Stockholms universitet

Hidden Markov Models
Theory and Simulation

André Inge

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2013:2
Matematisk statistik
Juni 2013

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2013:2**
http://www.math.su.se

Hidden Markov Models

Theory and Simulation

André Inge*

June 2013

Abstract

Markov chains describe stochastic transitions between states over time and the observations are the sequence of states. The assumption is that the state at the next step is dependent only on the current state. In many applications these states are not observable and the observations are instead outputs from another stochastic process which is dependent on the state of the unobservable process. These models are called hidden markov models (HMMs). This paper will provide a theoretical background for discrete-time, finite-state HMMs starting in ordinary markov chains. It will also answer questions on how to infer information about the hidden process and how to predict future distributions. It ends with simulations and a real data example where the covered material is put into use. Examples are also provided throughout the paper. The simulations showed that local maxima of the likelihood can be detected through assigning implausible starting values for estimation algorithms and that the precision of global decoding increase with smaller overlapping of the density/mass of the state dependent variables.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:andre.inge@hotmail.com . Supervisor: Mehrdad Jafari Mamaghani.

Sammanfattning

Markovkedjor beskriver stokastiska overgångar mellan tillstånd över tid och observationerna motsvaras av en serie tillstånd. Antagandet är att tillståndet efter nästa steg enbart beror på det nuvarande tillståndet. I många fall kan dock dessa tillstånd inte direkt observeras och observationerna kommer istället från en annan stokastisk process vars fördelning beror på tillståndet i den för oss gömda processen. Sådana modeller kallas for Hidden Markov models (HMMs). Denna uppsats tillhandahåller en teoretisk genomgång for HMMs i diskret tid och med ett begränsat antal tillstånd och tar sin start vid teori för vanliga Markovkedjor. Den kommer också att svara på frågor om hur man kan härleda fram information om den gömda processen och hur man kan prediktera framtida fördelningar. Uppsatsen avslutas med simuleringar och ett exempel med riktig data där vi använder teorin. Exempel finns genomgående i texten.

Simuleringarna visade att lokala maxpunkter av likelihoodfunktionen kan upptäckas genom att tilldela mindre troliga startvärden till den använda skattnings-algoritmen och att precisionen vid global dechiffrering ökar med minskat snitt av täthets/sannolikhets-funktionerna i de tillståndsberoende variablerna.

Acknowledgements

I would like to thank my supervisor Mehrdad Jafari Mamaghani for introducing me to the subject of Hidden Markov Models and for all help during this work. Also, a great thank you to the taxpayers of Sweden who paid for my education.

Contents

1 Introduction

A Markov chain is a stochastic process where the underlying mechanism is transitions between states [3]. It is based on the assumption that the value of the process (the state) at the next step is dependent only on the value at the time (a formal definition will be given in chapter 2). It could be used for modeling the future value of an asset, analyzing simple board games, or simply the probability of a certain event occurring when to process moves to the next stage.

In many situations though, even though one might be interested in what state a Markov chain is in, that variable cannot be observed. Instead the only observable information connected with that process is another stochastic variable which distribution depends on the state of the hidden process. Such models are called Hidden Markov models. The term hidden refers to the fact that the Markov chain driving the process is not visible to the observer, rather we observe emissions from a random variable connected with the current state.

As an example to illustrate this principle, imagine that no records of the historical weather were available. Suppose that the weather on a certain day could be either rainy or sunny and that the weather on the next day depends only on the weather on that particular day. You find a diary from a person and in it you read that on this specific day this person enjoyed a cool beverage at the beach. Even though the text mentions nothing about the weather, a not too far fetched conclusion would be that in betting whether it was a sunny or a rainy day the former would pay off better.

The state, here being sunny or rainy, is hidden from us in the sense that we cannot for sure know what the true weather was. For simplicity, assume that a person on a given day could either enjoy a cool beverage at the beach or stay home and read a book. The emission variable here would thus give one of the two as output. Given that we know that the person did the first we are lead to believe that the most probable weather was sunny.

In a sense Hidden Markov models are thus a sort of a statistician detective's work in that one draws conclusions about the most probable event based on observable information closely connected to the event.

Hidden Markov models have over the last decades become a highly useful tool for a growing number of engineering applications. One of the most prominent is different kinds of recognition and it is widely used in such for speech, writing etc. Chances are that when you use some sort of voice control, a Hidden Markov model is behind the result. Other areas are bioin-

formatics where it can be used for DNA decoding and in economics for modeling financial time series [1].

This paper will only regard finite, discrete time Markov chains. The structure will closley follow the book Hidden Markov Models for Time Series by Walter Zucchini and Iain L. MacDonlad and throughout we will borrow noations and equations from it.

2 Markov Chains

2.1 Definition and Chapman-Kolmogorovs Equations

This chapter will provide the necessary mathematics behind Markov chains needed to properly define and analyzing Hidden Markov models. It will emphasize properties associated with the type of Hidden Markov models that this paper covers, leaving some features less connected to the same out. We will end this chapter with an example where we put to use the covered material. We begin by defining a discrete-time Markov chain. Let

$$\{C_t : t \in \mathbf{N}\}$$

be a sequence of discrete random variables. It is said to be a discrete time Markov chain if for all

$$t \in \mathbf{N} \qquad P(C_{t+1}|C_t, ..., C_1) = P(C_{t+1}|C_t)$$

This is called the Markov property. In words it says that the distribution at time $t+1$ depends on the history of the process only through the value at time t. At each discrete time change, the process moves from one state to another or stays in the same state. Describing these events are the so called transition probabilities:

$$\gamma_{ij}(1) = P(C_{s+1} = j|C_s = i) \quad \text{for } i, j\text{=1,2,...,}m \text{ and } t \in \mathbf{N}$$

These denotes the probability that the process will in the next step move to state j from state i. For a chain consisting of m states these probabilities can be summarized into the so called transition probability matrix which has the form

$$\mathbf{\Gamma}(1) = \begin{pmatrix} \gamma_{11}(1) & \gamma_{12}(1) & ... & \gamma_{1m}(1) \\ . & & & \\ . & & & \\ . & & & \\ \gamma_{m1}(1) & \gamma_{m2}(1) & ... & \gamma_{mm}(1) \end{pmatrix}$$

The argument one refers to that this is the matrix consisting of the one step transition probabilities. Since the rows in $\mathbf{\Gamma}$ are probability distributions they must all sum to 1

$$\sum_{j=1}^{m} \gamma_{ij}(1) = 1 \quad \text{for } i\text{=1,2...m}$$

Expressed in terms of $\mathbf{\Gamma}$ this is equivalent to stating that the row vector $\mathbf{1}'$ is a right eigenvector of $\mathbf{\Gamma}$ with the eigenvalue 1 and we will henceforth use

this notation for similar cases [1].

