A** matrix. There are two additional probability sets that we have to define in order to complete the introduction of HMM. The first probability is the emission probabilities distribution and the second is the initial state probability distributions.

Let's take a look at the emission probability distribution that is the probability of observing $\mathbf{y}_t$ given the state and the parameters. We define the emission probability distribution in state j, $\mathbf{B} = \{b_j(t)\}$, where j is the state and t is the time index. We have:

$$b_j(t) = \Pr[\mathbf{y}_t \mid q_t = S_j, \Theta], \qquad 1 \leq j \leq N, \quad 1 \leq t \leq T \qquad \textbf{(3-5)}$$

For a multivariate Gaussian emission that we have mentioned in Eq. (3-1) we can write the emission probabilities as:

$$b_j(t) = \Pr[\mathbf{y}_t \mid q_t = S_j, \Theta]$$

$$= \frac{\left|\Sigma_j\right|^{-1/2}}{(2\pi)^{-M/2}} \exp\{-\frac{1}{2}[\mathbf{y}_t - \bar{\mathbf{y}}_t(\Theta)]^T \Sigma_j^{-1}[\mathbf{y}_t - \bar{\mathbf{y}}_t(\Theta)]\} \qquad \textbf{(3-6)}$$

For the initial state probability distribution $\pi = \{\pi_i\}$ we have:

$$\pi_i = \Pr[q_1 = S_i], \qquad 1 \leq j \leq N \qquad \textbf{(3-7)}$$

Now the main idea of HMM is that given the appropriate values of **A**, Θ, **Σ** and π, the HMM can be used to generate the observation sequence **Y**. The elements **A**, Θ, **Σ** and π can be used as both a generator of observations, and as a model for how a given observation sequence was generated by an appropriate HMM. For notational convenience, we use the compact notation of $\lambda = (\mathbf{A}, \Theta, \mathbf{\Sigma}, \pi)$ to indicate the complete parameter set of the model.

Given the parameter set of HMM and the observation sequence, there are three basic problems of interest that must be solved for the model to be useful. The problems are the following:

Given the observation sequence $\mathbf{Y} = (\mathbf{y}_1 \mathbf{y}_2 ... \mathbf{y}_T)$*, and the model* $\lambda = (\mathbf{A}, \Theta, \mathbf{\Sigma}, \pi)$*, how do we efficiently compute* $\Pr(\mathbf{Y}|\lambda)$*, the probability of the observation sequence, given the model?*

The answer to this question lies within a procedure called the forward-backward method or Baum-Welch algorithm [24, 25]. Consider the forward variable

$$\alpha_t(i) = \Pr(\mathbf{y}_1 \mathbf{y}_2 ... \mathbf{y}_t, q_t = S_i \mid \lambda) \tag{3-8}$$

i.e. the probability of the partial observation sequence $(\mathbf{y}_1 \mathbf{y}_2 ... \mathbf{y}_t)$ (until time t) and state S_i at time t, given the model λ. There exists a closed form derivation for finding $\alpha_t(i)$, inductively [4]. We have:

Initializa tion :

$$\alpha_1(i) = \pi_i b_i(t), \quad 1 \le i \le N$$

Induction :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_{t+1}(i) a_{ij} \right] b_j(t+1), \quad 1 \le t \le T-1, \quad 1 \le j \le N \tag{3-9}$$

Terminatio n :

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)$$

In a similar manner we can construct a backward variable $\beta_t(i)$, defined as:

$$\beta_t(i) = P(\mathbf{y}_{t+1}\mathbf{y}_{t+2}...\mathbf{y}_T \mid q_t = S_i, \lambda) \quad \textbf{(3-10)}$$

That is the probability of partial observation sequence from t+1 to the end, being in state S_i at time t and given the model λ. Note that the backward probability calculation will have important applications on training and calibrating the parameter set λ. We can solve for $\beta_t(i)$ inductively.

Initializa tion :

$$\beta_T(i) = 1, \quad 1 \le i \le N$$

Induction : **(3-10-a)**

$$\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij}\, b_j(t+1)\beta_{t+1}(j)\right], \quad t = T-1, T-2, \ldots, 1$$

$$1 \le i \le N$$

Now, let's look at the second problem of interest in HMM.

Given the observation sequence $\mathbf{Y} = (\mathbf{y}_1\mathbf{y}_2...\mathbf{y}_T)$*, and the model* $\lambda = (\mathbf{A}, \Theta, \mathbf{\Sigma}, \pi)$*, how do we choose the corresponding state sequence* $Q = (q_1q_2...q_T)$*, which is optimal in some meaningful sense (i.e. best explains the observation sequence)?*

A formal technique for finding the single best state sequence exists based on Bellman's dynamic programming methods, and is called the Viterbi algorithm [41, 55]. To find the single best state sequence $Q=(q_1q_2...q_T)$, for a given observation sequence $\mathbf{Y} = (\mathbf{y}_1\mathbf{y}_2...\mathbf{y}_T)$, we need to define a quantity

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} \Pr(q_1, q_2, \ldots, q_t = S_i, \mathbf{y}_1\mathbf{y}_2...\mathbf{y}_t \mid \lambda) \quad \textbf{(3-11)}$$

i.e., $\delta_t(i)$ is the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state S_i. There exists a closed form derivation for finding $\delta_t(i)$,

inductively in Rabiner [4]. To actually retrieve the state sequence we need to keep track of the argument, which maximizes Eq. (3-11) for each time t, and i. Then according to the Bellman's method, by backtracking we can find the highest probability sequence of states. By induction we have:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}].b_j(t+1) \qquad \textbf{(3-12)}$$

As we mentioned we need to keep track of the argument, which maximizes the Eq. (3-11). We do this via an array $\Psi_t(j)$. The recursive procedure is as follows:

$$\begin{aligned}
&\text{Initialization:} \\
&\delta_t(i) = \pi_i b_i(1), \qquad 1 \le i \le N \\
&\psi_t(i) = 0 \\
&\text{Recursion:} \\
&\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}].b_j(t), \quad 2 \le t \le T, \quad 1 \le j \le N \\
&\psi_t(j) = \operatorname{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad 2 \le t \le T, \quad 1 \le j \le N \\
&\text{Termination:} \\
&P^* = \max_{1 \le i \le N} [\delta_T(i)] \\
&q_t^* = \operatorname{argmax}_{1 \le i \le N} [\delta_T(i)] \\
&\text{Backtracking Path:} \\
&q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1
\end{aligned} \qquad \textbf{(3-13)}$$

Now lets get to the third, and perhaps the most important question of HMM that one needs to answer:

How do we adjust the model parameters $\lambda = (\mathbf{A}, \Theta, \mathbf{\Sigma}, \pi)$ *to maximize* $\Pr(\mathbf{Y} \mid \lambda)$*?*

This problem, which by far is the most challenging one, is to determine a method to adjust the model parameters $\lambda = (\mathbf{A}, \Theta, \mathbf{\Sigma}, \pi)$, to maximize the probability of the observation sequence given by the model. There is no known way to analytically solve for the model to maximize the probability. There are, however, ways to calibrate the model to locally maximize $\Pr(\mathbf{Y} \mid \lambda)$ through an iterative procedure such as Baum-Welch method (or equivalently EM (Expectation Maximization) method) [49], or using gradient techniques [6].

3.3 Extended Baum-Welch Algorithm

Let us discuss one possible solution to problem III of the HMM design, i.e., calibrating the model parameters $\lambda = (\mathbf{A}, \Theta, \mathbf{\Sigma}, \pi)$. This method is primarily based on the classic work of Baum and his colleagues for choosing, and calibrating the parameters of HMM. For the standard Baum-Welch algorithm the emission model is simply the elements of an emission matrix versus our approach in which the emission parameters Θ and $\mathbf{\Sigma}$ are the parameters of the model, described in Eq. (3-1) and (3-6**)**. The discussions in this section can be considered as a special application of EM algorithm, studied in the earlier section (2.6.2) for hidden Markov models. In order to describe the procedure for re-estimation (iterative update and improvement) of HMM parameters, we first define the parameter $\xi_t(i,j)$ as the probability of being on two specific consecutive states, (i.e., the probability of being in state S_i at time t, and state S_j, at time t+1) , given the model and observation sequence;

$$\xi_t(i,j) = \Pr(q_t = S_i, q_{t+1} = S_j \mid \mathbf{Y}, \lambda) \qquad \textbf{(3-14)}$$

It is easy to note that from the definitions of forward and backward variables in Eq. (3-8) and (3-10-a) we can write Eq. (3-14) as,

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(t+1)\beta_{t+1}(j)}{\Pr(\mathbf{Y} \mid \lambda)}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(t+1)\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(t+1)\beta_{t+1}(j)} \quad \textbf{(3-15)}$$

where the numerator term is just $\Pr(q_t = S_i, q_{t+1} = S_j, \mathbf{Y} \mid \lambda)$ and the denominator is the total probability and also a normalizing factor. We define the variable $\gamma_t(i)$ as the probability of being in state S_i at time t, given the observation sequence and model,

$$\gamma_t(i) = \Pr(q_t = S_i \mid \mathbf{Y}) = \frac{\alpha_t(i)\beta_t(i)}{\Pr(\mathbf{Y} \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \quad \textbf{(3-16)}$$

Combining Eq. ((3-15) and ((3-16) yields

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \quad \textbf{(3-17)}$$

If we obtain the sum of $\gamma_t(i)$ over time t, we get a quantity that can be interpreted as the expected number of times that state S_i is visited or equivalently as the expected number of transitions made from state S_i. For this intuition we just need to exclude the time T from the summation. Similarly summation of $\xi_t(i,j)$ over t (again from t=1 to t=T-1) can be interpreted as the expected number of transitions from state S_i to S_j. That is,

$$\sum_{t=1}^{T-1}\gamma_t(i) = \text{expected number of transitions from } S_i$$

$$\sum_{t=1}^{T-1}\xi_t(i,j) = \text{expected number of transitions from } S_i \text{ to } S_j \quad \textbf{(3-18)}$$

Using the intuition above we can give a method of re-estimation of the transition probability matrix, and the initial probabilities. A set of reasonably re-estimated **A** and π are:

$$\hat{\pi}_i = \gamma_1(i) = \text{expected frequency in state } S_i \text{ at time } t = 1$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} = \frac{\text{expected number of transitions from } S_i \text{ to } S_j}{\text{expected number of transitions from } S_i} \quad \textbf{(3-19)}$$

where " ^ " denotes the re-estimation. In the original work by Baum et *al*. [26], a formula was shown to estimate the density without assuming emission model. In this work, the same way as Shi's work [1], assuming the parametric nonlinear emission model of Eq. ((3-1), re-estimation of the emission parameters Θ, **Σ**, comes directly from maximizing the total likelihood function, or minimizing the total error. However, shown by Fraser et *al.* in [56] we know that maximizing the Eq. (2-13) is equivalent to maximizing the following quantity:

$$\sum_{t=1}^{T}\sum_{j=1}^{N} \gamma_t(j) \log\Pr(\mathbf{y}_t \mid q_t = S_j, \Theta_j, \Sigma_j) \quad \textbf{(3-20)}$$

Where Θ_j, Σ_j are the parameters of the emission model of the state *j*. According to Shi [1], Eq. (3-20), could be seen as a cost function for the emission model, where the computation of the parameters Θ, Σ depends entirely on the, in general, nonlinear emission model of Eq. (3-1). Here, we assume the error to be gaussianly distributed in Eq. (3-6), and use a *k*-th order multivariate autoregressive emission model. Therefore Θ is the parameter set for the *k*+1 intercept and the coefficients of the autoregressive model and Σ is the covariance of the added noise. Θ can be estimated by minimizing the total locally weighted squared error function:

$$\hat{\Theta}_j = \text{Argmin}\{\sum_{t=1}^{T} \gamma_t(j)(\mathbf{y}_t - \bar{\mathbf{y}}_t(\Theta_j))^2\} \quad \textbf{(3-21)}$$

Note that the time-local weights of the errors are the probabilities of each state. Now after re-estimating the Θ_j we can simply calculate the re-estimation of Σ_j based on the newly estimated parameters via:

$$\hat{\Sigma}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)(\mathbf{y}_t - \bar{\mathbf{y}}_t(\hat{\Theta}_j))^2}{\sum_{t=1}^{T} \gamma_t(j)} \quad \textbf{(3-22)}$$

Note that the right hand side of the Eq. (3-22) guarantees the covariance matrices to be symmetric and positive definite.

If we define the set of estimated model parameters by $\hat{\lambda}$ then it has been proven by Baum and his colleagues that the model $\hat{\lambda}$ is almost always more likely than λ in the sense that

$$\Pr(\mathbf{Y} \mid \hat{\lambda}) \geq \Pr(\mathbf{Y} \mid \lambda) \quad \textbf{(3-23)}$$

Note that implementing HMMs are tricky and there are some issues associated with them. One of these issues is when the sequence of observation and therefore the sequence of states are long and multiplication of the resulting probability values could numerically converge to zero. Therefore even the most advanced computers, currently available, cannot easily compute the likelihood value of a long sequence.

The solution to this is using log-likelihood instead of likelihood function and also scaling the probabilities. The details of these methods could be found in [4]. Two other common issues that can be challenging are initial parameter estimation, and choice of model size and type.

3.3.1 Robust Parameter Estimation

In the previous section we studied the fact that Θ can be estimated by minimizing the total locally weighted squared error function Eq. (3-21). In that case we minimize the sum of weighted square errors between the emission model and the data, where the weighting functions

are the state probabilities. In this way since we are squaring the errors, the outliers in the data set, if any, will have a larger effect on the parameter estimation, as they should. In other words the estimated parameters are going to be largely skewed towards the outliers. Borrowing the notion of "Robust Regression", and using the similarity of this concept, we introduce the robust parameter estimation by:

$$\hat{\Theta}_j = \text{Argmin}\{\sum_{t=1}^{T} \gamma_t(j) \left| \mathbf{y}_t - \bar{\mathbf{y}}_t(\Theta_j) \right|\} \qquad \textbf{(3-24)}$$

Using Eq. (3-24), to estimate the parameters in the EM algorithm, will have a fewer tendency to be skewed towards the outliers in the data set. However, note that implementing Eq. (3-24) may not be straightforward, since its lack of differentiability and convexity faces a problem for numerical solvers. Working with a numerical software like Matlab, it is possible to use functions such as; "`fmincon`" for constraint optimization, "`fminunc`" for unconstraint optimization, or even more sophisticated optimization toolboxes e.g. `CVX` optimization. One, however, must make sure that the objective function has the correct format suitable for Matlab or any other software that is used.

Calculating the covariance matrix from the Eq. (3-22) is straightforward, upon robustly finding the intercept and those coefficients of the AR model via Eq. ((3-24).

3.4 A Closed-Form Solution to EM Algorithm

We aim for a closed-form solution to the re-estimation procedure of the emission model of EM algorithm. For this purpose, we take a slightly different approach for the emission model. As per Eq. (3-1), the emission probability densities for each hidden state of HMM can be modeled by a nonlinear dynamic equation. The linearized version of Eq. (3-1) can be a multivariate Autoregressive model, where the parameters to be estimated are the intercepts, the AR coefficients, and the covariance matrix of the added noise. To accomplish the task of finding a closed-form solution, we need to model the density of each of the states directly. In other words, we assume parametric distributions for the densities of each hidden state, where the goal

of the closed-form solution is to directly find the parameters of the densities. Note that the closed-form solution can only be found for the special case where the parametric densities are assumed to be elliptically symmetrical. Elliptical symmetry refers to the case where the projection of the density function for any pair of multivariate random variables is an ellipse and is symmetric about its vector mean. Gaussian distribution satisfies the elliptical symmetric conditions.

3.4.1 Features of the closed-form re-estimations

Here we assume an elliptically symmetrical parametric density for each state of HMM, where the closed-form solution to the local maxima of its convex likelihood function exists.

The ellipsoidal symmetric densities are assumed to have the form:

$$b_j(t) = \Pr[\mathbf{y}_t \mid q_t = S_j, \Sigma_j, \mu_j] = \left|\Sigma_j\right|^{-1/2} f_j\{(\mathbf{y}_t - \mu_j)^T \Sigma_j^{-1} (\mathbf{y}_t - \mu_j)\}$$
$$1 \le j \le N, \quad 1 \le t \le T \tag{3-25}$$

Note that the observation sequence is an M-dimensional multivariate time series. Therefore, Σ is an $M \times M$ covariance matrix, and μ is an M-dimensional vector of expected values that could be estimated separately for each of the N state. For the densities that the expected value cannot be defined such as multivariate Cauchy, μ_j is set to zero. Other variables, however, need to be estimated. This section specifically addresses the class of ellipsoidal symmetric densities whose expected value and covariance can in fact be defined.