We now define the t-step transition probabilities

$$\gamma_{ij}(t) = P(C_{s+t} = j | C_s = i) \quad t \geq 0, i, j = 1, 2, ...m$$

These describes the probability that the process in t-steps moves from state i from state j. We can now extend the one step probability matrix $\mathbf{\Gamma}(1)$ and let $\mathbf{\Gamma}(t)$ be the matrix consisting of the elements $\gamma_{ij}(t)$. Computation of these probabilities are given by the Chapman-Kolmogorov equations:

$$\mathbf{\Gamma}(t+u) = \mathbf{\Gamma}(t)\mathbf{\Gamma}(u)$$

Which further implies that for all $t \in \mathbf{N}$,

$$\mathbf{\Gamma}(t) = \mathbf{\Gamma}(1)^t \tag{1}$$

In words, matrix containing the t-step probabilities is attained through the t th power of the one step probability matrix.

2.2 Basic Properties

We will here briefly cover some basic properties of Markov chains.

Communication of States
Two states i and j are said to communicate if it is possible to go from i to j and j to i. Expressed in terms of the transition probabilities we formally write this as $\gamma_{ij}(t) > 0$ and $\gamma_{ji}(t) > 0$ for some $t \geq 1$.

Class
If two states communicate we say that they belong to the same class and further if the Markov chain consists of only one class it is said to be irreducible. In this paper we will only concern ourselves with irreducible Markov chains

Reccurent and Transient States
For state i let g_i be the probability that the process will ever re-enter i given that it started in i. If $g_i = 1$ we call state i recurrent and else we call it transient. If state i is recurrent it follows fairly easy that given that the process starts in that state it will reenter the same infinitly many times as $t \to \infty$. Similarly if state i is transient there will be a positiv probability $(1 - g_i)$ that it will never again enter i when in that state. A reccurent state i is said to be possitive reccurent if the expected time until the process returns to i starting in i is finite and for a finite-state Markov chain all reccurent states are positive reccurent. Reccurence is a class propterty which means that if i is reccurent and communicates with j (i.e. they belong to the same class),

then j is also reccurent. A convinent way to check if a state i is reccurent is if [2]

$$\sum_{t=1}^{\infty} \gamma_{ii}(t) = \infty \tag{2}$$

and if the same sum is less than ∞ it is transient.

Periodicity
We call a state i *periodic* with period k if, starting in i it can only return to i in multiples of k time steps. If the period of state i is 1 we call the state aperiodic which implies that returns to i can occur at irregular times. Periodicity is a class property which means that if state i has the period k then all other states that communicates with i also has the period k. If the Markov chain is irreducible this implies that all states share the same period since all states communicates with one another.

Ergodic states
A state i is called ergodic if it is positive reccurent and aperiodic.

Uncontitional Probabilities
Before moving on to stationary distributions we shall define another important feature. The unconditional probabilities, which describes the probability of a Markov chain being in a certain state at a given time t. We denote these $P(C_t = j)$. For $j = 1, 2...m$ we can then create the row vector

$$\mathbf{u}(t) = (P(C_t = 1), ...P(C_t = m)), \quad t \in \mathbf{N} \tag{3}$$

for t=1 we call $\mathbf{u}$ the initial distribution. If this i known we can now compute the distrubtion at time $t = 2$ through $\mathbf{u}(2) = \mathbf{u}(1)\mathbf{\Gamma}$ and indeed the following holds:

$$\mathbf{u}(t+1) = \mathbf{u}(t)\mathbf{\Gamma} \tag{4}$$

2.3 Stationary Distributions

Consider the transition probability matrix described in (1). What will happen to $\mathbf{\Gamma}(t)$ as t grows large? It holds that for an irreducible ergodic Markov chain there exists a unique limit distribution equal to the rows in $\mathbf{\Gamma}(t)$ as t grows large. This distribution is what we call the stationary distribution.
Formally we say that, for a Markov chain with the above stated properties, the row vector $\boldsymbol{\delta}$ is the stationary distribution if

$$\boldsymbol{\delta}\mathbf{\Gamma} = \boldsymbol{\delta} \quad \text{and} \quad \boldsymbol{\delta}\mathbf{1}' = 1 \tag{5}$$

At the end of this chapter we will present an example of a Markov chain and compute its stationary distribution.

From (5) we can then conclude that a Markov chain starting in its stationary

distribution will at all subsequent time points have the same distribution i.e. the stationary dito and we define such a process as a stationary Markov chain. In other words de define a stationary Markov chain as having the property that the initial distribution $\mathbf{u}(1)$ is indeed $\boldsymbol{\delta}$ [1].

2.4 Estimation of the Transition Probabilities

There are a number of ways to estimate the transition probabilities and we will only concern ourselves with one which is a very straight forward way given a sequens of observations. This is also indeed the maximum likelihood estimate [1].

Imagine that we want to model a phenomenon using a 2 state Markov chain. We call the states 1 and 2. Typically a sequens of observations will have the form (11121122211112111122) where each number corresponds to the process being in either state 1 or 2 at different times and the vector being ordered with respect to time (here from t=1 to t=20).
We can directly see that given a 1 another 1 followed in 9 cases. We donote this number f_{11} and similarly we see that $f_{12} = 4$, f_{21}= 3 and f_{22}=3 and we combine this into the matrix:

$$\mathbf{F} = \begin{pmatrix} 9 & 4 \\ 3 & 3 \end{pmatrix}$$

The number of transitions from state 1 is the sum of the elements in row 1 and in the same way for the number of transitions from state 2 which is the sum of the elements in row 2.
A natural way of estimating the transition probabilities if therefore:

$$\hat{\gamma}_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}} \tag{6}$$

where m is the number of states.

2.5 An Illustrative Example

We will here give an example of a Markov chain. We will use most of the definitions in previous sections to illustrate them however the example is fictional and the data made up.

Assume that the value of an asset at the end of a trading day could be either low, average or high. To model this with a Markov chain we assume according to the Markov property that the value on a following day depends only on the value at that certain day and not the value of all days leading up to that. This will then be a 3 state Markov chain and we name the states low, average and high, 1 2 and 3.

The data will be collected and summarized into a vector consisting of the numbers 1, 2 and 3 as in the example from the previous section. Using (6) we estimated the transition probabilities and summarized them into the transition probability matrix

$$\mathbf{\Gamma}(1) = \begin{pmatrix} 0.67 & 0.22 & 0.11 \\ 0.50 & 0.30 & 0.20 \\ 0.09 & 0.31 & 0.60 \end{pmatrix}$$

Examining the above matrix we can directly see that all states communicate and so the chain only has one class and is thus irreducible. Knowing that we can use equation (2) and conclude that it is also reccurent (*see element (1,1) in stationary distribution below, the sum in* (2) *is here* ∞) and since it is a finite-state chain it also follows that it is positive reccurent. Since it is possible to enter any state from any other state at all times we also conclude that the chain is aperiodic. Thus our Markov chain is ergodic.

Assume now that the day we start ($t = 1$), the value is high. The initial distribution ($\mathbf{u}(1)$), is then (1,0,0). Using equation (4) we see that the distribution for day 2 is given by

$$\mathbf{u}(2) = \mathbf{u}(1)\mathbf{\Gamma} = (0.67, 0.22, 0.11)$$

and further

$$\mathbf{u}(3) = \mathbf{u}(2)\mathbf{\Gamma} = (0.5688, 0.2475, 0.1837)$$

So the the probability of being in state 1 in 3 days is 0.5688. Now note that

$$\mathbf{\Gamma}(2) = \begin{pmatrix} 0.5688 & 0.2475 & 0.1837 \\ 0.5030 & 0.2620 & 0.2350 \\ 0.2693 & 0.2988 & 0.4319 \end{pmatrix}$$

The distribution for day 3 given that we started in state 1 is the first row in $\mathbf{\Gamma}(2)$. Had the initial distribution been (0,1,0) i.e. starting the process in state 2 it would have instead being row 2 that gave the distribution for day 3.