We take the parameter set of the model as $\lambda = (\mathbf{A}, \mu, \mathbf{\Sigma}, \pi)$, where the re-estimation of the initial probabilities π, and the transitional probability matrix $\mathbf{A}$, could be conducted from Eq. (3-19). The aim of the closed-form solution for the re-estimation procedure is, therefore, to find the emission model parameters. In a standard continuous-time HMM, the emission parameters are estimated directly by maximizing the total likelihood function via Eq. (3-21). Thus, we take the parameter set to be $\lambda = (\mu, \mathbf{\Sigma})$.

The mathematical derivations of this section are strongly motivated by a theorem by Fan [57], which says if $|\Sigma|^{-1/2} f\{(\mathbf{y}-\mu)^T \Sigma^{-1}(\mathbf{y}-\mu)\}$ in our emission model of Eq. (3-25) satisfies the consistency conditions of Kolmogorov [58], then it also has the representation:

$$b(\cdot) = |\Sigma|^{-1/2} f\{(\mathbf{y}-\mu)^T \Sigma^{-1}(\mathbf{y}-\mu)\} = \int_0^\infty \mathcal{N}(\mathbf{y};\mu,v^2\Sigma)\, dG(v) \qquad \textbf{(3-26)}$$

for some probability distribution *G,* on [0,∞), where $\mathcal{N}(\mathbf{y};\mu,v^2\Sigma)$ is a multivariate Gaussian density with mean μ and the covariance matrix $v^2\Sigma$. Fan's theory essentially says that an elliptically symmetric density function can be represented as a continuous combination of related Gaussians.

We also have the total likelihood of the observation sequence $\mathbf{Y} = (\mathbf{y}_1\mathbf{y}_2...\mathbf{y}_T)$, conditional on a particular state sequence $Q=(q_1q_2...q_T)$, given the parameter set as:

$$\Im_\lambda(\mathbf{Y}) = \sum_{i=1}^{\mathcal{N}} \pi_i \prod_{t=2}^{T} a_{q_{t-1}q_t}.\ b_{q_t}(t)$$

where: **(3-27)**

$$(a_{q_{t-1}q_t} \mid q_{t-1} = S_i, q_t = S_j) \equiv a_{ij} \quad , 1 \le i, j \le \mathcal{N}$$

$$(b_{q_t}(t) \mid q_t = S_i) \equiv b_i(t)$$

Note that in this representation a_{ij} and $b_i(t)$ could be substituted from Eq. (3-3) and Eq. (3-5).

Combining E. (3-26) and Eq. (3-27) we have:

$$\Im_\lambda(\mathbf{Y},Q) = \int_0^\infty \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1}q_t}.\ \mathcal{N}(\mathbf{y}_t;\mu_{q_t},v_t^2\Sigma_{q_t})\, dG(v_1)...dG(v_T) \qquad \textbf{(3-28)}$$

where: $(\mu_{q_t},\Sigma_{q_t} \mid q_t = S_i) \equiv \mu_i,\Sigma_i \quad ,1 \le i \le \mathcal{N}$

Liporace in [27], however, shows that this integral can be represented as an average of likelihood functions for each state over the T-fold distribution of $G(v_1)G(v_2)....G(v_T)$. With the notation that "E_V" is the "expected value" over the space $V=(v_1,v_2,...,v_T)$, Eq. ((3-28) can be re-written as:

$$\Im_\lambda(\mathbf{Y},Q)=E_V[\Im_\lambda(\mathbf{Y},Q,V)]$$

where:

$$\Im_\lambda(\mathbf{Y},Q,V)\equiv \pi_{q_1}\prod_{t=2}^{T}a_{q_{t-1}q_t}.\,\mathcal{N}(\mathbf{y}_t;\mu_{q_t},v_t^2\Sigma_{q_t}) \qquad \textbf{(3-29)}$$

Based on the pioneering works of Baum et *al.* [23-26, 35] the re-estimation procedure requires an auxiliary function $\Omega(\lambda,\hat{\lambda})$ of the current parameter set λ and the re-estimated parameter set $\hat{\lambda}$.

$$\Omega(\lambda,\hat{\lambda})\equiv\sum_{Q}E_V[\Im_\lambda(\mathbf{Y},Q,V)\log\Im_{\hat{\lambda}}(\mathbf{Y},Q,V)] \qquad \textbf{(3-30)}$$

The utility of this auxiliary function has the property that increasing $\Omega(\lambda,\hat{\lambda})$ by re-estimation of the parameter set will monotonically increase the total likelihood function in Eq. (3-29).

Since in this section the distributions are assumed to be elliptically symmetric and, therefore, the existence of a local maxima is guaranteed over the convexity of the likelihood function, we can find the re-estimation of the parameter set by taking the M-dimensional derivatives of the $\Omega(\lambda,\hat{\lambda})$ function with respect to μ and Σ, and set it equal to zero. For re-estimation of the M-vector of multivariate μ we have:

$$\frac{\partial \Omega(\lambda, \hat{\lambda})}{\partial \mu_j} = 0 \qquad 1 \le j \le \mathcal{N}$$

$$\Rightarrow \sum_{Q} \mathrm{E}_V[\Im_\lambda(\mathbf{Y}, Q, V)] \sum_{t \in T_j(Q)} \left|\hat{\Sigma}_j\right|^{-1} (\mathbf{y}_t - \hat{\mu}_j) / v_t^2 = 0 \qquad \textbf{(3-31)}$$

where:

$$T_j(Q) = \{t : q_t = S_j\}$$

By interchanging the order of summation and multiplying both sides of the Eq. ((3-31) by $\left|\hat{\Sigma}_j\right|$, which $\left|\hat{\Sigma}_j\right|^{-1}$ is assumed to exist, we have:

$$\sum_{t=1}^{T} \sum_{q_t \in Q} \mathrm{E}_V[\Im_\lambda(\mathbf{Y}, Q, V) / v_t^2] \,.\, (\mathbf{y}_t - \hat{\mu}_j) = 0 \qquad \textbf{(3-32)}$$

Note that this equation was derived from Eq. (3-26), by replacing $b_j(t)$ with

$$\int_0^\infty v_t^{-2} \mathcal{N}(\mathbf{y}_t; \mu_j, v^2 \Sigma_j) \; dG(v_t) \qquad \textbf{(3-33)}$$

According to the Fan's theorem and Eq. (3-26), the Eq. (3-33) can simply be represented by:

$$-2 \frac{\partial \{b_j(t)\}}{\partial \{(\mathbf{y} - \mu_j)^T \Sigma_j^{-1} (\mathbf{y} - \mu_j)\}}\Bigg|_{\mathbf{y}=\mathbf{y}_t} \qquad \textbf{(3-34)}$$

Borrowing the forward and backward variables of Eqs. (3-8) and (3-10), we can solve for the re-estimation of μ_j:

$$\hat{\mu}_j = \left(\sum_{t=1}^{T} \varphi_t(j) \beta_t(j) \right)^{-1} \sum_{t=1}^{T} \varphi_t(j) \beta_t(j) . \mathbf{y}_t \qquad \textbf{(3-35)}$$

where:

$$\varphi_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \left[-2 \frac{\partial\{b_j(t)\}}{\partial\{(\mathbf{y}-\mu_j)^T \Sigma_j^{-1} (\mathbf{y}-\mu_j)\}} \bigg|_{\mathbf{y}=\mathbf{y}_t} \right] \tag{3-36}$$

Upon finding the expected values, for re-estimation of Σ_j, we obtain:

$$\begin{aligned} &\frac{\partial \Omega(\lambda, \hat{\lambda})}{\partial \Sigma_j} = 0 \qquad 1 \le j \le N \\ &\Rightarrow \hat{\Sigma}_j = \left(\sum_{t=1}^{T} \alpha_t(j) \beta_t(j) \right)^{-1} \cdot \sum_{t=1}^{T} \varphi_t(j) \beta_t(j) (\mathbf{y}_t - \mu_j)(\mathbf{y}_t - \mu_j)^T \end{aligned} \tag{3-37}$$

To efficiently implement these closed-form solutions, one can find the new variable we introduced in Eq. (3-36). We now take a look at two important special cases for elliptical symmetric densities.

3.4.1.1 Multivariate Gaussian

Let's start with:

$$b_j(t) = \left| \Sigma_j \right|^{-1/2} e^{\frac{-1}{2} (\mathbf{y}_t - \mu_j)^T \Sigma_j^{-1} (\mathbf{y}_t - \mu_j)} \tag{3-38}$$

where for this special case we have:

$$-2 \frac{\partial\{b_j(t)\}}{\partial\{(\mathbf{y}-\mu_j)^T \Sigma_j^{-1} (\mathbf{y}-\mu_j)\}} \bigg|_{\mathbf{y}=\mathbf{y}_t} = b_j(t) \tag{3-39}$$

and therefore:

$$\varphi_t(j) = \alpha_t(i) \tag{3-40}$$

3.4.1.2 Multivariate Cauchy

In this case we have:

$$b_j(t) = \left|\Sigma_j\right|^{-1/2} (\xi + \mathbf{y}_t^T \Sigma_j^{-1} \mathbf{y}_t)^{-(\delta + \frac{d}{2})} \tag{3-41}$$

where $\lambda = (\xi, \delta, d, \Sigma)$ is the parameter set to be re-estimated. For this special case we have:

$$\begin{aligned} &-2 \frac{\partial\{b_j(t)\}}{\partial\{(\mathbf{y} - \mu_j)^T \Sigma_j^{-1} (\mathbf{y} - \mu_j)\}}\Bigg|_{\mathbf{y}=\mathbf{y}_t} \\ &= 2(\delta + \frac{d}{2})\left|\Sigma_j\right|^{-1/2} . (\xi + (\mathbf{y}_t - \mu_j)^T \Sigma_j^{-1} (\mathbf{y}_t - \mu_j))^{-(\delta + \frac{d}{2} + 1)} \end{aligned} \tag{3-42}$$

Note that since the expected values cannot be defined for multivariate Cauchy densities, the derivations may need to be custom-tailored for these families of densities.

With this, we close this chapter and proceed to the next chapter for hidden Markov discussions.

4 Two-Step Hidden Markov Model

In this chapter we take a different approach in order to improve the "goodness of fit" of the model. In this regards, we estimate the probabilities of Markov switching with two separate definitions: *Transient Probabilities*, and *Steady Probabilities*. To best of our knowledge this is the first time that the transient and steady responses of probability estimations of HMM are studied with a 2-step HMM. We will first introduce a two-step HMM, where the first step uses a standard Viterbi. The second step combines state sequence and uses an extended Viterbi to model the transient and steady probabilities.

The first section introduces the first step standard HMM. Section 2 introduces the second step of the derivation based on a combined state sequence HMM. Next section introduces the steady versus transient probabilities. Section 3 gives an example to illustrate the theory, and conducts a comparison between the standard first-order HMM and our two-step derivation of HMM.

4.1 First Step: Standard HMM

Let $\boldsymbol{Q} = (q_1, q_2, \ldots, q_T)$ represent an N-state, t time-long Markov process, where q_t ($1 \leq t \leq T$) represents any of the N states. Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ represent an observation sequence, where $\mathbf{y}_t$ ($1 \leq t \leq T$) is a discretized measure of a continuous-time multivariate observation universe. A basic assumption of HMM holds where the observation is memoryless, i.e., for any t, the observation $\mathbf{y}_t$ depends only on the current state x_t.

Lets once again introduce the notation for HMM as adopted this chapter and the remainder of the document. Note that this notation is consistent throughout the dissertation, and we are repeating it here for the sake completeness.

Pr($\mathbf{Q}$): Total probability of the state sequence $\mathbf{Q}$;
Pr($\mathbf{Q}$,$\mathbf{Y}$): Joint probability of state sequence $\mathbf{Q}$ and the observation sequence $\mathbf{Y}$;
Pr($\mathbf{Q}$|$\mathbf{Y}$): Probability of state sequence $\mathbf{Q}$, conditional on the observation sequence $\mathbf{Y}$;
Pr($\mathbf{Y}$|$\mathbf{Q}$): Probability of observation sequence $\mathbf{Y}$, conditional on the state sequence $\mathbf{Q}$;
Pr(q_1): Initial state probability;
Pr($q_{t+1} \mid q_t$): State transition probability;
Pr($\mathbf{y}_t \mid q_t$): Probability that $\mathbf{y}_t$ is observed at time t, given the state q_t at the same time;

Our aim is to find a particular state sequence $\mathbf{Q}^*$, when the observation sequence $\mathbf{Y}$ is given, such that Pr($\mathbf{Q}^*$|$\mathbf{Y}$) is maximized. Solving this maximization problem is equivalent to solving the problem of maximizing Pr($\mathbf{Q}^*$,$\mathbf{Y}$)=Pr($\mathbf{Q}^*$|$\mathbf{Y}$).Pr($\mathbf{Y}$)**.** The solution to this problem via a standard Viterbi is:

$$\begin{aligned}\Pr(\mathbf{Q},\mathbf{Y}) &= \Pr(\mathbf{Q}).\Pr(\mathbf{Y}\mid\mathbf{Q}) \\ &= \Pr(q_1).\Pr(\mathbf{y}_1\mid q_1).\Pr(q_2\mid q_1).\Pr(\mathbf{y}_2\mid q_2)\ \ldots \\ &= \Pr(q_1).\Pr(\mathbf{y}_1\mid q_1).\prod_{t=2}^{T}\Pr(q_t\mid q_{t-1}).\Pr(\mathbf{y}_t\mid q_t)\end{aligned} \qquad \textbf{(4-1)}$$

Adopting the method, described in chapter 3, we can find a state sequence $\mathbf{Q}^* = (q_1^*, q_2^*, \ldots, q_T^*)$, which is basically the solution to the problem number 2 of HMM explained in Section 3.2. The main contribution of this chapter is to find another state sequence that has a larger maximum likelihood than the state sequence $\mathbf{Q}^*$.

4.2 Second Step: HMM with Combined State

Lets introduce a combined state sequence $\mathbf{\Delta Q} = (\Delta q_1, \Delta q_2, \ldots, \Delta q_{T-1})$, where:

$$\Delta q_t = q_t q_{t+1} \qquad 1 \le t \le k-1 \tag{4-2}$$

Note that $\mathbf{\Delta Q}$ is a N^2 state, $(T\text{-}1)$ time-long first-order HMM. For example, in the speech recognition sense, for the word "*seems*":

$\mathbf{Q} = (q_1, q_2, q_3, q_4, q_5) = (\text{s, e, e, m, s})$

$\mathbf{\Delta Q} = (\Delta q_1, \Delta q_2, \Delta q_3, \Delta q_4) = (\text{se, ee, em, ms})$

Therefore the N=3 states of $\mathbf{Q}$ versus N=9 states of $\mathbf{\Delta Q}$ for the HMM for a dictionary that only contains the 3 letters of the word "*seems*", i.e., *s*, *e*, *m* would be:

$$\mathbf{Q} = \begin{bmatrix} s \\ e \\ m \end{bmatrix} \Rightarrow \mathbf{\Delta Q} = \begin{bmatrix} ss & se & sm \\ es & ee & em \\ ms & me & mm \end{bmatrix} \tag{4-3}$$

The univariate observation sequence for the combined state HMM is $\mathbf{Z} = (z_1, z_2, \ldots, z_{k-1})$, where:

$$z_t = \Pr(q_{t+1}^* \mid \mathbf{Y}) - \Pr(q_t^* \mid \mathbf{Y}) \qquad 1 \le t \le k-1 \tag{4-4}$$

Note that the two elements of the RHS of the Eq. (4-4) are already estimated from the Eq. (3-16) of the standard HMM in the previous chapter.

The objective of the second step HMM in this section is to find the particular state sequence $\mathbf{\Delta Q}^*$ so that $\Pr(\mathbf{\Delta Q}^*|\mathbf{Z})$, or equivalently $\Pr(\mathbf{\Delta Q}^*,\mathbf{Z}) = \Pr(\mathbf{\Delta Q}^*|\mathbf{Z})$. $\Pr(\mathbf{Z})$ is maximized. We have:

$$\begin{aligned}\Pr(\Delta\mathbf{Q},\mathbf{Z}) &= \Pr(\Delta\mathbf{Q}).\Pr(\mathbf{Z}\,|\,\Delta\mathbf{Q}) \\ &= \Pr(\Delta q_1).\Pr(z_1\,|\,\Delta q_1).\Pr(\Delta q_2\,|\,\Delta q_1).\Pr(\Delta z_2\,|\,\Delta q_2)\ \ldots \\ &= \Pr(\Delta q_1).\Pr(z_1\,|\,\Delta q_1).\prod_{t=2}^{\mathcal{T}}\Pr(\Delta q_t\,|\,\Delta q_{t-1}).\Pr(z_t\,|\,\Delta q_t)\end{aligned} \qquad \textbf{(4-5)}$$

The state sequence Δq_t^* is defined as $q_t^* q_{t+1}^*$. Additionally keep in mind that although the states for this HMM may look like a second-order HMM, they are not. Note that each combination of Δq belongs to a separate state, since the second step HMM has N^2 states, e.g., Eq. (4-3). In other words each Δq_t can be considered as a single state, associated with a N^2 state, (k-1) time-long first-order HMM.

4.2.1 Steady vs. Transient Probabilities

Δq explains the steady state versus transient probabilities. If in $q_{t-1}^* q_t^*$ both consecutive parts (i.e. q_{t-1}^* and q_t^*) are the same, then it implies that there is more likelihood that the Markov state at the next time step stays on the same state. On the other hand if the two consecutive parts of $q_{t-1}^* q_t^*$ are not the same, then the Markov switching is on the transient from q_{t-1}^* to q_t^*. In the speech recognition sense, again for the word "*seems*":

$\mathbf{\Delta Q} = (\Delta q_1, \Delta q_2, \Delta q_3, \Delta q_4)$ = (se, ee, em, ms)

shows that the second state is a steady state of state "e", and every other state is transient.

Referring to the notation used in Eq. (4-3), one can relate the diagonal elements of $\mathbf{\Delta Q}$ matrix as steady probabilities and the off-diagonal elements as the transient probabilities. Introducing the steady and transient states enables to claim that if the state goes to a transient response it will soon end up on the second state of the two combined states.

We can now introduce our final single state sequence, which is always the second part of the combined state sequence. If the combined state sequence is steady, then the second part of it

is indeed the current state. If the combined state is transient, then the second part is the state in which the transition will end up. This state sequence estimation is in general ahead of time.

Note that, once again finding the $\mathbf{\Delta Q}^*$ is the solution to a separate standard HMM that was explained in Chapter 3. Lets introduce a single state sequence $\hat{\mathbf{Q}} = (\hat{q}_1, \hat{q}_2, ..., \hat{q}_T)$, so that:

$$\begin{aligned} &\hat{q}_{t+1} = q_{i+1} \quad \text{where:} \\ &q_{i+1} = \text{Argmax}\{\Pr(q_i^* q_{i+1}^* \mid \mathbf{Z})\} \quad 1 \le i \le k-1 \\ &\text{and } \hat{q}_1 = q_1 \end{aligned} \tag{4-6}$$

$\Pr(\Delta q_t^* | \mathbf{Z}) = \Pr(q_t^* q_{t+1}^* | \mathbf{Z})$ is already estimated as the output of our second step HMM via maximization of Eq. (4-5). The joint probability of states and observation sequence for the second step HMM can be noted as:

$$\begin{aligned} &\Pr(\hat{\mathbf{Q}}, \mathbf{Z}) = \\ &= \Pr(\hat{q}_1).\Pr(\hat{q}_2 \mid \hat{q}_1).\Pr(z_1 \mid \hat{q}_2) \ ... \\ &= \Pr(\hat{q}_1).\prod_{t=2}^{T} \Pr(\hat{q}_t \mid \hat{q}_{t-1}).\Pr(z_{t-1} \mid \hat{q}_t) \end{aligned} \tag{4-7}$$

Adding a z_0 with $\Pr(z_0)=1$ to the observation sequence of Eq. (4-7), we can linearly transform $\mathbf{Z} = (z_0, z_1, ..., z_{T-1})$, to $\hat{\mathbf{Z}} = (\hat{z}_1, \hat{z}_2, ..., \hat{z}_T)$ by shifting the time index forward by 1, yielding:

$$\begin{aligned} &\Pr(\hat{\mathbf{Q}}, \hat{\mathbf{Z}}) = \\ &= \Pr(\hat{q}_1).\prod_{t=2}^{T} \Pr(\hat{q}_t \mid \hat{q}_{t-1}).\Pr(\hat{z}_t \mid \hat{q}_t) \end{aligned} \tag{4-8}$$

Note that the state sequence of $\hat{\mathbf{Q}} = (\hat{q}_1, \hat{q}_2, ..., \hat{q}_T)$ is already calculated from Eq. (4-6), and will not be estimated from maximizing Eq. (4-7) nor from Eq. (4-8). Using Eq. (4-6), however, we can see that the characteristics of the states $\hat{q}$ are the same as states q. However, the

sequence of $\hat{\mathbf{Q}}$ is different than the sequence of $\mathbf{Q}$. Therefore, the transition probabilities of the two states different. In other words:

$$\begin{aligned} &\hat{q} \equiv q \\ &\Pr(\hat{q}_t \mid \hat{q}_{t-1}) \neq \Pr(q_t \mid q_{t-1}) \\ &\Pr(z_t \mid \hat{q}_t) \neq \Pr(y_t \mid q_t) \end{aligned} \qquad \textbf{(4-9)}$$

There are two differences between the standard HMM state sequence estimation of Eq. (4-1) and the extended HMM state sequence estimation in Eq. (4-8): First, the observation sequence and the transition probabilities are different. Second, the length of Eq. (4-8) is less than Eq. (4-1) by 1 probability measure because the first state at Eq. (4-8) doesn't have an observation associated with it.

Due to the differences between the transition and observation probabilities, we cannot directly compare the total maximized likelihood of our extended 2-step HMM of Eq. (4-8) with the standard HMM of Eq. (4-1). It is, however, for Eq. (4-8) to be greater than Eq. (4-1) because of its shorter length. Note that the probabilities are positive values and are less than 1, so shorter sequences have larger total likelihoods. To confirm this hypothesis, we conduct a series of experiments, as we will describe in the next section.

4.3 A 2-Step HMM Experiment

To evaluate and compare the extended two step HMM algorithm with the standard first-order HMMs, we conduct a series of experiments. The result of one of the experiments is given in this section. Figure 4-1 shows a 2 state (i.e. N=2) HMM with its associated time series of length T=1000$^{\text{sec}}$. Note that this time series is generated with an AR(1) difference model with different set of parameters at each state.

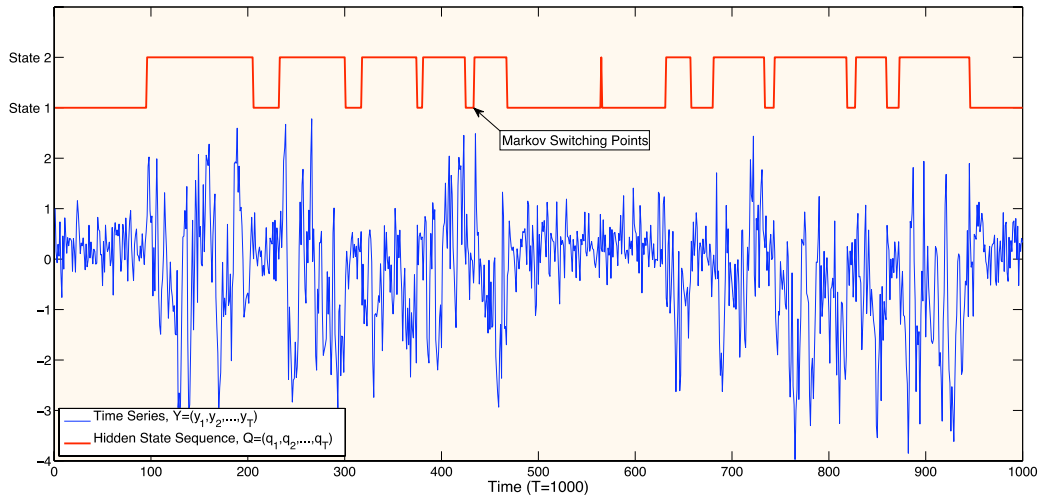

Figure 4-1: A sample 2-state HMM with its univariate time series

More specifically, the generative models at both states of Figure 4-1 are:

$$\begin{cases} y_{t+1} = +0.2 + 0.1\,y_t + \varepsilon_1 \;\; : \varepsilon_1 \sim \mathcal{N}(0,0.2), \;\; \text{if } x_t=1 \\ y_{t+1} = -0.2 + 0.7\,y_t + \varepsilon_1 \;\; : \varepsilon_1 \sim \mathcal{N}(0,0.9), \;\; \text{if } x_t=2 \end{cases} \qquad \textbf{(4-10)}$$

where: $y_1 = 0$

The sequence of hidden states **Q** is shown at the top of Figure 4-1. **Q** is essentially what we are trying to estimate, having **Y** as observation sequence. Figure 4-2 shows the hidden state sequence **Q** as a solid red curve on the top exhibit), and its estimate that is obtained from a standard one-step first-order HMM as a dotted blue. Our goal is to find a state sequence that has less error and higher total likelihood than the estimate described by the blue line.

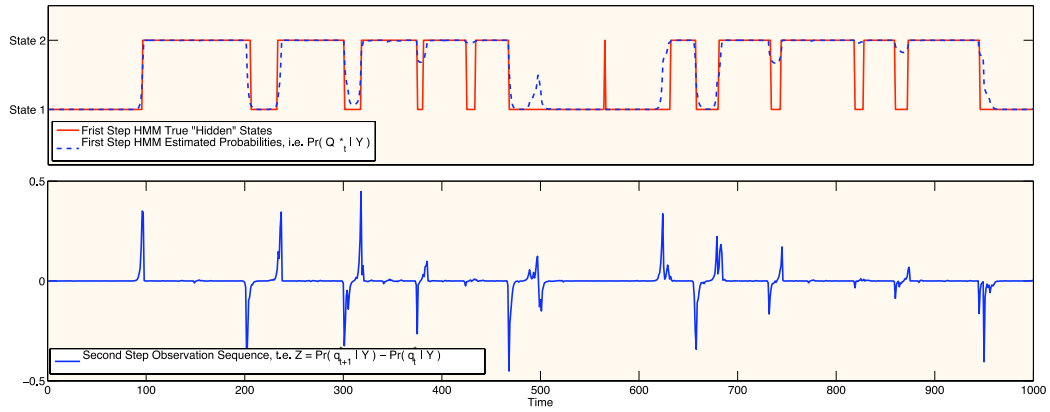

Figure 4-2: (a) The true 2-state sequence and the first step estimation of Markov probabilities, (b) The second step observation sequence, calculated as the difference of the first step Markov probabilities

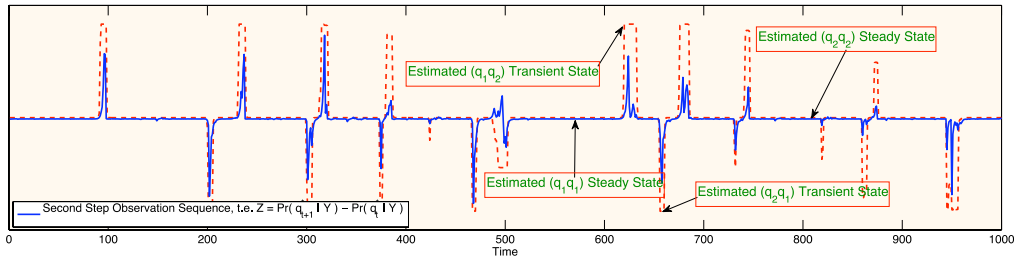

Figure 4-3: The second step estimated 4-states for the second step observation sequence

Figure 4-2-b shows the observation sequence for the second step HMM of our derivation, calculated from the estimated probabilities of the first step HMM in Eq. (4-4). The next step is the second step HMM that we run for this problem. Here we conduct a N^2=4 step HMM on the observation sequence in Figure 4-2-b.

Figure 4-3 shows the estimated four states of the second step of our HMM, using Eq. (4-5). Note that the hidden states for the second step are defined by Eq. (4-3).

After the second step is complete, from the dotted red curve of the Figure 4-3, we can calculate our extended Viterbi solution $\hat{\mathbf{Q}} = (\hat{q}_1, \hat{q}_2, ..., \hat{q}_T)$ via Eq. (4-6).

Figure 4-4-a, shows both the first step and the second step state probability estimations. The first step, shown by blue dotted curve is the standard Viterbi solution via a standard first orders HMM, whereas the green solid line is the extended Viterbi solution via a two-step HMM.

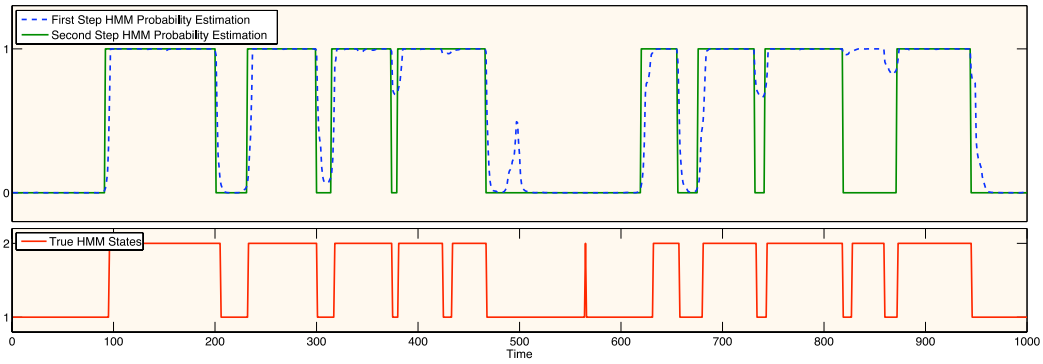

Figure 4-4 Top. (a) The first step versus second step state probability estimations. Bottom. (b) The true state estimate, shown for comparison purposes

It's worth noting that in Figure 4-4-a, the green solid line is still the estimated probabilities, not the state sequences. Figure 4-4-b is the state sequence. The fact that the second step probabilities are a lot cleaner than the first step estimation shouldn't be misunderstood. Figure 4-4-b shows the state sequence that we intended to estimate. This figure clearly shows the advantages of our 2-step state estimation method versus the standard HMM.

4.3.1 A Comparison

The extended Viterbi algorithm via a two-step HMM that we discussed in this chapter has one major advantage and one major weakness comparing to the standard one step HMM:

Weakness:

The computational complexity of our extended Viterbi is larger than the standard Viterbi. This is due to two reasons: First, the extended Viterbi runs the HMM algorithm twice. Second, the second step of our extended HMM runs on a N^2 state Markov chain. In general the computational complexity of the standard Viterbi (one-step HMM) versus the extended Viterbi (two-step HMM) is $O\{TN^2\}$ versus $O\{T(N+N^2)^2\}$, respectively, where T is the length of time series. The difference in the computational complexity is particularly noticeable for systems with larger number of states.

Advantage:

The main advantage of this method versus the standard Viterbi is its greater accuracy of estimation. As shown in Figure 4-4-a, the state estimation of the extended Viterbi is more accurate and cleaner than the standard Viterbi. Furthermore, there are multiple times that the standard Viterbi with a one step HMM misses the Markov switching points at time t=380, t=740, and t=830, whereas the 2-step HMM catches the Markov switching accurately and consistently.

With this method we are sacrificing some computational efficiency to get a more accurate state estimation. To quantify the advantages of the 2-step HMM method over the standard HMM,

a series of 30 experiments each with the length of k=1000^{sec} are conducted and the differences between the outputs of the two methods are observed. On the average the extended Viterbi method gives 29% improvement over the standard Viterbi. Along the series of experiments, a range of 14% to 41% improvement over the standard Viterbi was observed. The particular example in the previous section showed 21% improvement for the state estimation over the standard Viterbi algorithm.

5 Duration Dependent HMM

We will discuss a novel derivation of HMM that involves duration dependency of the state probabilities. This approach diverges from the Markovity assumptions to some extent, and derives the transition probabilities conditional on the joint density of what the current state is and how long the current state has been occupied. Because there have been a large number of studies related to this topic during the past decade, we will first review the past researches and then present our approach.