Now what happens to the rows in $\mathbf{\Gamma}(t)$ as t grows? We have already referred to the stationary distribution for convenience when stating that the chain was reccurent however that could have been shown without knowing it but we conclude that a stationary distribution must exist since the chain is irreducible and ergodic. Using computer software we can calculate any power of $\mathbf{\Gamma}(\mathbf{1})$ and by some computations we see that

$$\boldsymbol{\Gamma}(26) = \boldsymbol{\Gamma}(27) = \begin{pmatrix} 0.4727825 & 0.2648016 & 0.2624159 \\ 0.4727825 & 0.2648016 & 0.2624159 \\ 0.4727825 & 0.2648016 & 0.2624159 \end{pmatrix}$$

The rows here form the stationary distribution which we called $\boldsymbol{\delta}$ and we can directly see that both conditions in (5) are satisfied.

3 Hidden Markov Models

3.1 A General Approach

In the introductory chapter we mentioned some basics for Hidden Markov models to give a sense of the structure. We will in this section, before moving on to defining the models mathematically, aim to give a more solid idea through a general discussion. We will do so mainly using examples which we will return to in later sections of this chapter and in later chapters.

Hidden Markov models are a form of Dynamic Bayesian Network which are types of model used to describe conditional dependencies of a set of random variables. More precisely, the prefix dynamic refers to Baysian networks for sequences of variables i.e development over time.

Although Hidden Markov models and similar models go by different names such as Hidden Markov Processes, Markov-dependent mixtures or Markov-switching models, sometimes depending on the applications and sometimes on the author, we will only refer to them as Hidden Markov models [1].

Hidden Markov models (HMM) are models where the distribution of the output variables, or emission variables, are dependent on the state a Markov process that cannot be observed directly. We will refer to these distributions as the state dependent distributions. As a first example consider a 2-state Markov chain described in the previous chapter. Whenever the process enters a state we observe an outcome from a stochastic variable whose distribution depends on whether the process is in state 1 or 2. Let the output be either A or B for both states and let $P_1(A)$ and $P_1(B)$ be the probability distribution when the process is in state 1 and $P_2(A)$ and $P_2(B)$ be the same in state 2 and of course $\sum_{i=1}^{2} P_j(i) = 1$ for $i = A, B$ and $j = 1, 2$. The state dependent variables are here Bernoulli distributed with different probabilities depending on the state in which the process is in. This could in some sense be considered the simplest form of a Hidden Markov model and we refer to it as a Bernoulli-HMM. A sequens of observations would then typically have the appearance of a vector consisting of A and B for example $\mathbf{X} = (AABAABA)$.

As a second example, consider a phenomenon that produces an output at discrete times. We know from theory that this phenomenon over time moves between periods of high and low activity such that in the former the output results in high values and the latter low. We cannot observe in what state (high or low) the process is in but rather just observe the output of the process. A typical observation sequens could then look something like this $\mathbf{X} = (23, 21, 24, 12, 11, 26, 24, 9, 9, 7)$. If we were to ignore that the process

could be in different states we could calculate the mean of $\mathbf{X}$ (or whatever information we are after) and be done with it. However given the theory we could instead try to fit this into a model that takes into account that the observations could belong to a period of high respective low activity which would give us two sets from which we could calculate two different means each belonging to each state. Further we could also be interested in the transition probabilities describing the transitions between the two states.
If the state dependent variables in this example are Poisson distributed we call this a Poisson-HMM and we shall later examine this model by simulations.

There are numerous questions that emerges directly in regards to these two examples. Given an observation sequence $\mathbf{X}$:

- What state sequence is most likely to have produced $\mathbf{X}$?
- What is the most probable state the process is in at time t given the history up to t?
- What values of the transition probabilities and the parameters of the emission variables fits the data best?
- Given estimates of the parameters, what can we say about future states and distributions?

We shall try to answer these questions and illustrate them using simulated data in the coming chapters. The following sections of this chapter will aim at mathematically define and desribe the models as such.

3.2 Definition

For a Hidden Markov model $\{X_t : t \in \mathbf{N}\}$ we denote the history up to t with $\mathbf{X}^{(t)}$ and $\mathbf{C}^{(t)}$. The first is the history of the observable variabels and the second the unobservable Markov chain. This model can then be summarized into the two parts:

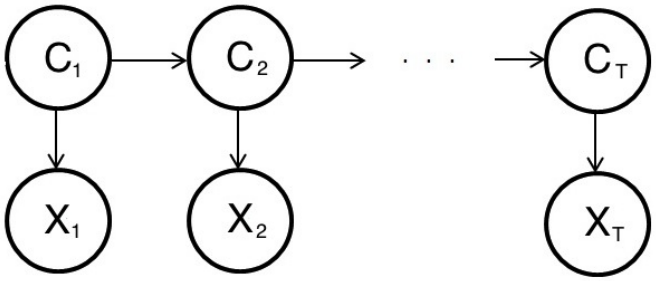
$$P(C_t|\mathbf{C}^{(t-1)}) = P(C_t|C_{t-1}) \quad t = 2, 3, ... \tag{7}$$

$$P(X_t|\mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = P(X_t|C_t) \quad t \in \mathbf{N} \tag{8}$$

The first expression describes a process satisfying the Markov property described in chapter 2. This is the unobserved process. The second expression describes the process $\{X_t : t = 1, 2, ...\}$ and from it, it is clear that the distribution of X_t depends solely on the current state C_t and is thus independent of earlier observations and states. It is worth stating that a HMM itself is not by necessity a Markov process [1] so in general

$$P(X_t|\mathbf{X}^{(t-1)}) \neq P(X_t|X_{t-1})$$

Which can be proved with simple counter examples, the term Markov in HMM simply refers to the unobserved process satisfying the Markov property. A HMM consisting of m hidden states is refered to as an m-state HMM. The structure of the model is illustrated in the graph below, a so called Trellis Diagram.



We now introduce some notations. For a specific HMM $\mathbf{\Gamma}$ is the transition probability matrix consisting of the elements γ_{ij} for $i, j = 1, 2...m$ describing the transitions between states in the unobservable Markov chain.

The state dependent distributions we denote $p_i(x) = P(X_t = x|C_t = i)$ for $i = 1, ..., m$ in the discrete case. This is the probability mass function of X_t when the process is in state i. If the state dependent variables are continous then $p_i(x)$ is the density function of $X_t|C_t = i$. We can conveniently sum up these in matrix form as

$$\mathbf{P}(x) = \begin{pmatrix} p_1(x) & & & 0 \\ & . & & \\ & & . & \\ & & & . \\ 0 & & & p_m(x) \end{pmatrix}$$

At last we define $u_i(t) = P(C_t = i)$ as the probability that the Markov chain at time t is in state i and we create the vector $\mathbf{u}(t) = (u_1(t), ..., u_m(t))$. For $t = 1$, $\mathbf{u}$ is the initial distribution of the Markov chain and we will denote this as $\boldsymbol{\delta}$.