5.1 Background of the Duration Dependent HMM

Since the inception of HMM, scientists have tried to explore various ways of increasing the "goodness" of the fit of Markov models. Experimental evidence demonstrates that the inclusion of duration modeling can improve the training rate and ultimately the goodness of the fit [51, 59]. Vaseghi in [60] showed that in a data set of spoken English alphabet, durational modeling improves the recognition accuracy by 5.6%. There exists some arguments, however, against explicit-duration modeling since it increases the complexity of the HMM. An

improvement of 5.6% in magnitude of the state probabilities may not even change the output of the Viterbi state estimation [41, 55, 61]. Adding more complexity to the likelihood function of observations, conditioned on the state sequence may result in some inevitable errors during the maximization process as others have discovered [1]. Vaseghi introduces the duration-dependent transition probability matrix where the transition probability matrix in Eq. (3-3) becomes a three dimensional matrix with each entry a_{ij} having a different value for each discrete measure of duration d, on the 3rd dimension of the matrix [60]. At the end, fitting a polynomial to the 3rd dimension brings back the transition probability matrix into the two-dimensional space where each element a_{ij} now is a function of d; i.e. $a_{ij}(d)$. Durland and McCurdy also introduce the duration dependence directly into the transition probability matrix with the difference that the matrix elements follow an exponential kernel rather than a low order polynomial [62]. Lam extended the method by combining Hamilton [63] and Durland and McCurdy's approach [62] to incorporate the duration dependence in [64]. Hirokuni uses a Bayesian interface via MCMC and incorporates duration dependence for estimating the business cycles in Japan [65]. Finally Pelagatti uses Gibbs sampling for duration dependent Markov switching models and applied it to the US business cycles [66].

Russell and Cook address the property of the underlying model of state duration within the context of speech pattern modeling [59, 67]. To accomplish this, they presented an experimental evaluation of two extensions: Hidden semi-Markov models (HSMMs) and extended state HMMs (ESHMMs), where each state of HMM is modeled by a separate sub-HMM that outputs the pdf of the duration of that state [67]. The distributions considered in this research are Poisson and Gamma, and the method is theoretically extended to other distributions. All of these studies have improved the modeling of duration probability estimation.

Mitchell, et al. in [68] looked at the complexity of explicit duration HMMs. By introducing a new recursion method that significantly reduced the cost of training, as formulated to lower than other HMMs with duration modeling.

Burshtein introduced a robust parametric modeling of durations in HMMs [69]. He proposed a modified Viterbi algorithm in speech recognition in such a way that through incorporating both state and word duration modeling. He proved that the error rate in speech recognition could be reduced by 29% to 43% when compared to previous methods.

Djuric, et al. introduced an MCMC sampling approach for estimation of non-stationary HMMs [70]. They also considered a time dependent transition probability structure that indirectly models state durations by a probability mass function. More recently, Johnson addressed the capacity and complexity of HMM duration modeling techniques [71]. Johnson studied the standard and extended HMM methods with specific duration-focused approach.

Lately, during the past three years, from 2009 to the current date, several studies present HMM with duration dependency [72-78]. Also we would like to address two valuable literature reviews in duration-dependent HMMs, one by Ostendorf, et al. [79], and another one by Yu [80], which is a unified review of more recent works.

Reviewing more than 300 articles, published during the past two decades [80], we observe that most of these studies include the duration in the HMM structure for improving the goodness of fit of the model. Such an approach, however, increases the computational complexity. Many of these approaches can also estimate the expected duration of each state. All of these studies without exception, however, somehow include the duration of stay in the state transition probabilities, either directly or indirectly, implicitly or explicitly.

For out-of-sample applications, however, in which the current state self-loops during some time frame residency, the re-estimation of the state transition probabilities increases the probability of staying on that state. The reason for this is that no matter what method we use to re-estimate the state transition probabilities, at the end of the day, we are essentially counting the number of the time frames that HMM has been on that state, and divide it by the total number of time frames. Keeping this division in mind, a self-loop on any state, increases the numerator of such a division and thus, increases the fraction; i.e., the self-transition probability. As such, in an out-of-sample framework during the residency on any state the magnitude of the corresponding self-loop transition probability increases, and causes the probability of switching to other states to decreases. Some of the hidden semi Markov models (HSMMs) set the self-transition probabilities to zero, and assume that each state has multiple observations that correspond to the duration. These methods essentially have the same kind of problem in their emission model structure for out-of-sample applications.

According to Eq. (3-3) and (3-4) rows of the transition probability matrix always add up to 1, and therefore according to the Perron-Frobenius theorem [81, 82], this matrix always has at least one eigenvector with all N elements being 1, which corresponds to the eigenvalues of

magnitude 1. This fact makes the dynamics of HMM unstable, where the unstable modes correspond to the magnitude 1 eigenvalues. For such dynamics, during the short-term response (for small durations) the stable modes are dominant. In the long run (for long durations), however, the stable modes damp out and are overtaken by the unstable modes, leading to unreliable results. Clearly a better estimation of duration probability is needed.

We know that a good measure of the duration probability is the one that can increase the transition-to-other states' probability with passing time so that we can expect the state to switching [60, 62].

Solving this problem for out-of-sample frameworks is the main motivation of this chapter. There are two major differences between this research and previous studies:

> The main purpose of this research is not to increase the accuracy of estimation nor increasing the goodness of fit. The purpose of this research is to solve the dilemma of estimating the expected duration of stay on each state without using conventional methods.
>
> This study incorporates the duration of stay in the state probabilities independent of the state transition probabilities. In other words, we will have two different transition probabilities, 1- State transition probabilities and 2- Duration transition probabilities.

From the system dynamics standpoint, the previous studies include the duration in the open loop dynamics, whereas, we include the duration as a closed loop dynamic. Therefore, our derivation *controls* the state transitions to follow a certain duration distribution. Since the parameters of that duration distributions are unknown, we use a separate EM algorithm to estimate them. Therefore, the computational complexity of the EM algorithm for duration modeling is of the same size as the state modeling, which is significantly lower than the methods described in the past studies. For the past studies the computational complexity increases by the power of 2. Additionally this method estimates the distribution of the durations and thus, the probability of duration at each time step; as compared to the past studies that estimate the expected values and/or lower and upper bounds of the durations.

5.2 A Novel Derivation of Duration-Dependency

Let q_t be the actual hidden state at time t, and let d_t^i be the latest duration of stay on state q_t. Table 5-1 shows a possible realization for q_t and d_t^i to illustrate the concept.

Table 5-1: A possible realization of 2-state process states and corresponding durations

t	1	2	3	4	5	6	7	8	9
q_t	s_1	s_1	s_1	s_2	s_2	s_2	s_2	s_1	s_1
d_t^1	1	2	3	0	0	0	0	1	2
d_t^2	0	0	0	1	2	3	4	0	0

We are trying to build a model that the probability of q_{t+1} being on a certain state s_i depends on q_t and d_t^i independently. For each state, note that d_{t+1} is deterministic when d_t^i and s_{t+1} are given:

$$\begin{cases} d_{t+1}^i = d_t^i + 1 & if \quad q_{t+1} = q_t = s_i \\ d_{t+1}^i = 1 & if \quad q_{t+1} \neq q_t = s_i \\ d_t^i = 0 & if \quad q_t \neq s_i \end{cases} \tag{5-1}$$

This also shows that to find d_t^i we don't need to remember the sequence of $q_{1,\,2,\,\ldots,t}$. We only need to have the d_{t-1}^i and the q_t.

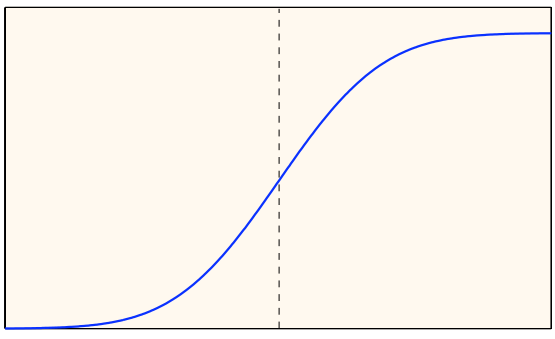

Figure 5-1: Kernel of a Sigmoid function

Table 5-2: Elements of the Duration Dependent Model

N	Number of distinct states
$\mathbf{A} = [a_{ij}]$	i,j: [1,N], a_{ij} are the elements of the state transition probability matrix
$\mathbf{B} = [d_{ij}]$	i,j: [1,N], d_{ij} are the elements of the duration transition probability matrix
s_i	i: [1,N], Distinct states
q_t	Actual state at time t
d_t^{i}	The consecutive duration of stay on state i, at time t
S_t^i	Probability of actual state at time t being on the state s_i : Pr(q_t=s_i)
$d_t^{\mathrm{i}} \quad E$	Excitatory d_t^{i} to switch to other states
$d_t^{\mathrm{i}} \quad I$	Inhibitory d_t^{i} to switch to other states
D_t^i	Probability of d_t^{i} being excitatory, given the state at time t : Pr(d_t^{i} E | q_t =s_i)
$\mathbf{S}_t$	(1 x N) Vector form of the probabilities S_t^i for i:[1,N]
$\mathbf{D}_t$	(1 x N) Vector form of the probabilities D_t^i for i:[1,N]
$\bar{d}^i$	Expected duration of stay on state i, given the Gaussian assumption
σ_{di}^2	Variance of estimated duration of stay on state i, given the Gaussian assumption
$\mathbf{y}_t$	Observation at time t
$\mathbf{Y}$	Observation sequence from time 1 to T
λ	Markov model parameter set
θ	Emission model parameter set
$\alpha_t(i)$	Extended Forward variable in the Baum-Welch method for joint densities of state and duration
$b_j(t)$	Emission probabilities of jth observation at time t
$\beta_t(i)$	Extended Backward variable in the Baum-Welch method for joint densities of state and duration
$\gamma_t(i)$	Probability of being on state i, while duration of stay at state i is excitatory at time t
$\xi_t(i,j)$	Probability of going to state j at time t+1, conditional on being on the state i, and having the duration of d_t^{i} at time t, given the parameter set of the Markov model and the observation sequence

Table 5-2 introduces the elements and notations of our duration derivation. We assume that the durations $d_t^i \mid q_t = s_i$ follow a certain parametric distribution for example Gaussian, Gamma, etc. Therefore, in a parametric framework we try to estimate the parameters of that distribution. For now to illustrate the theory let's assume that the distribution of the durations is Gaussian. Implementing this method for any other parametric distribution follows the same steps as the Gaussian distribution. We have:

$$d_t^i \mid q_t = s_i \sim \mathcal{N}(\bar{d}^i, \sigma_{di}^2) \tag{5-2}$$

Where the parameters to be estimated are:

$$\begin{aligned} \bar{d}^i &= \text{Mean}(\max(d_t^i)) \\ \sigma_{di}^2 &= \text{Var}(d_t^i \neq 0) \end{aligned} \tag{5-3}$$

Therefore, we have:

$$\Pr(d_t^i \mid q_t = s_i) = \frac{1}{\sqrt{2\pi}\sigma_{di}} e^{\frac{-(d_t^i - \bar{d}^i)^2}{2\sigma_{di}^2}} \tag{5-4}$$

We further truncate the distribution of Eq. (5-4), or any other possible form of $d_t^i \mid q_t = s_i$ to allow only for positive values of d_t^i. The duration of zero timeframe might correspond to a non-zero probability. This means that by our definition $\Pr(d_t^i | q_t \neq s_i)$ could be also non-zero. Thus, the cumulative density function (cdf) of the duration of stay for each state follows a sigmoid kernel of Figure 5-1. Note that if the distribution is non-Gaussian, the cdf function will be a skewed version of the sigmoid function. With this notation, as time passes, if the current state dose not switch, the cumulative probability of d_t^i goes up on a sigmoidal trajectory. The dotted line in Figure 5-1 denotes the inflection point of the sigmoid function that corresponds to the $\Pr(d_t^i | q_t = s_i) = 0.5$, and is in the middle of the distribution for the Gaussian case. This feature makes the dynamics of probabilities similar to the excitatory (E) and inhibitory (I) interactions in

a localized population of neuron synapses. Wilson and Cowan showed that the excitation and inhibition of neurons follows a sigmoidal kernel, in such a way that if the excitatory neurons are more than 50% of the population, the localized neuron population fires [33]. Borrowing that notation and applying it to the Eq. ((5-4), by definition we have:

$$\begin{aligned} &\text{If} \quad \Pr(d_t^i \mid q_t = s_i) > 0.5 \quad \text{then} \quad d_t^i : \text{Excitatory } (E) \\ &\text{If} \quad \Pr(d_t^i \mid q_t = s_i) < 0.5 \quad \text{then} \quad d_t^i : \text{Inhibitory } (I) \\ &\Rightarrow D_t^i = \Pr(d_t^i \equiv E \mid q_t = s_i) = 1 - \Pr(d_t^i \equiv I \mid q_t = s_i) \end{aligned} \tag{5-5}$$

Which means that if the $\Pr(d_t^i \mid q_t = s_i) > 0.5$, say $\Pr(d_t^i \mid q_t = s_i) = 0.6$, then there is a 60% probability that the q_{t+1} would switch to another state, conditional on its current duration. This is so far the Markov switching conditional ONLY on the duration of stay and independent of the current state (i.e., current duration status, "*E*" or "*I*"). In other words as time passes, $\Pr(d_t^i \mid q_t = s_i)$ follows the sigmoidal kernel of Figure 5-1, and therefore until it hits the inflection point there is more probability of "stay" and after the inflection point there is more probability of "switch". The only difference here with the original localized neural population is that in neural system when the population of "*E*" neurons hits the inflection point, the total population fires deterministically. In our derivation, however, while the d_t^i *I*, (i.e. $\Pr(d_t^i \mid q_t = s_i) < 0.5$, and the probability trajectory has not reached the inflection point yet, the probability of "Stay" is increasing and at the inflection point the probability of "Stay" is maximized. After that the probability of d_t^i *I* increases and therefore the probability of d_t^i *E* decreases. Therefore at the inflection point there is no guarantee that the regime will switch deterministically. The probability of regime switching conditional ONLY on the duration of stay is, however, maximized. With the notation that $S_t^i = \Pr(q_t = s_i)$ and $D_t^i = \Pr(d_t^i \equiv E)$ in our model we are trying to derive S_{t+1}^i conditional on the joint density of S_t^i and D_t^i. Note that based on Markov properties of HMM S_t^i is independent from D_t^i.

Now lets derive the Markov switching probabilities (i.e. $S_{t+1}^i = \Pr(q_{t+1} = s_i)$), conditional on the joint densities of current state (i.e. $S_t^i = \Pr(q_t = s_i)$) and current duration status (i.e. $D_t^i = \Pr(d_t^i \equiv E)$). We have:

$$\begin{aligned}\Pr(q_{t+1} = s_j) &= \sum_{i=1}^{\mathcal{N}} \Pr(q_t = s_i, d_t^i \equiv E) \Pr(q_{t+1} = s_j \mid q_t = s_i, d_t^i \equiv E) \\ &= \sum_{i=1}^{\mathcal{N}} \Pr(q_t = s_i) \Pr(d_t^i \equiv E \mid q_t = s_i) \Pr(q_{t+1} = s_j \mid q_t = s_i, d_t^i \equiv E) \\ &= \sum_{i=1}^{\mathcal{N}} \Pr(q_t = s_i) \Pr(q_{t+1} = s_j \mid q_t = s_i) \qquad \text{for } j = 1:\mathcal{N}\end{aligned} \tag{5-6}$$

On the other hand we have:

$$\begin{aligned}\Pr(q_{t+1} = s_j) &= \sum_{i=1}^{\mathcal{N}} \Pr(q_t = s_i, d_t^i \equiv E) \Pr(q_{t+1} = s_j \mid q_t = s_i, d_t^i \equiv E) \\ &= \sum_{i=1}^{\mathcal{N}} \Pr(d_t^i \equiv E) \Pr(q_t = s_i \mid d_t^i \equiv E) \Pr(q_{t+1} = s_j \mid q_t = s_i, d_t^i \equiv E) \\ &= \sum_{i=1}^{\mathcal{N}} \Pr(d_t^i \equiv E) \Pr(q_{t+1} = s_j \mid d_t^i \equiv E) \qquad \text{for } j = 1:\mathcal{N}\end{aligned} \tag{5-7}$$

Combining Eqs. (5-6) and (5-7), we have:

$$\Pr(q_{t+1} = s_j) = \sum_{i=1}^{\mathcal{N}} \frac{1}{2} \left\{ \Pr(d_t^i \equiv E) \Pr(q_{t+1} = s_j \mid d_t^i \equiv E) + \Pr(q_t = s_i) \Pr(q_{t+1} = s_j \mid q_t = s_i) \right\} \tag{5-8}$$

In the matrix format of the probability space for say, a 2-state HMM we will have:

$$\begin{aligned}&\mathbf{S_{t+1}} = \frac{1}{2}\left\{\mathbf{S_t}.\mathbf{A} + \mathbf{D_t}.\mathbf{B}\right\} \\ &\Rightarrow \begin{bmatrix} S_{t+1}^1 & S_{t+1}^2 \end{bmatrix} = \frac{1}{2}\left\{ \begin{bmatrix} S_t^1 & S_t^2 \end{bmatrix} . \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} D_t^1 & D_t^2 \end{bmatrix} . \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \right\}\end{aligned} \tag{5-9}$$

where a_{ij} is defined by Eq. (3-3), and consequently d_{ij} can be defined as:

$$d_{ij} = \Pr(q_{t+1} = s_j \mid d_t^i \equiv E) \tag{5-10}$$

Eq. (5-9) from the statistical and also system dynamics standpoint has intuitive meanings. From statistical standpoint, it means that the probability of Markov switching comes from two sources:

What state we are currently on,

The duration of stay on that state.