Every unique HMM is hence determined by these three entities.

- The transition probability matrix $\mathbf{\Gamma}$
- The state dependent distributions $\mathbf{P}(x)$
- The initial distribution $\boldsymbol{\delta}$

3.3 The Joint Probability Mass Function and Likelihood

For a set of variables the joint probability mass function of $(X_1, ..., X_T, C_1, ..., C_T)$ $= (\mathbf{X}^{(T)}, \mathbf{C}^{(T)})$ is given by

$$P(\mathbf{X}^{(T)}, \mathbf{C}^{(T)}) = P(C_1)P(X_1|C_1)\prod_{k=2}^{T} P(C_k|C_{k-1})P(X_k|C_k) \tag{9}$$

Using the notations from the previous section, the likelihood function in matrix form is hence L_T given by

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}' \tag{10}$$

We will now give an example on how direct computation of the likelihood can be used given a specified model and a set of two observations.

Example 3.1 Consider the following model.

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{pmatrix}$$

$$\delta = (0.4, 0.6) \quad p_i(x) \in Po(2i) \text{ for } i\text{=1,2}$$

$\boldsymbol{\delta}$ is here the initial distribution as well as the stationary distribution. Using (10), the likelihood would then be $\boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\mathbf{1}'$. Which can be expressed as

$$\sum_{i=1}^{2}\sum_{j=1}^{2} \delta_i p_i(x_1)\gamma_{ij}p_j(x_2) \tag{11}$$

Say now that we have the observations $x_1 = 1$ and $x_2 = 5$ and want to know what state sequence maximizes the likelihood, that is what combination of i and j maximizes the expression in the dubbel-sum in (11)? For $i = 1$ and $j = 1$ we get that $\delta_1 = 0.4$, $p_1(1) = 2e^{-2}$, $\gamma_{11} = 0.25$, $p_1(5) = \frac{2^5 e^{-2}}{5!}$ and the product is 0.00098. The table below shows the computaions over all i:s and j:s. and from it wee se that the answer to the question is the state sequence

i	j	δ_i	$p_i(1)$	γ_{ij}	$p_j(5)$	product
1	1	0.4	0.271	0.25	0.036	0.00098
1	2	0.4	0.271	0.75	0.156	0.01268
2	1	0.6	0.073	0.5	0.036	0.00079
2	2	0.6	0.073	0.5	0.156	0.00342

(1,2). This should not come as a surprise given the transition probabilities and the fact that $P(p_1(x) \leq 4) = 0.9473$. What we did here was a so called global decoding and we will return this in chapter 4.

Looking at the table again we see that each of the 4 terms (the elements last column) consists of 4 factors, in fact for a m-state HMM with T observations, computations of the likelihood will consist of a sum of m^T terms each with $2T$ factors.

3.4 The Forward and Backward probabilites

In chapter 4 we will discuss methods through which we can answer the questions stated in the beginning on this chapter. Before we do so we must introduce two more features, the forward and backward probabilities which are used to estimate unknown parameters as well as for decoding.

Forward Probabilities
For a set of T observations we define the vector of forward probabilities $\boldsymbol{\alpha}_t$ as

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad t = 1, 2, ..., T \tag{12}$$

where $\boldsymbol{\delta}$ is the initial distribution of the Markov chain. The elements in $\boldsymbol{\alpha}_t$ are what we call the forward probabilities. If $t = T$ then the sum of the elements in $\boldsymbol{\alpha}_t$ is the likelihood.

It holds that the j th element in $\boldsymbol{\alpha}_t$ has the joint probability

$$\alpha_t(j) = P(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j) \quad \text{for j=1,...,m} \tag{13}$$

Backward Probabilities
For a set of T observations we define the vector of backward probabilities $\boldsymbol{\beta}'_t$ as

$$\boldsymbol{\beta}'_t = \boldsymbol{\Gamma}\mathbf{P}(x_{t+1})\boldsymbol{\Gamma}\mathbf{P}(x_{t+2})\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}' \quad t\text{=1,2,...,}T \tag{14}$$

It then holds for the j th element in $\boldsymbol{\beta}'_t$ that

$$\beta_t(j) = P(X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, ..., X_T = x_T | C_t = j) \tag{15}$$

While the forward probabilities are joint probailities the backward probabilities are conditional ones, the conditional probability of $\mathbf{X}$ from t+1 up to T given that we at time t are in state j. Obviously $\beta_T(j) = 1$ for all $j = 1, ..., m$.

Combining the forward and backward probabilities we get the following results which will be needed in the coming chapter. For a proof we refer to Zucchini & MacDonald.

Proposition 1

- $\alpha_t(j)\beta_t(j) = P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = j) \quad t = 1, ..., T$

- $\boldsymbol{\alpha}_t\boldsymbol{\beta}_t' = P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = L_T$
- $P(C_t = j|\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \alpha_t(j)\beta_t(j)/L_T \qquad t = 1, ..., T$
- $P(C_{t-1} = j, C_t = k|\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \alpha_{t-1}(j)\gamma_{jk}p_k(x_t)\beta_t(k)/L_T \qquad t$=2,...,$T$

In short, the first says that the joint probability of $\mathbf{X}$ and $C = j$ at time t is attained through the product of the forward and backward probability at time t. The second follows as a consequens of the first. The third and fourth describe conditional probabilities of $\mathbf{C}$ given the history of $\mathbf{X}$. For more information about the properties of these equations we again refer to Zucchini & MacDonald.

Example 3.2 Consider again the model in example (3.1). We will now compute $\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}_2$ when $\mathbf{X} = (2, 2, 4)$.

$$\boldsymbol{\alpha}_2 = \begin{pmatrix} 0.4 & 0.6 \end{pmatrix} \begin{pmatrix} 0.2707 & 0 \\ 0 & 0.1465 \end{pmatrix} \begin{pmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} 0.2707 & 0 \\ 0 & 0.1465 \end{pmatrix}$$

$$\boldsymbol{\beta}_2' = \begin{pmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} 0.0902 & 0 \\ 0 & 0.1954 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

We get that $\boldsymbol{\alpha}_2 = (0.01922427, 0.01833889)$ and $\boldsymbol{\beta}_2 = (0.1690810, 0.1427952)$ from which we conclude that

$$P(X_1 = 2, X_2 = 2, X_3 = 4, C_2 = j) = (0.003250458, 0.002618705) \quad \text{for } j = 1, 2$$

4 Parameter Estimation and Inference

In this chapter we will discuss methods of estimating the parameters in a HMM through the Baum Welch algorithm. We also discuss how to draw conclusions about the hidden state at various times and the full state sequence of states given observations. We begin with an algorithm for estimating the parameters based on the EM-algorithm called the Baum-Welch algorithm.

4.1 The EM and Baum-Welch Algorithm

The Baum-Welch algorithm, named after Leonard E. Baum and Lloyd R. Welch, is one of the most common ways to estimate unknown parameters in a HMM using only the observed data as training [1]. It is based on the expectation maximization algorithm which is among other things can be used for maximizing the likelihood when some data is missing which in the case of HMMs corresponds to the hidden states of the Markov chain. It consists of two steps. The E-step calculates the expectation of the missing data given the observations and the current estimation of the parameters and the M-step maximizes that function with respect to the parameters. This procedure is repeated until the changes in the estimates are smaller than some predetermined threshold. In the context of HMMs the expectation maximization algorithm is knows as the Baum-Welch algorithm and it uses the forward and backward probabilities described in the previous chapter [1].

When applying the EM-algorithm the log-likelihood function is referred to as the incomplete log-likelihood. Incomplete since we are missing the values of the Markov process. In contrast we call the complete log-likelihood the log-likelihood of the data if we instead could see hidden data. The reason is that the former could be somewhat hard to maximize [7]. We therefore define the complete log-likelihood as

$$\log\left(P(\mathbf{x}^{(T)}, \mathbf{c}^{(T)})\right) = \log\left(\delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1},c_t} \prod_{t=1}^{T} p_{c_t}(x_t)\right) \tag{16}$$

Where $\boldsymbol{\delta}$ is the initial distribution of C_1. Since $c_1, ..., c_T$ is missing we somehow need to replace them. We therefore introduce the two variables:

- $u_j(t) = 1$ iff $c_t = j$ otherwise 0 $\qquad t = 2, ..., T$
- $v_{jk}(t) = 1$ iff $c_{t-1} = j$ and $c_t = k$ otherwise 0 $\qquad t = 1, ..., T$

Exanding the expression in (16) we get

$$\log\left(P(\mathbf{x}^{(T)}, \mathbf{c}^{(T)})\right) = \log \delta_{c_1} + \sum_{t=2}^{T} \log \gamma_{c_{t-1},c_t} + \sum_{t=1}^{T} \log p_{c_t}(x_t) \tag{17}$$

and with our new variables we can express this as

$$\sum_{j=1}^{m} u_j(1) \log \delta_j + \sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{t=2}^{T} v_{jk}(t) \right) \log \gamma_{jk} + \sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log p_j(x_t) \tag{18}$$

The complete-data log-likelihood is thus made up of three terms where the first depends only on the intial distribution, the second only on the transitions probabilities and the third only on the state dependet distributions and each is maximized with respect to its parameters.

To do so we need an expression for $u_j(t)$ and $v_{jk}(t)$ which can be achived by the E-step. We simply replace $u_j(t)$ and $v_{jk}(t)$ with their conditional expectations given $\mathbf{x}^{(T)}$ which are the third and fourth equation in proposition 1 in chapter 3 so that

- $\hat{u}_j(t) = P(C_t = j|\mathbf{x}^{(T)}) = \alpha_t(j)\beta_t(j)/L_T$
- $\hat{v}_{jk}(t) = P(C_{t-1} = j, C_t = k|\mathbf{x}^{(T)}) = \alpha_{t-1}(j)\gamma_{jk}p_k(x_t)\beta_t(k)/L_T$

These estimates are based on the current parameter estimates. When this step is done, (18) is maximized with respect to the three sets of parameters that makes up a unique HMM, the initial disitribution $\boldsymbol{\delta}$, the transition probability matrix $\mathbf{\Gamma}$ and the parameters of the state dependent distributions $p_j(x)$. The new parameter estimates are then used in the E-step again and repetition of this procedure is done until desired convergence at which point the value of the parameters will be at a stationary point of the likelihood of the observed data [1]. This point however is not a guaranteed global maximum and there is no known way to ascertain such point. In the next chapter we will simulate data and try different starting vaules for the paramters to explore this. We will end this section with an example on how to perform the maximization step for the parameters of the state conditional distribution when these are distributed according to an exponential distribution.

Example 4.1 Let $p_j(x) \in Exp(\lambda_j)$ so that $p_j(x) = \frac{1}{\lambda_j} exp(\frac{-x}{\lambda_j})$. The third term in (18), $\sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log p_j(x_t)$ then becomes for any $j = 1, ...m$

$$\sum_{t=1}^{T} \hat{u}_j(t) \log \left(\frac{1}{\lambda_j} e^{\frac{-x}{\lambda_j}} \right) = - \left(\hat{u}_j(1) \frac{x_1}{\lambda_j} + \hat{u}_j(1) \log(\lambda_j) + ... + \hat{u}_j(T) \frac{x_T}{\lambda_j} + \hat{u}_j(T) \log(\lambda_j) \right)$$

which upon differentiation becomes

$$\left(\frac{\hat{u}_j(1)x_1}{\lambda_j^2} + ... + \frac{\hat{u}_j(T)x_T}{\lambda_j^2} \right) - \left(\frac{\hat{u}_j(1)}{\lambda_j} + ... + \frac{\hat{u}_j(T)}{\lambda_j} \right) = \frac{1}{\lambda_j^2} \sum_{t=1}^{T} \hat{u}_j(t)x_t - \frac{1}{\lambda_j} \sum_{t=1}^{T} \hat{u}_j(t)$$

and this expression set to zero yields

$$\hat{\lambda}_j = \frac{\sum_{t=1}^{T} \hat{u}_j(t)x(t)}{\sum_{t=1}^{T} \hat{u}_j(t)} \tag{19}$$

Also this expression has the limit zero as $\lambda_j \to \infty$ but that is clearly not a solution here.

4.2 Decoding

Given a model with estimated parameters and a sequence of observations we shall now see how one can go about to infer information about the hidden states. Although there are several questions that can be answered we will only focus on two. The most likely state at a given time (local decoding) and the most likely state sequence (global decoding) given a set of observations.

4.2.1 Local Decoding

The goal here is to for a given time t find the most likely state of the Markov chain, that is we need $P(C_t = i|\mathbf{X}^{(T)} = \mathbf{x}^{(T)})$. From the first and second equation in Propostition 1 we had that $P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i) = \alpha_t(i)\beta_t(i)$ and $P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = L_T$. Using elementary probability theory we can directly see that

$$P(C_t = i|\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i)}{P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)})} = \frac{\alpha_t(i)\beta_t(i)}{L_T} \quad \text{for } i = 1, ..., m \tag{20}$$

So at time t, the most likely state i is the one that maximizes the above expression.

Example 4.2
Using the model and the observations in example 3.1 we have that $\boldsymbol{\alpha}_2 = (0.01922427, 0.01833889)$ and $\boldsymbol{\beta}_2 = (0.1690810, 0.1427952)$. The full likelihood of the observations $L_3 = \boldsymbol{\alpha}_2\boldsymbol{\beta}_2'$ is 0.005869163 which gives us

i	$P(C_2 = i\|\mathbf{X}^{(T)} = \mathbf{x}^{(T)})$
1	0.5538197
2	0.4461803

and so we conclude that the most likely state at time $t = 2$ is 1.