At the end, Eq. (5-9) takes the average of the probabilities of two sources as the total probability. From the system dynamics standpoint, however, Eq. (5-9) corresponds to a closed-loop version of a traditional HMM. In a traditional HMM the first part of the right hand side of Eq. (5-9) is all it matters since the probability of Markov switching is only conditioned on what state we are currently at. The second part of the right hand side of Eq. (5-9), however, introduces a feedback from the history of Markov switching, and basically *controls* the HMM probabilities so that the duration of stay on each state follows a distribution denoted by Eq. (5-4). Note that Eqs. (5-2) and (5-4) are based on the assumption that the parametric structure of the distributions of durations are known a priori, (e.g., they are Gaussian) and the values of the parameters have to be estimated. The Gaussian assumption is particularly a good assumption if we have enough regime switching in our history so that the law of large numbers applies. The excitatory and inhibitory probabilities of duration, however, can be calculated based on any distributions. For instance, one may assume that the durations have a Gamma distribution, being skewed towards the more recent history.

Now for implementation of this method we have to derive the re-estimation probabilities and inductions. Based on the EM algorithm to maximize the total likelihood of the observation sequence, conditional on the joint density of the states and durations we can derive the re-estimation procedure [4]. First in the **E-Step** we derive our Extended Forward-Backward procedure. Consider the forward variable $\alpha_t(i)$ as:

$$\alpha_t(i) = \Pr(\mathbf{y}_1\mathbf{y}_2...\mathbf{y}_t,\ q_t = s_i, d_t^i \equiv E \mid \lambda) \tag{5-11}$$

where λ is the entire Markov model parameter set. Therefore, based on Eq. (5-8) we can solve for $\alpha_t(i)$ inductively according to:

$$\begin{aligned}&\text{Initializa tion :}\\&\quad \alpha_1(i) = \pi_i b_i(1) \;\; : \;\; 1 < i < N\\&\text{Induction :}\\&\quad \alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \frac{1}{2}(\alpha_t(i)a_{ij} + D_t^i d_{ij})\right] b_j(t+1)\end{aligned} \qquad \textbf{(5-12)}$$

where $b_j(t+1)$ is the emission probability, based on the emission model. For instance, one can use an Autoregressive model AR(p) for emission probabilities with parameter set θ. The details of this emission model were discussed earlier. Now consider the backward variable $\beta_t(i)$ defined as:

$$\beta_t(i) = \Pr(\mathbf{y}_{t+1}\mathbf{y}_{t+2}\ldots\mathbf{y}_T \mid q_t = s_i, d_t^i \equiv E, \lambda) \qquad \textbf{(5-13)}$$

Again based on Eq. (5-8) we can solve for $\beta_t(i)$ inductively, but backwards this time, as follows:

$$\begin{aligned}&\text{Initialization:}\\&\quad \beta_T(i) = 1 \quad : \quad 1 < i < N\\&\text{Induction:}\\&\quad \beta_t(i) = \left[\sum_{j=1}^{N} \frac{1}{2}(a_{ij}\beta_{t+1}(j) + d_{ij}D_{t+1}^j)\right] b_j(t+1)\end{aligned} \qquad \textbf{(5-14)}$$

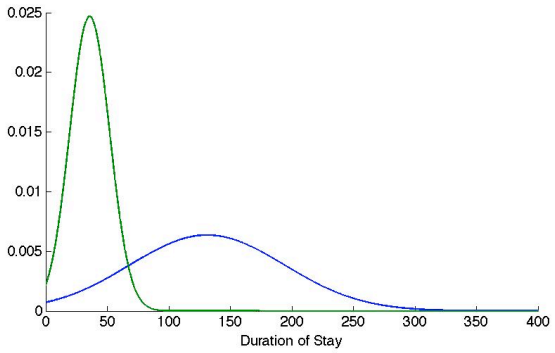

Figure 5-2: The truncated distribution of the duration of stay at each regime for a 2-state HMM

Note that in our extended forward-backward method, the induction is over the α and β probabilities only, but not the D probabilities. This is due to the fact that, while S_t^i and D_t^i are independent random variables, given d_t^i and q_{t+1}, d_{t+1} is deterministic from Eq. (5-1).

Now, the probabilities of being on each state while the duration probability is excitatory at each time can be denoted by the variable $\gamma_t(i)$, which is defined and calculated as follows:

$$\gamma_t(i) = \Pr(q_t = s_i, d_t^i \equiv E \mid \mathbf{Y}, \lambda)$$
$$\Rightarrow \gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{\mathcal{N}} \alpha_t(i)\beta_t(i)} \qquad 1 \le i \le \mathcal{N} \tag{5-15}$$

Also the state probabilities could be calculated from the Eq. (5-15) and Eq. (5-4), and (5-5). We have:

$$\Pr(q_t = s_i) = \frac{\Pr(q_t = s_i, d_t^i \equiv E \mid \mathbf{Y}, \lambda)}{\Pr(d_t^i \equiv E \mid q_t = s_i, \mathbf{Y}, \lambda)}$$
$$\Rightarrow S_t^i = \frac{\gamma_t(i)}{D_t^i} \qquad 1 \le i \le \mathcal{N} \tag{5-16}$$

Remember that $\mathbf{D}_t$ is re-calculated from Eq. (5-2) to (5-5) at the beginning of each E-Step. Also to implement the Eq. (5-11) to (5-16) effectively, we need to avoid numerical

instability, which corresponds to working with log-probabilities and a procedure called, scaling. The details of scaling and log-probabilities are included in [4, 6].

Now in order to fully describe the re-estimation process we need to define one more probability, which is: $\xi_t(i,j)$. Here we define it as follows:

$$\xi_t(i,j) = \Pr(q_t = S_i, d_t^i \equiv E, q_{t+1} = S_j \mid \mathbf{Y}, \lambda) \tag{5-17}$$

Thus from the Eq. ((5-8) we can calculate $\xi_t(i,j)$ as follows:

$$\xi_t(i,j) = \frac{[\frac{1}{2}(\alpha_t(i)a_{ij} + D_t^i d_{ij})]b_j(t+1)\beta_{t+1}(j)}{\sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{\mathcal{N}}[\frac{1}{2}(\alpha_t(i)a_{ij} + D_t^i d_{ij})]b_j(t+1)\beta_{t+1}(j)} \tag{5-18}$$

The E-Step of the EM algorithm is essentially calculating the probabilities of Eq. (5-11) to Eq. (5-18). Now we can reconstruct the transition probabilities and re-estimate the model parameters in the **M-Step**. For the state and duration transition probabilities we have:

$$\Pr(q_{t+1} = s_j \mid q_t = s_i, d_t^i \equiv E) = \frac{\Pr(q_t = s_i, d_t^i \equiv E, q_{t+1} = s_j \mid \mathbf{Y}, \lambda)}{\Pr(q_t = s_i, d_t^i \equiv E \mid \mathbf{Y}, \lambda)} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{5-19}$$

Also, according to Eq. (5-7) for the re-estimation of the state transition probabilities, $\hat{a}_{ij}$, we have:

$$\Pr(q_{t+1}=s_j \mid q_t=s_i)=\frac{\Pr(q_{t+1}=s_j,q_t=s_i)}{\Pr(q_t=s_i)}$$

$$=\frac{\sum_{d_t^i:\ i=1:\mathcal{N}}\Pr(q_{t+1}=s_j,q_t=s_i,d_t^i\equiv E)}{\sum_{d_t^i:\ i=1:\mathcal{N}}\Pr(q_t=s_i,d_t^i\equiv E)} \qquad \textbf{(5-20)}$$

$$\Rightarrow \hat{a}_{ij}=\frac{\sum_{t=1}^{\mathcal{T}-1}\{\sum_{i=1}^{\mathcal{N}}\xi_t(i,j).D_t^i\}}{\sum_{t=1}^{\mathcal{T}-1}\{\sum_{i=1}^{\mathcal{N}}\gamma_t(i).D_t^i\}}$$

And for the re-estimation of the duration transition probabilities, $\hat{d}_{ij}$, we also have:

$$\Pr(q_{t+1}=s_j \mid d_t^i\equiv E)=\frac{\Pr(q_{t+1}=s_j,d_t^i\equiv E)}{\Pr(d_t^i\equiv E)}$$

$$=\frac{\sum_{S_t^i:\ i=1:\mathcal{N}}\Pr(q_{t+1}=s_j,q_t=s_i,d_t^i\equiv E)}{\sum_{S_t^i:\ i=1:\mathcal{N}}\Pr(q_t=s_i,d_t^i\equiv E)} \qquad \textbf{(5-21)}$$

$$\Rightarrow \hat{d}_{ij}=\frac{\sum_{t=1}^{\mathcal{T}-1}\{\sum_{i=1}^{\mathcal{N}}\xi_t(i,j).S_t^i\}}{\sum_{t=1}^{\mathcal{T}-1}\{\sum_{i=1}^{\mathcal{N}}\gamma_t(i).S_t^i\}}$$

Re-estimation of the emission probability parameters, θ, is through maximization of the total likelihood of the observation sequence conditioned on the states and durations joint densities.

$$\hat{\theta}=\text{Argmax}\{\sum_{t=1}^{\mathcal{T}}\sum_{j=1}^{\mathcal{N}}\gamma_t(j)\log\Pr[\mathbf{Y}\mid q_t=s_j,d_t^j\equiv E,\theta_j]\} \qquad \textbf{(5-22)}$$

When the EM algorithm is complete with the method that was explained earlier, there are two sets of variables that are available, which didn't exist in the traditional HMM. The

first set of variables is the parameters of the duration distributions for each hidden state. With the Gaussian assumption of Eq. (5-2) to (5-4) the parameters are the expected value and the variance of the duration for each state, i.e. $\overline{d}^{\,i}$ and σ_{di}^2. The second set of variables in hand is a time series of the cumulative probabilities of duration of stay for each regime, along the observation sequence, i.e. D_t^i. These variables resembles the sigmoid function for each regime with the expected duration of stay for each regime; $\overline{d}^{\,i}$, happens at the probability of 0.5, where the inhibitory probabilities, i.e. the probabilities of "stay" are maximized and after that the excitatory probabilities increases, i.e. the probability of "switch" to other states.

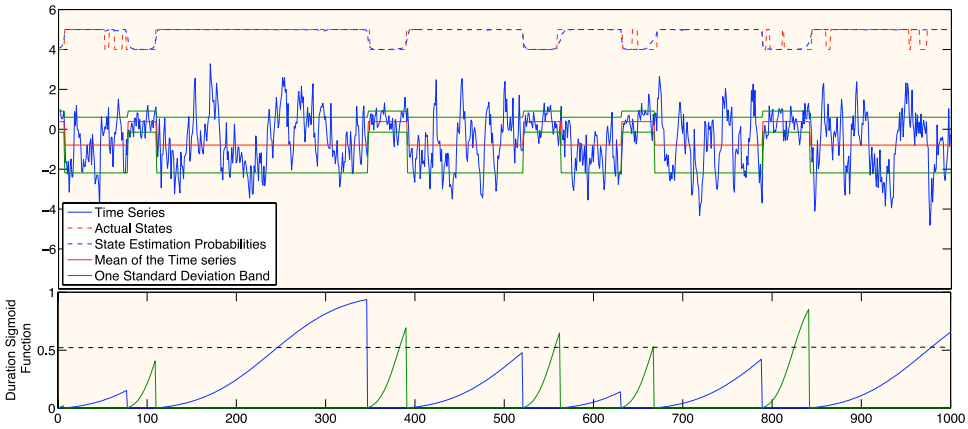

Figure 5-3: (a) A sample univariate time series, as well as the Markov states and their estimations. (b) Duration Sigmoid functions are shown to illustrate the theory

5.3 Example

Figure (5-3) shows a random time series of 1000 data points. The random data points come from two sources (hidden states) with different means and different standard deviations. The expected values, (plus/minus one standard deviation band around the mean) is shown on the time series to illustrate the states. The switching between the two hidden states follows a Markovian dynamics. To show the effectiveness of our model we showed the actual hidden states in Figure 5-3, so we can compare the model's estimation of states with the actual hidden

states. Note that in our derivations we never used any information about the hidden states, since they are supposed to be hidden and unknown.

Figure 5-2, shows the distribution diagrams of the duration of stay for the two states. The expected duration of stay for the two regimes (states) are about 40 and 140 consecutive time frames. The standard deviation around the expected values also shows the uncertainty of the expected values. In the Figure 5-3-b, the time series of D_t^i is shown. This is the sigmoid function that is inhibitory before the inflection point and excitatory after the inflection point.

5.3.1 Remarks

As observed in Figure 5-3, around time frame of 830, and 970, the actual regime switches, although the change in the probabilities is not large enough to change the output of the Viterbi algorithm for the state estimation. Therefore the estimated state does not catch the regime switching at time frame 830 and 970. The state duration probabilities on the sigmoid function, in Figure 5-3-b, however, reach the inflection point at those time frames. Therefore it shows a good measure of the expected values of duration of stay, even when the state probabilities do not catch them.

5.3.2 Comparison

In this approach the durations are not needed to be members of a finite set. Furthermore, one does not need them to be bounded, with the bounds (expected values and variances of the durations) estimated from an EM algorithm.

Yu in [80] indicates that Explicit Duration HMM via a hidden semi-Markov modeling framework faces the magnitude of $O(T(N^2+ND+ND^2))$ computational complexity where T is the length of time series and D is upper bound of duration. Variable Transition HMM faces $O(TN^2D)$ computational complexity, and Residential Time HMM faces the magnitude of $O(T(N^2+ND))$ computational complexity. Since our method uses the same size of EM algorithm

for durations as states, it faces the computational complexity of $O(2TN^2)$. This is typically significantly lower than previous methods since the size of D is typically much larger than N.

Also to illustrate the improvements that the duration dependency has over the goodness of fit of the model, compare the two state estimation of Figure 5-4 below.

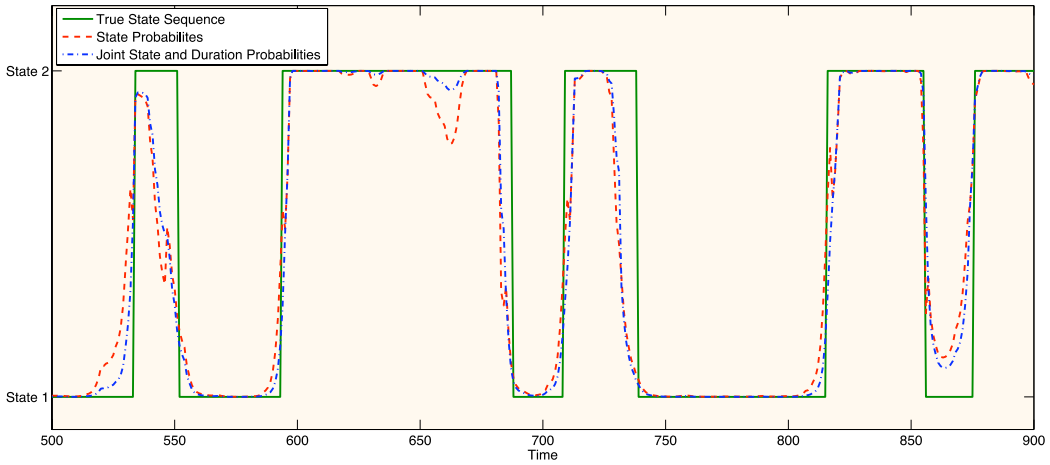

Figure 5-4: State probabilities versus joint State and Duration probabilities

As we can see in this figure, including durations in the Markov probabilities improves the goodness of fit of the HMM. Figure (5-4) clearly shows that the duration dependent HMM has smoother state probabilities that are closer to the true state sequence.

6 HMM in Bifurcation theory

Bifurcation theory is the mathematical phenomenon of changes in the qualitative or topological structure of a given family of dynamic nonlinear systems, such as a family of nonlinear vector fields, and the solutions of a family of differential equations. Most commonly applied to the mathematical study of dynamical systems, a bifurcation occurs when a small smooth change made to the parameter values (the bifurcation parameters) of a system causes a sudden qualitative or topological change in its behavior [83].