4.2.2 Global Decoding

Often one is not merely interested in the state at one perticular time but rather what sequence of states is most likely to have produced a sequence of observations (this is the case in for example speech recognition [1] where

the hidden states corresponds to the abstract syllable and the observations to the spoken sound) . The task is to find the state sequence $c_1, ..., c_T$ that maximizes the conditional probability

$$P(\mathbf{C}^{(T)} = \mathbf{c}^{(T)} | \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) \tag{21}$$

This undertaking is called global decoding and even though the results are often smiliar to local decoding they are not by necessity identical [1]. Returning again to example 3.1 we saw an occurence of this. There we had a model with 2 states and 2 observations. This gave us a sum of 4 terms each with 4 factors and in general the computation of the likelihood over all combination of states consists of a sum of m^T terms each with $2T$ factors and obviously this makes direct computations unfeasible [1]. The solution is a case of dynamic programming algorithm known as The Viterbi Algoritm [3]. Though we will use this in the next chapter we will not discuss the detalis of this but instead refer to Zucchini & MacDonald or Stark & Woods for a theoretical explanation.

4.3 Predicting Future States and Distributions

In many applications one might not only be interested in decoding past states but rather make predictions about the future given the history of the process. We will cover two aspects of this which we will use to predict the future rate of inflation in Sweden. These two aspects are the most likely state h steps after T and the distribution of $\mathbf{X}$ h steps after T.

4.3.1 Predicting Future States

The task is to find $P(C_{T+h} = i | \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$. Consider first the case where $h = 1$. Looking at (20) we see that when $t = T$ the right hand side becomes $\frac{\alpha_T(i)}{L_T}$ since $\boldsymbol{\beta}_T = 1$. So the distribution of C at time T is

$$P(C_T = i | \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{\alpha_T(i)}{L_T}$$

Going back to the theory of Markov chains the distribution at the next step is the above expression multiplied with the transition probability matrix. We can write this as

$$P(C_{T+1} = i | \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{\boldsymbol{\alpha}_T \boldsymbol{\Gamma}(, i)}{L_T}$$

where $\boldsymbol{\Gamma}(, i)$ is the i th column of $\boldsymbol{\Gamma}$. Expanding this to h steps and we have the following

$$P(C_{T+h} = i | \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{\boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h(, i)}{L_T} \tag{22}$$

As $h \to \infty$ the right hand side converges towards the stationary distribution of the Markov chain indicating that if the initial distribution is in fact the stationary the ride hand side of (22) is always the stationary distribution. This is also clear if we interpret the stationary distribution as the percent of time spent in each state in the long run.

4.3.2 Predicting Future Distributions of X, (Forecast)

In many cases one is more interested in the output variables rather than the states themselves though the states are driving the evolution of the process. For example one could model the volatility of stock returns and want to predict the value at some point in the future where the states corresponds to periods of high and low volatility [4]. One way to go about it would be to predict the most probable state at that time and use the estimated state dependent variable for that state. The downside to this is that the state dependent variables could differ a lot. For example if we have two Gaussian distributed variables with the same variance but the means 1 and 100 and we predict that the most likely state is state 1 with probabiility 0.51 it is not very reasonable to be use the mean 1 over the mean 100. Instead it would be more wise to use a weighted average based on the distribution of predicted states. Moreover if one were to only use the most probable state one would not be using all information avaiable. Instead we calculate the vector of probabilities for each state at time $T+h$ and use those as weights. We get

$$P(X_{T+h} = x|\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{\boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h \mathbf{P}(x)\mathbf{1}'}{L_T} \tag{23}$$

Let

$$\frac{\boldsymbol{\alpha}_T}{\boldsymbol{\alpha}_T \mathbf{1}'} = \boldsymbol{\phi}_T$$

and we get

$$P(X_{T+h} = x|\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \boldsymbol{\phi}_T \boldsymbol{\Gamma}^h \mathbf{P}(x)\mathbf{1}'$$

The vector of weights for the h th step after T if then

$$\boldsymbol{\phi}_T \boldsymbol{\Gamma}^h \tag{24}$$

Equation 24 is the vector of probabilities of the Markov chain being in different states at h steps after T so the idea is simply to use those as weights on our state dependent variables to predict future distributions.

4.4 Choosing Model

As in many applications in statistics the range of possible models is vast. In the case of HMM we have to decide how many states the Markov chain

should have and what state conditional distributions fit the data best. One problem that could easily arrise is overparameterization. In the case of a multiple linear regression model for example, the R^2 always increases with the addition of another explanatory variable but at the cost of higher model complexity. In the case of HMMs the fit of the model similarly increases with the addition of another state but at the expense of a quadratic increase in the number of parameters [1]. Of course in some applications we know from theory the exact number of possible states but in many cases this number is rather arbitrary. We could flip two unfair coins with a Markov chain driving which coin to flip next and either pat our friend on the head or pull his tail depending on whether we ended up with heads or tails. If our friend (who cannot see the coin) would want to model this using a HMM he would know that the only possible number of states is 2 (coin 1 and coin 2). In another case we could for example consider an economy and divide its current condition into the states of boom and recession. However we could also add a state called depression to distinguish periods of moderate recession from periods of very high dito. We end up having to ask ourselves what number of states is the optimal? Many suggestions have been made and this question is by no mean settled [1]. We will in the coming chapter for our simulations use the Akaike information criterion which is defined as

$$AIC = -2 \log L + 2p$$

and the Bayesian information criterion which is defined as

$$BIC = -2 \log L + p \log T$$

Where $\log L$ is the log-likelihood of the fitted model and p the number of parameters of the model. For both criterions, the 'best' model is the one that minimizes the information criterion. Adding a state will increase the number of parameters quadraticly with respect to the transition probabilities and linearly with state dependent variables which is why the order of the models with regards to the information criterion only depends on the former.

5 Simulation

In this chapter we will simulate data from specified models. We will then use a form of reverse engineering to try to fit this data into different models. We will focus manily on

- Choosing models using the AIC and BIC
- Examining parameter estimations with different starting values (the problem of local maximum of the likelihood)
- Decoding

We will always assume non-stationarity in our models. LL denotes the log-likelihood of of the observations using the estimated parameters from the Baum-Welch algoritm and n denotes the numbers of iterations before convergence.

5.1 A Poisson-HMM

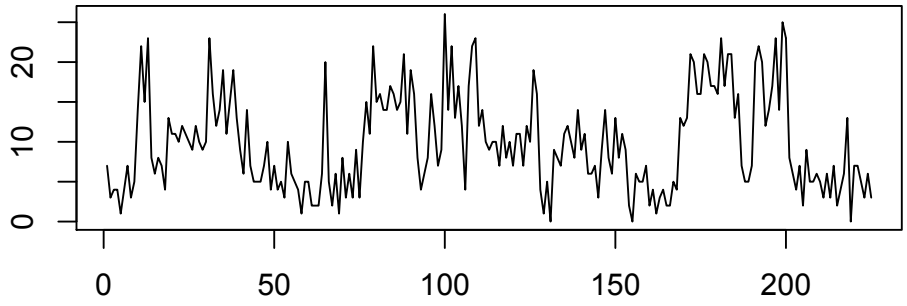
The Model
We will simulate 225 observations from the following model where $\boldsymbol{\delta}$ is the initial distribution.

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9 & 0.02 & 0.05 & 0.03 \\ 0.10 & 0.75 & 0.06 & 0.09 \\ 0.05 & 0.10 & 0.55 & 0.30 \\ 0.10 & 0.10 & 0.65 & 0.15 \end{pmatrix} \qquad \mathbf{P}(x) = \begin{pmatrix} p_1(x) & 0 & 0 & 0 \\ 0 & p_2(x) & 0 & 0 \\ 0 & 0 & p_3(x) & 0 \\ 0 & 0 & 0 & p_4(x) \end{pmatrix}$$

$p_1(x) \in Po(5) \quad p_2(x) \in Po(10) \quad p_3(x) \in Po(15) \quad p_4(x) \in Po(20)$

$\boldsymbol{\delta} = (0.2, 0.4, 0.1, 0.3)$

The simulated process is shown in the graph below.



We will now try to fit this data into 4 different models with 2, 3, 4 and

5 states. As initial values we will for the transition probability matrix use uniformly distributed rows and. The starting values of $\boldsymbol{\lambda}$ (parameters of the Poisson distributions) and the initial distribution we will allow to depend on the number of states. In the estimated matrices ϵ is a number such that $0 < \epsilon < 3.74 \cdot 10^{-7}$. The reason for using this number instead of just round of to 0 is that the latter would imply an impossibility of transitions from that i to j whereas this allows for this but with a very low probability.

2 states
Starting values: $\boldsymbol{\lambda} = (5, 15)$

$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.8857 & 0.1143 \\ 0.1051 & 0.8949 \end{pmatrix}$, $\hat{\boldsymbol{\delta}} = (1, 0)$, $\hat{\boldsymbol{\lambda}} = (5.0032, 14.3895)$, $LL = -661.8274$, $n = 15$

3 states
Starting values: $\boldsymbol{\lambda} = (5, 10, 15)$.

$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.9116 & 0.0229 & 0.0655 \\ 0.0393 & 0.8984 & 0.0623 \\ 0.0736 & 0.0673 & 0.8591 \end{pmatrix}$, $\hat{\boldsymbol{\delta}} = (1, 0, 0)$, $\hat{\boldsymbol{\lambda}} = (4.6064, 9.3283, 17.0601)$, $LL = -630.2404$, $n = 27$

4 states
Starting values: $\boldsymbol{\lambda} = (3, 9, 16, 21)$

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.90368 & 0.01945 & 0.07687 & \epsilon \\ 0.0358 & 0.9073 & \epsilon & 0.0569 \\ 0.17462 & 0.05102 & 0.00001 & 0.77435 \\ \epsilon & 0.07194 & 0.75248 & 0.17558 \end{pmatrix}$$

$\hat{\boldsymbol{\delta}} = (1, 0, 0, 0)$, $\hat{\boldsymbol{\lambda}} = (4.5401, 9.3051, 15.3738, 18.4456)$, $LL = -626.3862$, $n = 99$

5 states
Starting values: $\boldsymbol{\lambda} = (3, 8, 12, 16, 21)$

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} \epsilon & 0.9292 & 0.0577 & 0.0131 & \epsilon \\ 0.9015 & \epsilon & \epsilon & 0.0985 & \epsilon \\ \epsilon & 0.0414 & 0.8931 & \epsilon & 0.0655 \\ \epsilon & 0.1658 & 0.0043 & \epsilon & 0.8299 \\ \epsilon & \epsilon & 0.1142 & 0.5655 & 0.3203 \end{pmatrix}$$

$\hat{\boldsymbol{\delta}} = (0, 1, 0, 0, 0)$, $\hat{\boldsymbol{\lambda}} = (3.4838, 5.6538, 9.3633, 15.0641, 18.4124)$, $LL = -618.4638$, $n = 214$

To decide which model is preferable we summarize the results in the table below.

m	p	$-LL$	AIC	BIC
2	4	661.827	1331.655	1345.318
3	9	630.2404	**1278.481**	**1309.226**
4	16	626.3862	1284.772	1339.43
5	25	618.4638	1286.928	1372.33

As expected we see that the likelihood increases with the number of states or equivalent, the negative log-likelihood decreases. But, taking into account the number of parameters, we find that both the AIC and BIC select the model with 3 states. For this data we thus choose that model.
We cannot be sure however that the estimates in our model as well as in the rejected models really correspond to a global maximum of the likelihood function. One way to at least increase our confidence is to try different starting values for the Baum-Welch algorithm but of course, given the number of possible values this is absolutely no guarante. What is certain though is that if two different set of starting values returns two different values of the likelihood then the one with the smaller is not a global maximum. For this simulation, different values were tried and all led to the same estimates.

Having chosen our model it could be interesting to determine som features of the Hidden Markov chain based on the estimation.

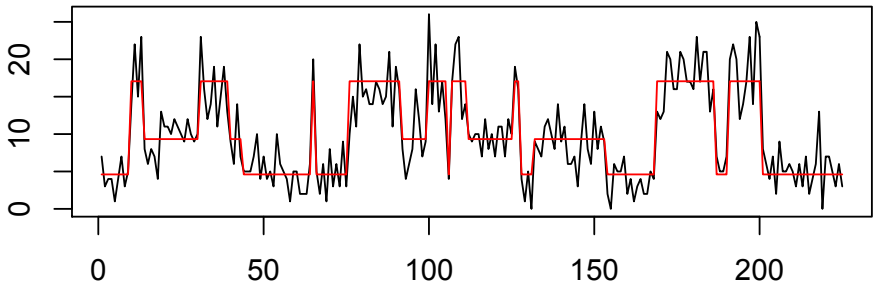
$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.9116 & 0.0229 & 0.0655 \\ 0.0393 & 0.8984 & 0.0623 \\ 0.0736 & 0.0673 & 0.8591 \end{pmatrix}$$

is obviously irreducible since all states communicate. Further it is aperiodic since all states can be reached from all other states at all times. It has a stationary distribution which we compute by

$$\lim_{t\to+\infty} \hat{\mathbf{\Gamma}}(t) = \begin{pmatrix} 0.3917604 & 0.2954757 & 0.3127639 \\ 0.3917604 & 0.2954757 & 0.3127639 \\ 0.3917604 & 0.2954757 & 0.3127639 \end{pmatrix}$$

and so we conclude that it is also positive reccurent and hence ergodic and its stationary distribution is equal to rows in the above matrix.

For our model the graph below shows the most likely state sequens according to the Viterbi algorithm where the red line is λ_i for state i.



Conclusion

Though the data was simulated from a 4 state HMM we were led to choose a model with 3 states given our four model-setups and the information criterion . We found that our estimate of $\mathbf{\Gamma}$ was ergodic and we illustrated the most likely state sequence using the Viterbi algorithm.

5.2 2 Gaussian HMMs

We will here use two different Gaussian HMMs. In the first one we will explore the problem of local maximum of the likelihood function and in the second we will simulate from a model with a small overlapping of the density functions in the state dependent variables.