To introduce the relationship between hidden Markov modeling and the bifurcation theory, we study one type of bifurcation: local bifurcations.

6.1 Local Bifurcation

Local bifurcation happens when a change of parameter in the nonlinear system causes the change of stabilities among the equilibriums of the dynamic system. To relate this fact to HMM, we take the nonlinear true generative model of our time series to have multiple invariant sets. In this case the time series is the trajectory of the nonlinear dynamic system and multiple invariant

sets are multiple distributions that we are modeling our time series with them. The number of invariant sets in our model is basically the number of states of HMM.

Assume that each state is being modeled by a distribution with an expected value and a distribution around it. Markov switching model assumes that at each time one of these distributions is stable and the rest are unstable.

Figure 6-1: The trajectory and its nonlinear model with four equilibriums

Figure 6-1 illustrates the concept. HMM with four states can model this time series, with 4 Gaussian distributions as their emission models. Then the state with higher Markov probability will be chosen as the stable state, where its expected value is the stable equilibrium point and the distribution around it is its basin of attraction. All other equilibriums are unstable (unattractive). Figure 6-1 shows the location of the stable equilibrium with a solid red line, versus the unstable ones that are shown by dotted red lines.

Any Markov switching point is a local bifurcation point in time. This means that while the number of equilibriums is constant, some "hidden" change of parameter causes topological change of model where the stable equilibrium loses its stability and one of other equilibriums becomes stable. In HMM this is called a Markov switching point.

Note that because of the noise and uncertainty, while the trajectory is traveling through time, the positions of the equilibriums are re estimated.

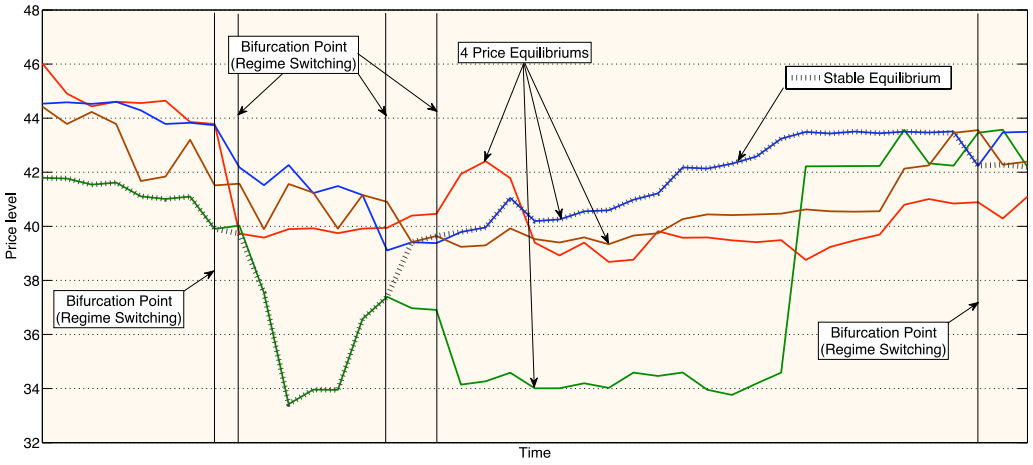

Figure 6-2: Locus of 4 equilibriums in a financial time series

Figure 6-2 shows the locus of the estimated equilibriums over time. In this figure, the solid curves are the locus of the equilibriums while the dotted curve is the position of the STABLE one, while switches back and forth among the equilibriums at the bifurcation points. The vertical lines are the bifurcation points in time.

Note that each equilibrium has its own basin around it that when we find out which one is the stable one, that basin is considered as the basin of attraction.

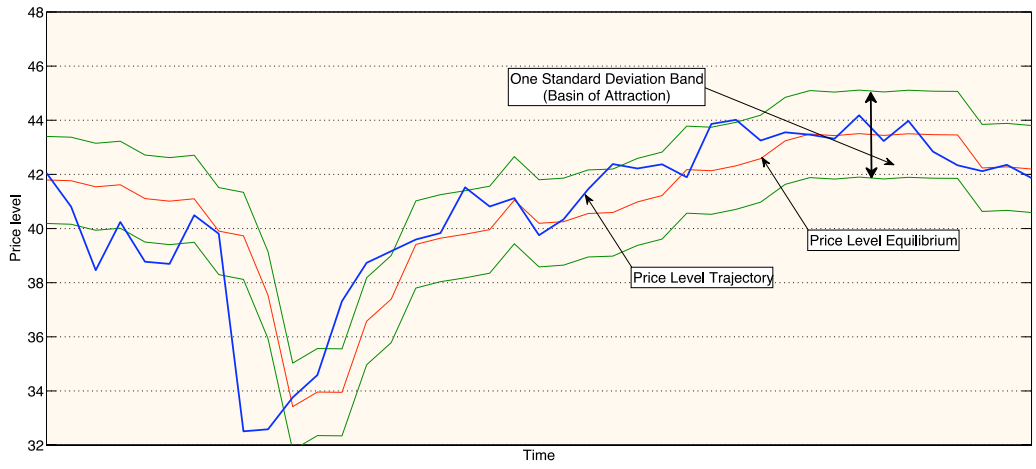

Figure 6-3: Locus of the stable equilibrium and its basin of attraction

Figure 6-3 shows the locus of the stable equilibrium point, i.e. the dotted curve in Figure 6-2, and its basin of attraction.

It's important to notice that when HMM estimates the equilibrium points and their basins of attraction, the actual trajectory at that time is not available. Note that the true time series trajectory (blue curve in Figure 6-3) travels within the basin of attraction and basically oscillates around the stable equilibrium point.

7 Application of HMM in Mean-Reverting processes

Mean reverting processes are commonly used to model different types of phenomena in statistical interfaces. For this purpose the z-score measure is typically used to estimate how far away the time series is from its historical mean, which depends on how mean-reverting the process is. One has to make some basic assumptions related to the distribution and the density function of the time series. In other words estimating the z-score typically follows the Gaussian assumption. The original time series has to be de-trended first. The reason for this is that the distribution of the data points on a time series with a trend is likely to be skewed towards the trend direction, and thus would not necessarily be Gaussian [84-88].

In this chapter we will study the applications of HMM in estimating and forecasting the mean reverting processes through applying to 1- the card game of baccarat and 2- estimating the metric of trustworthiness in cyber security.

7.1 The Game of Baccarat

In this work we first need to construct a mean reverting signal and then study its regime switching with HMM. As an application of the theory we study the game of Baccarat. To generate a mean reverting signal in any card game, the fundamental theorem of card counting is being used [89]. Basically, the idea is to break the card numbers down to discrete groups and then assign symmetric constant numbers to each group. For instance, in Blackjack we can divide the cards to two sets; {2, 3, 4, 5, 6, 7, 8} and {9, 10, J, Q, K, A} and assign a number -1 to the first set and a +1 to the second set, except for Ace that has +2 score. Then we keep a variable "x", initially zero, in our mind by counting the numbers, where we subtract 1 from x, if the card draw is from the first set and we add a 1 (or 2 if the card draw is Ace) to x if the card draw is from the second set [89]. Clearly the variable x is mean reverting with its long-term mean being zero. However if the variable deviates from its mean temporarily, e.g. x=−15, then it says that the remaining cards in the deck have to be more from the second set to eventually bring the mean back down to zero. Therefore the player can bet more on upcoming cards being from the second set. In the game of Baccarat, however, counting cards is more complicated, since the 3rd draw for the Player hand and the Bank hand follows more complicated rules and there are lot of If-Then statements involved. Therefore, upon betting, if the x=−16 for example, it is still uncertain that the cards from the second set will land in the player's hand or the bank's hand, versus Blackjack that the player can decide to draw the next card or not.

Taking a look at the Baccarat game, we know that the output of the game after each round is one of the three: Bank, Player, or Tie. We use a database of Baccarat consist of 1000 Baccarat shoes, with each shoe containing six decks and dealt until the number of cards remaining in the shoe is less than 6 [90]. For our simulation we picked, 75 shoes randomly out of the 1000 shoe in the database. 75 Baccarat shoes, is equivalent to 5135 rounds of dealt hand in Baccarat. Based on historical data of 5135 round of game we have:

$$\begin{aligned} &\Pr(\text{Bank}|\text{Data})=0.4563 \\ &\Pr(\text{Player}|\text{Data})=0.4487 \\ &\Pr(\text{Tie}|\text{Data})=0.0950 \end{aligned} \tag{7-1}$$

Based on fundamental theorem of card counting, we assign symmetric numbers to each outcome of the game [89]. Note that here we don't count the cards, but we count the outcomes. Because counting cards in Baccarat could be very complicated and not necessarily faithful to the player. Therefore, we are constructing a mean reverting signal based on the outcome of each round of the game and not the card draws. If we assign a, +1 to Bank, -1 to Player and 0 to Tie, and keep a variable $x(t)$, equal to the draw values; Bank, Player, tie we can have the time series of outcome. Figure 7-1 shows a sample length of 50 game outcomes to illustrate the concept. Calculating the expected value of $x(t)$ yields:

$$\begin{aligned} \mathrm{E}[x(t)] &= \sum_{i=1}^{3} x_i(t)\Pr(x_i(t)) \\ &= +1\times(0.4563)+0\times(0.0950)-1\times(0.4487) \\ &= 0.0076 \cong 0 \end{aligned} \tag{7-2}$$

Now keep going forward with the theory of card counting, the first signal of interest for us is going to be the cumulative sum of all $x(t)$.

$$y(t) = \sum_{\tau=1}^{t} x(\tau) \tag{7-3}$$

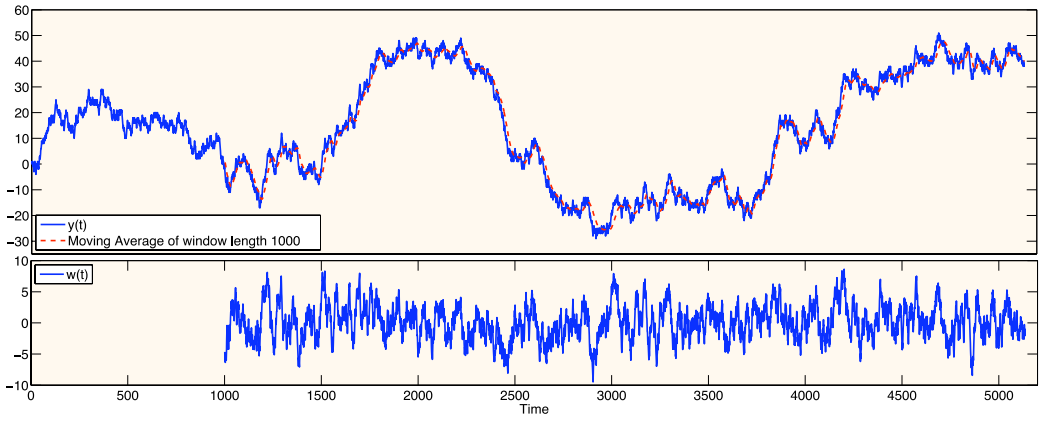

Figure 7-1: (a) The time series of $y(t)$, the moving average of length 1000, and (b) $w(t)$. For illustration purposes $y(t)$ and $w(t)$ are also shown in the time interval of [0 5200].

Figure 7-1 shows the time series of the signal $y(t)$ for our sample Baccarat of 75 shoes. This signal is equivalent to what we described for Blackjack in the section 7.1. In the theory of card counting this variable is easy to track at the table since after each hand based on the outcome of the game to be Bank, Player or Tie, we add a +1, -1, or a 0 to the variable, respectively. The initial value of $y(t)$ is zero. Note that, unlike $x(t)$, the overall mean of the variable $y(t)$ is not zero. This is because the expected value of $x(t)$ is not exactly zero and is skewed towards the positive side a little bit. Intuitively this means that after 5135 sample draw, the difference between the Banks and Players becomes more significant, lifting up the mean.

To build a mean reverting process we consider the deviation of the signal $y(t)$ from its moving average. By definition, the moving average of $y(t)$ crosses $y(t)$ frequently, and it means that at the intersection points the deviation signal is zero. The deviation of $y(t)$ from its moving average oscillates around zero and therefore is mean reverting. Let's build the deviation signal of interest, $w(t)$:

$$\begin{aligned} w(t) &= y(t) - \underset{\tau=\tau_0}{\overset{t}{\mathrm{E}}}[y(\tau)] \\ &= y(t) - \left(\frac{\sum_{\tau=\tau_0}^{t} y(\tau)}{\tau_0} \right) \end{aligned} \qquad \textbf{(7-4)}$$

Figure 7-1 also shows the signal $w(t)$; the deviation of the signal $y(t)$ from its moving average of rolling window of length 1000. This means that at each instant of time t, the variable $(t - \tau_0)$ is kept constant at 1000. This implies that as time passes and t gets larger, τ_0 also moves forward and gets larger; therefore the difference (the window length of the moving average) staying constant, e.g. 1000. Figure 7-1-b is $w(t)$, the mean reverting signal of interest in this research.

7.1.1 How to use HMM to play?

Note that since the moving average of rolling window of length 1000, starts from the data point 1000, our mean reverting signal of interest, $w(t)$, also starts from 1000. When the mean is temporarily positive, the HMM is in state 1 in Figure 7-2 and 7-3, this means that the mean reverting signal; $w(t)$, is temporarily above zero. This also means that in Figure 7-1, the $y(t)$ signal is temporarily above its Moving Average and basically the Moving Average signal is trying to catch up with the signal $y(t)$, although it is not able to reach it temporarily. This means that $y(t)$ is on an upward trend, and essentially the number of Banks (i.e. +1) is more than Players (i.e. -1). Simply, put State 1 means more likelihood of Bank than player, and State 2 means more likelihood of Players than Banks. Therefore the idea is that if HMM is on State 1, we bet on Bank, and if HMM is on State 2 we bet Player.

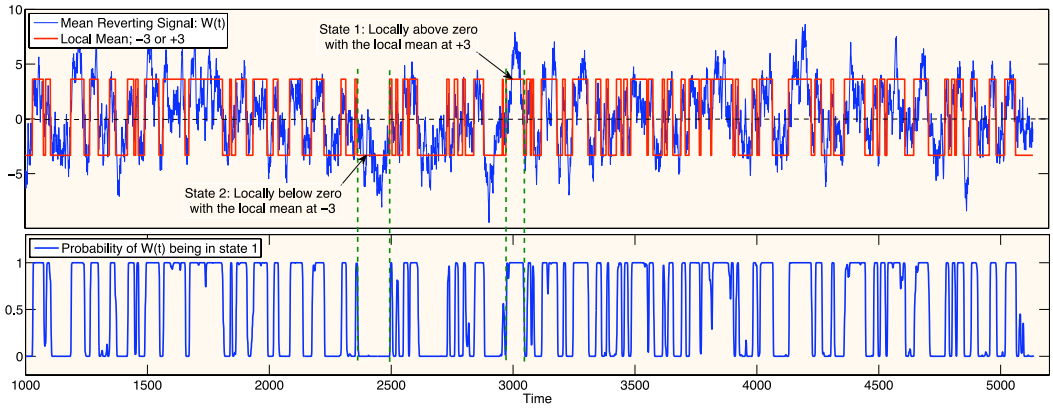

Figure 7-2: (a) The output of the regime switching model on the input signal $w(t)$. The top figure shows the signal $w(t)$ and its regime switching between 2 states, one with a local mean of +3 and one with the local mean of -3. (b) Markov probabilities.

Note that in Figure 7-2-a the local means also are estimated as parameters of the EM estimation. Figure 7-2-b shows the probabilities of being on state 1 at any instant in time. The Probability of 1 means 100% probability of being on State 1, whereas probability of 0, means 100% probability of being on State 2. For illustration purposes, the figures are zoomed for the time span of [1500,5200].

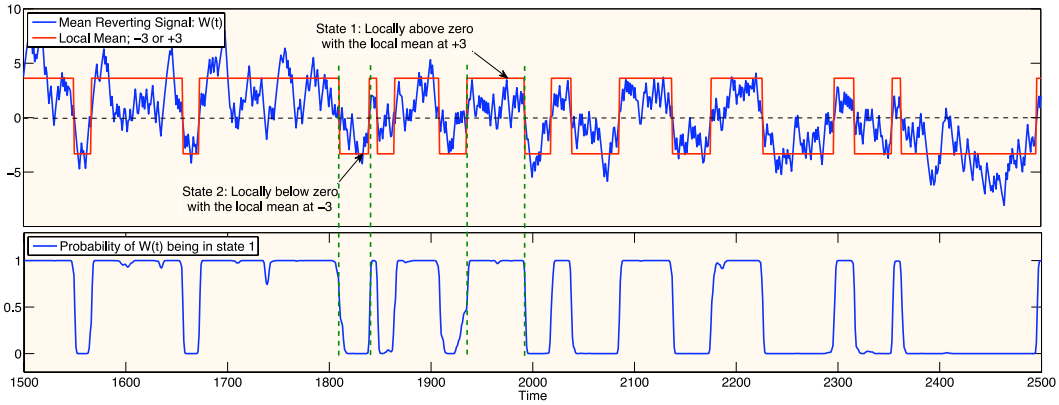

Figure 7-3: **The zoomed-in version of the last figure, at the time span of [1500,2500], for better clarity of the concept**

7.1.2 Optimal Leverage Factor

The question of interest to answer here in this section is the following: In the game of Baccarat, the State 1 is Bank, and State 2 is Player. If we know which state will come next, we can have a strong advantage over the house (casino). But we don't have that information, since we simply cannot predict the future for sure. However, we have the *probabilities* of State 1 and 2 (i.e. Bank and Player) in the next round. This is $\gamma_t(i)$ for i =1, 2. Clearly we know that we want to bet on the state that has the probability of greater than 50%. If the probability is less than 50% we simply bet on the other state that has more probability. If, however, the probability of one side, say Bank, is 99% we are willing to bet more on that particular round than if the Probability was 51%. Therefore *If we knew the probability of winning, to bet on the state, say Bank, then what is the optimal value of your money that you have to bet at that round to maximize your winnings over time?*

Let's answer this question:

Let $0.5 < \gamma(t) < 1$ be the probability of winning at time t. Clearly if the probability is less that 0.5, the other state is of interest. Let $0 < L(t) < 1$ be the leverage factor, that is the percentage of your money that you want to bet at time t. Let $Q(t)$ be the payoff function. For simplicity we ignore the house commission and/or transaction costs. We have:

$$\begin{aligned} Q(t) &= +L(t).\gamma(t) - L(t).(1-\gamma(t)) \\ &= +L(t).[2\gamma(t)-1] \end{aligned} \tag{7-5}$$

The problem of interest is to maximize $Q(t)$:

$$\begin{aligned} L(t) &= \text{Argmax}\{Q(t)\} \\ &= \text{Argmax}\{\ L(t).[2\gamma(t)-1]\ \} \\ &\quad \text{subject to:}\quad 0.5<\gamma(t)<1 \\ &\qquad\qquad\quad 0<L(t)<1 \end{aligned} \tag{7-6}$$

The solution to this constraint maximization is simple:

$$\begin{aligned} &\dot{Q}(t) = \dot{L}[2\gamma-1] + L[2\dot{\gamma}] = 0 \\ &\Rightarrow \frac{\dot{L}}{L} = \frac{-2\dot{\gamma}}{2\gamma-1} \Rightarrow L = -k(2\gamma-1) \end{aligned} \tag{7-7}$$

where $k \neq 0$ is a constant. Now let's find k:

$$\begin{aligned} &0.5 < \gamma < 1 \\ &\Rightarrow 0 < 2\gamma - 1 < 1 \\ &\Rightarrow \begin{cases} 0 < -k(2\gamma-1) < 1 & \text{if} \quad k < 0 \\ -1 < -k(2\gamma-1) < 0 & \text{if} \quad k > 0 \end{cases} \end{aligned} \tag{7-8}$$

Using the second constraint in Eq. (7-6) and the result of Eq. (7-7) we have:

$$0 < L < 1$$

$$\Rightarrow 0 < 2\gamma - 1 < 1$$

$$\Rightarrow \begin{cases} 0 < -k(2\gamma - 1) < -k & \text{if} \quad k < 0 \\ 0 < -k(2\gamma - 1) < 1 & \text{if} \quad k < 0 \end{cases} \tag{7-9}$$

$$\Rightarrow k = -1 \Rightarrow L(t) = 2\gamma(t) - 1 \therefore$$

This means that if the probability of winning is, say, 60%, the optimal leverage factor is (2x0.6-1)=20% of your wealth at time *t*. Now if the probability of winning is 90%, the optimal leverage factor is (2x0.9-1)=80%. Note that in this derivation the person at the table theoretically will never run out of money since its bet is always a fraction of its cash.

7.2 Trustworthiness in Cyber Security

A major challenge for a trust model is that trust is inherently application-dependent. The important question is whether an overarching unified trust model can be made adaptable to changing application environments. We have chosen *trust* (subjective-human centric) and *trustworthiness* (objective) as factors to assess the success of a complex system, such as a network comprised of sources of information, knowledge, hardware and software. Therefore, a complex system can be considered a heterogeneous network with multiple trustworthiness measurements [91]. A unified trust model allows trustworthiness assessment and trust judgment of complex systems, such as a distributed sensing network, since it can address issues such as risk, vulnerability, uncertainty, and confidence [91-93].

The studies by [94-97] provided definitions of trust and trustworthiness of a complex system and established features of a mathematical model that could assure trustworthiness measurements of a complex system. One specific example of the application of "Trust Model" is the assurance of secure data query processing in wireless sensor networks in both commercial and defense sectors. Trust is the essential component of the in-network decision making among the sensors. These networks depend on electronic connectivity in their operation, which is

subject to physical wear and tear as well as malicious attacks. For these networks to be efficient as well as secure it is shown that cryptography or authentication alone is not sufficient to provide trustworthy networks [98, 99]. A trust framework using both reputation and trustworthiness data of sensor nodes is the strategy to protect distributed complex systems against malicious attacks, tamper, and exploitation of intellectual property.

The studies by [96, 97] extend the definition of trust for sensor networks to reflect the dual functions of nodes in sensor networks versus traditional ad-hoc networks. The nodes in sensor networks relay data and also generate and collaboratively process information [98-100]. The definition provided in [100] is: "Trust is the node's *belief* in the competence and the reliability of another node. In other words, trust is the subjective probability by which node A depends on node B to fulfill its promises in performing an action and at the same time being reliable in reporting its sensor data".

We provide our definition of the metric of trustworthiness:

The probability that a data point at time $t=T$ belongs to the distribution of satisfactory behaviors (e.g. Beta, Gaussian, Gamma, etc.) during $t=1, \ldots, T-1$, is a metric of trustworthiness.

This definition is system agnostic and does not depend on any particular distribution or underlying factors, e.g., Subject, Agent, Action. The estimation and prediction methods presented in this dissertation would successfully accommodate the desired features of a mathematical trust model presented in [96, 97], which are: it must (1) support a heterogeneous network, (2) could accommodate multiple trustworthiness measures: e.g. multidimensional time series with different underlying characteristics in different Trust regimes, (3) be carried out with computational ease without extensive computational power from the sensor network, (4) be conceptually simple but have a firm basis in theory, and (5) be application independent. These properties were gleaned from trust models in wireless sensor networks, social networks, e-commerce, mobile ad-hoc, peer-to-peer, and distributed network services.

Moreover, we have shown in Chapter 3 that the cost of implementation is linear in this approach as opposed to supervised machine learning [101]. Liu et al. [101] showed that the number of service transactions must stay at 2000 for the successful transaction rates to remain at higher than 90% for a 50-node system in [101]. The scalability of this approach is therefore questionable for a large system, e.g. a 1000-node system.

7.2.1 Trust Model

We will study the applications of hidden Markov machine on modeling "trustworthiness". For this purpose, we generate a hypothetical scenario where the notion of trust is defined between an online vendor and its customers. In this case we take the simulated data to be the number of customers per day as the input signal X(t). Therefore by definition the percentage change of the number of customers per day is a measure of trust; Y(t)=[X(t)-X(t-1)]/X(t-1). In other words, if the number of customers for today has increased comparing with yesterday; Y(t)>0, we say that the vendor is more trusted today. And on the other hand, if the number of customers for today has decreased from yesterday; Y(t)<0, we say that the trust to the vendor has decreased. For simple cases we can look at the sign of Y(t) at each day and decide if the vendor is in the Trusted state or in the Untrusted state. In practice the problem is far more complicated. Various factors such as product availability or reliability, word of mouth, purchase power, negative feedback, privacy concern, customers' disposition, and vendor's reputation could impact trust in online transactions [102, 103]. In our approach, the factors that change Y(t) are collectively modeled as noise.

Trust evolves over time and is time sensitive; more recent actions should have more impact on the trust value. Time series analysis or autoregressive models offer tools to take care of the serial correlation present in trust scenarios. In autoregressive model of lag 1 or AR (1), the regression of a deviation on all previous deviations depends only on the most recent one. Equation (7-10), therefore, represents a general form for trust model accounting for autocorrelation terms and the uncertainties (noises). Note that extension of AR(q) with q=1, to any q>1, via a standard autoregressive framework in Eq.(7-10) is straightforward.

$$\begin{aligned} \Upsilon(t,\theta_1) &= \alpha_1 + \Upsilon(t-1)\beta_1 + \varepsilon_1 \quad \therefore \quad \varepsilon_1 \sim N(0,\sigma_1^2) \\ \Upsilon(t,\theta_2) &= \alpha_2 + \Upsilon(t-1)\beta_2 + \varepsilon_2 \quad \therefore \quad \varepsilon_2 \sim N(0,\sigma_2^2) \end{aligned} \tag{7-10}$$

where the parameter set is defined as $\theta = \{\alpha, \beta, \sigma\}$. Borrowing the concept of regime switching that we explained in chapter 3 we can simulate a trust scenario.

In this dynamic simulation we use Metropolis-Hastings Markov Chain Monte Carlo (MCMC) to sample from two Normal distributions, one for trusted regime and one for untrusted regime. Here we assume that if customers trust the online vendor, they are 98% likely to stay trusted to the vendor. Intuitively this assumption means that a vendor that has achieved a level of trustworthiness would work hard to keep its reputation that way. In 2% of the times, however, for whole lot of reasons, the customers might lose their trust for the vendor. In that case they are 95% likely to remains untrusted of the vendor. In other words it takes a lot of time and efforts from the vendor's side to win that trust back from its customers. Based on model shown in Eq. (7-10), we assume the parameter set is taken as:

$$\begin{aligned} &\alpha_1 = +0.1,\ \ \beta_1 = +0.4,\ \sigma_1^2 = +0.2 \\ &\alpha_2 = -0.1,\ \ \beta_2 = +0.7,\ \sigma_2^2 = +0.6 \end{aligned} \qquad \textbf{(7-11)}$$

Here, $\alpha_1 = +0.1$, means that in the trusted regime, the Y(t) is supposed to increase day by day. However we have; $\beta_1 = +0.4,\ \sigma_1^2 = +0.2$, which imposes an autoregressive model with Gaussian noise that takes all other factors into account. On the other hand, $\alpha_2 = -0.1$, means that in the untrusted regime, the Y(t) is supposed to decrease day by day. Again $\beta_2 = +0.7,\ \sigma_2^2 = +0.6$, takes all other factors into account including noise. Note that the variance of the noise in the untrusted regime is assumed to be 3 times the variance of the noise in the trusted regime. This is due to the imperfect information distribution among the customers. Intuitively this means that there might be many customers that don't realize that the vendor has lost its trustworthiness and still keep going back to the vendor for a while.

The blue time series on the top part of the Figure 7-4 shows the input signal to our regime switching model; Y(t). One important characteristic of Markov switching models is that they are NOT application-dependent and they are unsupervised learners. This means that to model this trust system, our hidden Markov framework does NOT need ANY of the assumption and/or parameters that we explained in Section 7.2.

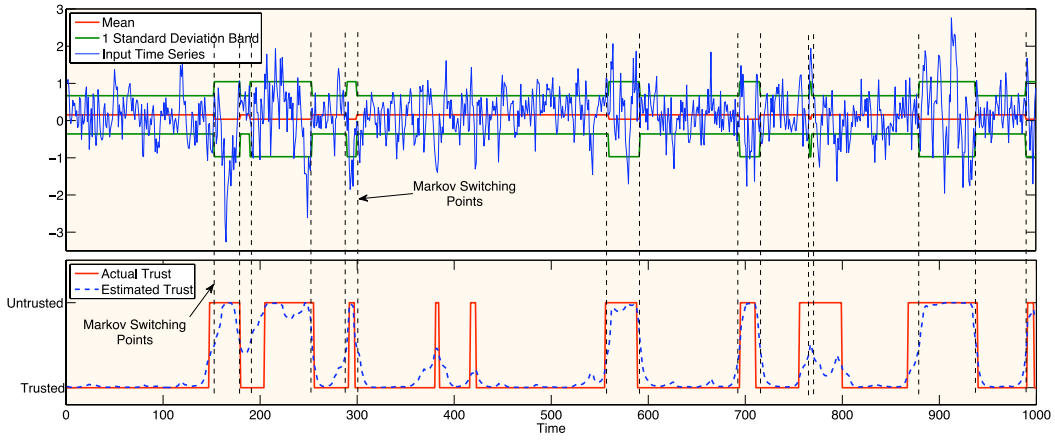

Figure 7-4: (a) The input time series Y(t) and (b) its regime switching

The only input that the HMM needs to model this trust system is the blue time series on the top part of the Figure 7-4. This characteristic is very important since it makes this trust framework application-independent. All the parameters that we assumed in Eq. (7-11) and their intuitions and all the philosophies that we talked about in Figure 7-4 could be assumed to be hidden. In fact it is the HMM's job to not only estimate the regime switches, but also estimate the parameters of Eq. (7-11) and also the percentages of Figure 7-4.

The red step function in Figure 7-4-b, shows that the actual regime switches based on the MCMC modeling of the dynamics of Figure 3-1. Note that this actual regime switching as a function of time is hidden and has to be estimated by HMM. The blue dotted plot in Figure 7-4-b, is the HMM estimation of the regime switching between a trusted regime and an untrusted regime. Note that there are times around t=400 and also t=780 that the HMM is not able to perfectly estimate the regime switches.

Also note that the HMM is able to estimate the parameters of the underlying autoregressive models that generated the time series. These essentially are the parameters of Eq. (7-11). And based on the Mean and Variance of an AR(1) process the HMM is able to estimate the expected values and the standard deviations of the time series Y(t) for each regime, as shown by the solid red and green step plots in Figure 7-4-a. As shown there the expected value of Y(t) (i.e., the percentage change of returning customers) compared with the day before in the trusted region is estimated to be positive as was expected. Additionally the noise in the trusted regime is

estimated to be smaller again as was expected. Conversely, in the untrusted region the expected value of $Y(t)$ is estimated to be negative and the noise is larger.

The Markov switching on Trust models, studied here is based on duration-dependent hidden Markov machine. As an unsupervised machine learning method, this framework is independent of assumption and nature of the Trust model. In other words, the only input needed to model the Trust system with HMM is a relevant time series that switches regimes from Trusted to Untrusted periods of times.

HMM successfully estimated the parameters of Eq. (7-11) and could accommodate the desired features of the trust model specified in this section. This successful estimation occurred despite various noises in the input time series. The noise or uncertainties in Figure 7-4 could be due to factors that impact trust in on-line transactions such as product availability or reliability, word of mouth, purchase power, negative feedback, privacy concern, customers' disposition, and vendors' reputation [102, 103].

8 Conclusion

This study addresses improvements to the theory and application of continuous-time multivariate hidden Markov modeling, beyond a large body of past studies. To effectively study this subject, we divided this dissertation into three parts.

The first part, documented in Chapter 1 and 2 was an introduction about the linear models, system identifications and their associate maximum likelihood techniques. The second part, which is the heart of this work, is included in Chapters, 3 though 5. Chapter 3 studies three main points. First is the theory of multivariate continuous-time HMM. This is essentially an extension of what Shi has done for his PhD dissertation in [1]. Shi's work was continuous time HMM with two different types of emission model, one being an AR process, and the other being a neural network. Chapter 3 extends Shi's work form univariate emission models to multivariate observations. Chapter 3 also includes two other enhancements to the theory of multivariate continuous-time HMM: one being a robust parameter estimation for the emission model of EM algorithm and the other being a closed-form solution to the emission parameter estimation of the EM algorithm for some special cases.

Chapter 4 introduced an extended Viterbi algorithm by studying a combined state sequence for a two-step HMM, in order to increase the goodness of fit of this mode.

Chapter 5 introduced a duration-dependent hidden Markov model. The main contribution of this chapter is providing a better estimation of the expected duration of stay at each regime. The durations were modeled by excitatory and inhibitory interactions of neurons in central nervous system of human body, based on a pioneering work of Wilson and Cowan in 1972.

Chapters 6, and 7 took a look at some of the applications of HMM in different branches of science and technology. Chapter 6 compared the HMM with the bifurcation theory in nonlinear dynamics. Chapter 7 studies the application of HMM in card games via the fundamental theory of card counting and provided an example of how the method can be applied to the game of Baccarat. This chapter also studied estimating "Trustworthiness" in cyber security. A scenario of Trust for an online vendor and the estimation of Trustworthiness through HMM were provided.

Therefore the significant contributions of this entire study can be summarized. This work claims that it solves some fundamental problems of HMM that the Markovian dynamics have been struggling with for decades.

1) The first contribution of this study was to achieve computational efficiency on the order of 50 times faster than the other methods. While time consuming calculations and computational costs of HMM has always been one of the biggest obstacles on its way, the closed-form solution to the local maxima of the likelihood surface improved the existing answers to the problem of computational complexity.
2) Another contribution of this work is to have smoother and better state probability estimations. This is particularly important for out-of-sample application where HMM will typically have some delay to catch a potential regime switches.
3) The third contribution of this work is to effectively estimate the expected duration of stay on each regime with a set of sigmoid functions.
4) Applying these enhancements to a few science and engineering problems is another aspect of this study's contributions.

8.1 Recommendation for Future studies

Futures extensions of this work can be done in several directions. One direction is to extend the theory in the area of Automatic Controls. Different branches of linear/nonlinear Controls can accompany HMM and generate new branches of Controls methods.

Another direction for future advances in this research is the applications of HMM in the industrial vibrations' energy harvesting. When the magnitude and frequency of the mechanical vibration changes from time to time, the parameters of the energy harvester might need to be adjusted to maximize the energy generating. HMM can estimate the regime switching in the mechanical vibrations and forecast them on real time effectively.

Another application of HMM would be in estimating the infrastructure quality of railroads. Monitoring the vertical acceleration of the train wheel sets on real time, when analyzed with HMM can estimate the "good quality infrastructure" regime switching to "bad quality infrastructure".

References

[1] S. Shi and A. S. Weigend, "Taking time seriously: hidden Markov experts applied to financial engineering," in *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*, 1997, pp. 244-252.

[2] S. Roweis and Z. Ghahramani, "A Unifying Review of Linear Gaussian Models," *Neural computation,* vol. 11, pp. 305-345, 1999.

[3] R. E. Kalman, *Contributions to the theory of optimal control* vol. [S.l.], 1960.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE,* vol. 77, pp. 257-286, 1989.

[5] L. R. Rabiner, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *The Bell System technical journal,* vol. 62, p. 1075, 1983.

[6] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell System technical journal,* vol. 62, p. 1035, 1983.

[7] G. E. Hinton, M. Revow, and P. Dayan, "Recognizing Handwritten Digits Using Mixtures of Linear Models," *The Press (Christchurch, N.Z.),* p. 1015, 1994.

[8] J. Baker, "The DRAGON system--An overview," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 23, pp. 24-29, 1975.

[9] J. Baker, "Stochastic modeling for automatic speech understanding," *Speech Recognition (D. R. Reddy editor), Academic Press, New York,* 1975.

[10] F. Jelinek, "Self-organizing Language Modeling for Speech Recognition," *IBM Research Report,* 1985.

[11] F. Jelinek, "Statistical Methods for Speech Recognition," *The Press (Christchurch, N.Z.),* 1998.

[12] A. Averbuch, L. Bahl, R. Bakis, P. Brown, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, E. Epstein, D. Fraleigh, F. Jelinek, B. Lewis, R. Mercer, J. Moorhead, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman, P. Spinelli, D. Van Compernolle, and H. Wilkens, "Experiments with the Tangora 20,000 word speech recognizer," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, 1987, pp. 701-704.

[13] S. U. H. Qureshi, "Adaptive equalization," *Proceedings of the IEEE,* vol. 73, pp. 1349-1387, 1985.

[14] J. Zhang, "The mean field theory in EM procedures for Markov random fields," *Signal Processing, IEEE Transactions on,* vol. 40, pp. 2570-2583, 1992.

[15] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *Speech and Audio Processing, IEEE Transactions on,* vol. 1, pp. 431-442, 1993.

[16] R. Chen and J. Liu, "Mixture Kalman Filters."

[17] A. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, 1982, pp. 1291-1294.

[18] A. B. Poritz, "Hidden Markov models: a guided tour," *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing,* pp. 7-13, 1988.

[19] L. P. Neuwirth, "Unpublished Lectures," 1970.

[20] J. Ott, "Counting methods (EM algorithm) in human pedigree analysis: Linkage and segregation analysis," *Annals of human genetics,* vol. 40, pp. 443-454, 1977.

[21] P. Billingsley, "Statistical inference for Markov processes / Patrik Billingsley," ed: Chicago, 1961.

[22] J. D. Ferguson, "Unpublished Lectures," 1974.

[23] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities,* vol. 3, pp. 1-8, 1972.

[24] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Society,* vol. 73, p. 212, 1967.

[25] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of mathematical statistics,* vol. 37, p. 1554, 1966.

[26] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of mathematical statistics,* vol. 41, p. 164, 1970.

[27] L. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE transactions on information theory,* vol. 28, p. 729, 1982.

[28] J. K. Baker, "Trainable grammars for speech recognition," *The Journal of the Acoustical Society of America,* vol. 65, p. S132, 1979.

[29] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures " *Math. Inst. Hungarian Acad. Sci., Budapest,* 1982.

[30] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM review,* vol. 26, p. 195, 1984.

[31] B. Templeton, "A Polynomial Chaos Approach to Control Design," Doctor of Philasophy, CVeSS, Mechanical Engineering, Virginia Tech, Blacksburg, 2009.

[32] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*: Prentice-Hall, Inc., 1988.

[33] H. R. Wilson and J. D. Cowan, "Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons," *Biophysical Journal,* vol. 12, pp. 1-24, 1972.

[34] J. N. Juang, *Applied System Identification*. Upper Saddle River, NJ: Prentice Hall PTR, 1994.

[35] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pacific journal of mathematics,* vol. 27, p. 211, 1968.

[36] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME,ÄìJournal of Basic Engineering,* vol. 82, pp. 35-45, 1960.

[37] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Transactions of the ASME. Series D, Journal of Basic Engineering,* vol. 83, pp. 95-107, 1961.

[38] R. E. Kalman, P. L. Falb, and M. A. Arbib, *Topics in mathematical system theory*: McGraw-Hill, 1969.

[39] H. Rauch, "Solutions to the linear smoothing problem," *IEEE Transactions on Automatic Control,* vol. 8, pp. 371-372, 1963.

[40] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIA journal,* vol. 3, p. 1445, 1965.

[41] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE transactions on information theory,* p. 260, 1967.

[42] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. New York: Springer-Verlag, 1995.

[43] J. Oud, "SEM State Space Modeling of Panel Data in Discrete and Continuous Time and its Relationship to Traditional State Space Modeling," in *Recent developments on structural equation models*, K. van Montfort, J. Oud, and A. Satorra, Eds., ed: Springer, 2004, pp. 13-40.

[44] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*: The MIT Press, 1983.

[45] R. H. Shumway and D. S. Stoffer, "AN APPROACH TO TIME SERIES SMOOTHING AND FORECASTING USING THE EM ALGORITHM," *Journal of Time Series Analysis,* vol. 3, pp. 253-264, 1982.

[46] R. Shumway and D. Stoffer, *Time series analysis and its applications*: Springer, 2000.

[47] Z. Ghahramani and G. E. Hinton, "Switching state-space models," ed, 1996.

[48] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," *Tech. Rep. CRG-TR-96-2, Dept. of Computer Science, University of Toronto, Toronto, CA* 1996.

[49] A. P. Dempster, "Maximum likelihood from incomplete data via the EM algorithm," *Applied statistics,* vol. 39, p. 1, 1977.

[50] R. M. Neal and G. E. Hinton, *A review of the EM algorithm that justifies incremental, sparse and other variants* Dordrecht, MA: Kluwer: Learning in Graphical Models, 1998.

[51] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer speech & language,* vol. 1, p. 29, 1986.

[52] E. P. Neuburg, "Markov Models for Phonetic Text," *The Journal of the Acoustical Society of America,* vol. 50, p. 116, 1971.

[53] F. Jelinek, "Fast sequential decoding algorithm using a stack," *IBM journal of research and development,* vol. 13, p. 675, 1969.

[54] L. R. Bahl, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. PAMI-5, pp. 179-190, 1983.

[55] G. D. Forney Jr, "The viterbi algorithm," *Proceedings of the IEEE,* vol. 61, p. 268, 1973.

[56] A. M. Fraser and A. Dimitriadis, "Forecasting probability densities by using hidden Markov models " *Addison-Wesley Reading, MA,* vol. Time Series Prediction: Forecasting the future and Understanding the Past, pp. 265-282, 1994.

[57] Fan. K., "Les fonctions définies-positives et les fonctions complètement monotones," *Sci. Math. L'Acad. Sci. de Paris,* 1950.

[58] J. L. Doob, "Heuristic approach to the Kolmogorov-Smirnov theorems," *The Annals of mathematical statistics,* vol. 20, p. 393, 1949.

[59] M. Russell and R. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, 1985, pp. 5-8.

[60] S. V. Vaseghi, "State duration modelling in hidden Markov models," *Signal processing,* vol. 41, p. 31, 1995.

[61] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*: John Wiley & Sons, 584pp., 2001.

[62] J. M. Durland and T. H. McCurdy, "Duration-Dependent Transitions in a Markov Model of U.S. GNP Growth," *Journal of Business & Economic Statistics,* vol. 12, pp. 279-288, 1994.

[63] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica,* p. 357, 1989.

[64] P. S. Lam, "A Markov-switching model of GNP growth with duration dependence," *International economic review (Philadelphia),* p. 175, 2004.

[65] I. Hirokuni, "Duration dependence of the business cycle in Japan: A Bayesian analysis of extended Markov switching model," *Japan and the World Economy,* vol. 19, pp. 86-111, 2007.

[66] M. Pelagatti, "Gibbs sampling for a duration dependent Markov switching model with an application to the US business cycle," *statistica.unimib.it,* vol. QD2001/2, 2001.

[67] M. Russell and A. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, 1987, pp. 2376-2379.

[68] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMM's," *Speech and Audio Processing, IEEE Transactions on,* vol. 3, pp. 213-217, 1995.

[69] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 548-551 vol.1.

[70] P. M. Djuric and C. Joon-Hwa, "An MCMC sampling approach to estimation of nonstationary hidden Markov models," *Signal Processing, IEEE Transactions on,* vol. 50, pp. 1113-1123, 2002.

[71] M. T. Johnson, "Capacity and complexity of HMM duration modeling techniques," *Signal Processing Letters, IEEE,* vol. 12, pp. 407-410, 2005.

[72] C. Wei-ho and Y. Kung, "Modified hidden semi-markov model for modelling the flat fading channel," *Communications, IEEE Transactions on,* vol. 57, pp. 1806-1814, 2009.

[73] S. Winters-Hilt and J. Zuliang, "A Hidden Markov Model With Binned Duration Algorithm," *Signal Processing, IEEE Transactions on,* vol. 58, pp. 948-952, 2010.

[74] S. Winters-Hilt, Z. Jiang, and C. Baribault, "Hidden Markov model with duration side information for novel HMMD derivation, with application to eukaryotic gene finding," *EURASIP J. Adv. Signal Process,* vol. 2010, pp. 1-11, 2010.

[75] M. Dong and Y. Peng, "Equipment PHM using non-stationary segmental hidden semi-Markov model," *Robot. Comput.-Integr. Manuf.,* vol. 27, pp. 581-590, 2011.

[76] L. M. Lee, "High-Order Hidden Markov Model and Application to Continuous Mandarin Digit Recognition " *Journal of Information Science And Engineering,* vol. 27, pp. 1919-1930, 2011.

[77] X. Yi, Y. Shun-zheng, T. Shensheng, and H. Xiangnong, "A Two-Layer Hidden Markov Model for the Arrival Process of Web Traffic," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on*, 2011, pp. 469-471.

[78] S. Calinon, A. Pistillo, and D. G. Caldwell, "Encoding the time and space constraints of a task in explicit-duration Hidden Markov Model," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, 2011, pp. 3413-3418.

[79] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," *Speech and Audio Processing, IEEE Transactions on,* vol. 4, pp. 360-378, 1996.

[80] S. Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence,* vol. 174, pp. 215-243, 2010.

[81] A. Householder, *The theory of matrices in numerical analysis*, 1964.

[82] P. Lancaster, *Theory of Matrices* New York-London: Academic Press, 1969.

[83] P. Blanchard, R. L. Devaney, and G. R. Hall, *Differential Equations*. London: Thompson, 2006.

[84] R. Elliott, P. Fischer, and E. platen, "Hidden Markov model filtering for a mean reverting interest rate model," *The Canadian applied mathematics quarterly,* vol. 7, p. 381, 1999.

[85] M. J. Dueker, "Markov switching in GARCH processes and mean-reverting stock-market volatility," *Journal of Business & Economic Statistics,* p. 26, 1997.

[86] B. M. Friedman and D. I. Laibson, "Economic implications of extraordinary movements in stock prices," *Brookings papers on economic activity,* vol. 1989, p. 137, 1989.

[87] M. Elliott, "Model Averaging Methods for Weight Trimming," ed.

[88] R. J. Elliott, *An introduction to latent variable models* London: Chapman & Hill, 1984.

[89] E. O. Thorp and W. E. Walden, "The Fundamental Theorem of Card Counting with applications to Trente-et-Quarante and Baccarat," *International Journal of Game Theory,* vol. 2, pp. 109-119, 1973.

[90] "http://en.wikipedia.org/wiki/Baccarat."

[91] J. Erickson, "Trust Metrics," presented at the International Symposium On Collaborative Technologies And Systems, IEEE, Chicago, Il., , 2009.

[92] J. B. Lyons and C. K. Stokes, "Predicting Trust in Distributed Teams: Dispositional Influences," presented at the 24th Society for Industrial and Organizational Psychology Symposium, New Orleans, Louisiana, 2009.

[93] Y. Sun, W. Yu, Z. Han, and K. J. R. Liu, "Information Theoretic Framework of Trust Modeling and Evaluation for Ad Hoc Networks," *IEEE Journal on Selected Area in Communications,* vol. 24, pp. 305-317, 2006.

[94] S. Hall and W. McQuay, "Review of Trust Research from an Interdisciplinary Perspective - Psychology, Sociology, Economics, and Cyberspace," in *Proceedings of National Aerospace and Electronics Conference*, Dayton, Ohio, July, 14-16 2010.

[95] S. Hall, W. McQuay, and K. Ball, "Initial results from an interdisciplinary review of trust research," in *Proceedings of the ASME 2010 International Mechanical Engineering Congress & Exposition IMECE2010*, Vancouver, British Columbia, Canada, November 12-18, 2010.

[96] S. Hall and W. McQuay, "Fundamental features of a unified trust model for distributed systems," in *Proceedings of National Aerospace and Electronics Conference*, Dayton, Ohio, July, 20-22, 2011.

[97] S. Hall, W. McQuay, and E. Vance, "Features of a trust model for a complex system," in *Proceedings of the ASME 2011 International Mechanical Engineering Congress & Exposition IMECE2011*, Denver, Colorrado Nov. 2011.

[98] H. Deng, G. Jin, R. Xu, W. Shi, and F. Harlow, "Ensuring data integrity through trust in wireless sensor networks," in *Proceedings of SPIE Wireless Sensing, Localization, and Processing V Conference*, Orlando, FL, 2010.

[99] K. Thirunarayan, P. Anantharam, C. A. Henson, and A. P. Sheth, "Some Trust Issues in Social Networks and Sensor Networks," in *Proceedings of 2010 International Symposium on Collaborative Technologies and Systems, IEEE,*, Chicago, IL, 2010.

[100] T. Sobh, K. Elleithy, A. Mahmood, and M. Karim, *Modeling Trust in Wireless Sensor Networks from the Sensor Reliability Prospective*, 2007.

[101] Z. Liu, S. S. Yau, D. Peng, and Y. Yin, "A Flexible Trust Model for Distributed Service Infrastructures," presented at the 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing (ISORC), 2008.

[102] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in Online Shopping: An Intergrated Model," *MIS Quarterly,* vol. 27, pp. 51-90, 2003.

[103] D. Gefen, E. Karahanna, and D. W. Straub, "nexperience and Experience with Online Stores: The Importance of TAM and Trust," *IEEE Transactions on Engineering Management,* vol. 50, pp. 307-321, 2003.