5.2.1 Model 1

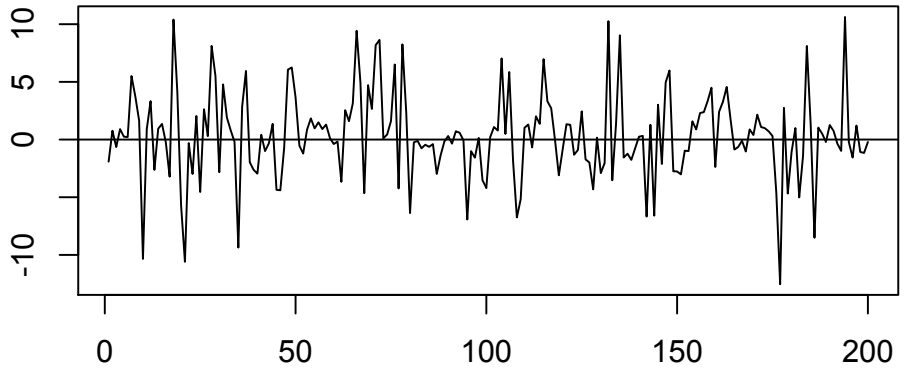
For this simulation we will assume that we have reason to believe that a 2 state model is the only possible model and we will thus try to fit the data into a model with two states. The simulation comes from the below specified model and the plot shows the observations.

$$\mathbf{\Gamma} = \begin{pmatrix} 0.8 & 0.20 \\ 0.25 & 0.75 \end{pmatrix}, p_1(x) \in N(0,1), p_2(x) \in N(0,4), \boldsymbol{\delta}=(0.8,0.2)$$

We will now try out different values of parameters of the state dependent variables for fixed transition and initial probabilities. We therefore let

$$\mathbf{\Gamma} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \boldsymbol{\delta} = (0.25, 0.75)$$

We now assume that we have no knowledge of the model that generated these observations except for the number of states and that the state dependent variables are assumed to be normally distributed. If we relax that for a moment one, from looking at the graph might want to suggest that 3 states with the means low, around zero and high, would result in a good



fit. It looks as if once the process is in the high state it tends to quickly move to the low state and once there quickly return to the high state and occationally visit the zero state for a period of time. But since we know that the number of states can only be 2 we should instead be led to belive that the low and high states are actually the same state and that the quick transitions between them instead reflects a high variance of that very same state. We should therefore believe that we are dealing with 2 variables that differ in variance. As for the means we notice that the sample mean is 0.3101073 and it looks as though the high and low valued observations are somewhat equally distributed around 0.

By this reasoning we choose as starting values $\boldsymbol{\mu} = (-1, 1)$ and $\boldsymbol{\sigma} = (0.5, 2)$. After 20 iterations the Baum-Welch algorithm returns the following estimates:

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.7716 & 0.2284 \\ 0.1354 & 0.8646 \end{pmatrix}, \hat{\boldsymbol{\delta}} = (0.9999999, 0.0000001)$$

$$\hat{\boldsymbol{\mu}} = (0.1344, 0.4162), \hat{\boldsymbol{\sigma}} = (1.0255, 4.6851)$$

and LL=-521.4165768 and so the likelihood is 3.561705e-227.
Other values close to these where tried and all resulted in the same estimates.

Now let's try a set of values for which there is no reason to belive to be true based on the graph and the data. Let $\boldsymbol{\mu} = (10, 1)$ and $\boldsymbol{\sigma} = (10, 10)$. Running this through the Baum-Welch algorithm we get the following after 38 iterations:

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.8645 & 0.1355 \\ 0.2284 & 0.7716 \end{pmatrix}, \hat{\boldsymbol{\delta}} = (0, 1)$$

$$\hat{\boldsymbol{\mu}} = (0.4164, 0.1341), \hat{\boldsymbol{\sigma}} = (4.6856, 1.0260)$$

and LL=-521.4165773 and so the likelihood is 3.561703e-227.

Looking at these two estimates it seems as if state 1 in the first corresponds to state 2 in the second such that all elements in the estimated matrices and vectors have changed place. The slight differences in the estimates comes from the threshold for convergence. If that was set to a smaller quantity the two models would be nearly identical since the order of the terms in the likelihood function does not change its structure.

Now let's try $\boldsymbol{\mu} = (1, 1)$ and $\boldsymbol{\sigma} = (1, 0.05)$ as starting values which makes no sense given a quick look at the graph. Convergence was reached after 32 iterations and the estimates are

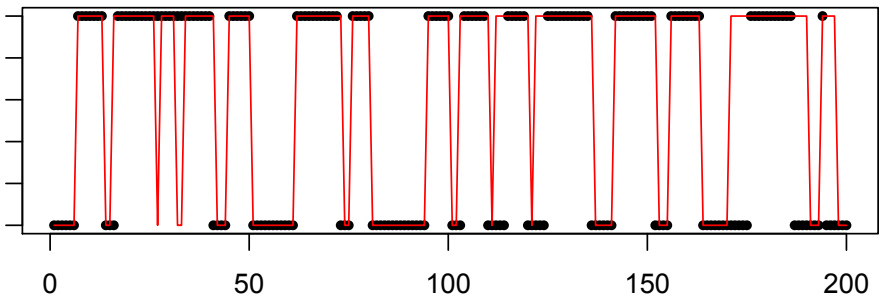
$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.9453 & 0.0547 \\ 0.9999 & 0.0001 \end{pmatrix}, \hat{\boldsymbol{\delta}} = (1, 0)$$

$$\hat{\boldsymbol{\mu}} = (0.2781, 0.8973), \hat{\boldsymbol{\sigma}} = (3.8535, 0.0914)$$

With a LL of -543.1396356 and thus the likelihood value 1.310549e-236 $<$ 3.561705e-227 we conclude that this point is no global maximum of the likelihood function and so we discard these estimates.

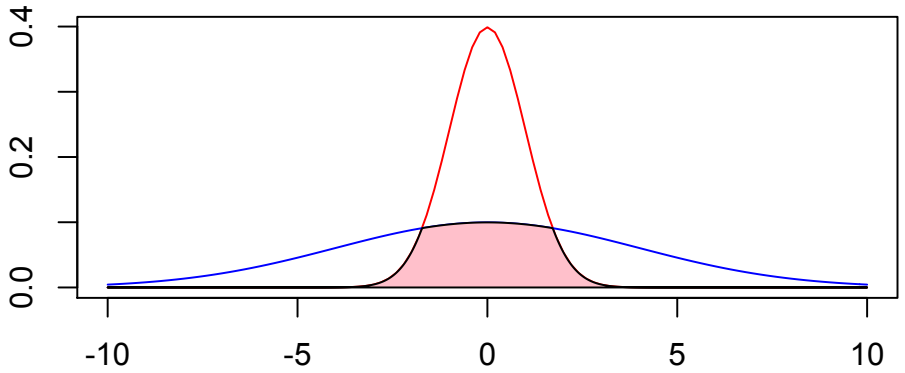
We also identify another local maximum point with $LL = -545.0098570$ using the starting values $\boldsymbol{\mu} = (-10, 10)$ and $\boldsymbol{\sigma} = (1, 10)$.

Even though we cannot guarantee a global maximum we strongly belive that the first suggested model is the best and we therefore choose that model. Based on this model the most likely state sequens as determined by the Viterbi algorithm is shown below in the graph below. The black dots represents the Viterbi path and the red line is the actual state



We see that in 171 out of 200 cases the Viterbi algorithm suggested the actual state of the Markov chain. Intuitively it seems reasonable to suggest that the number of 'correct' states will increase with smaller overlapping of the density functions of the state dependent variables. If the sample spaces

are disjoint the Viterbi algorithm will always suggest the actual state and further it will render the HMM a Markov process as defined by the Markov property [1]. In this simulation the overlapping is not negligible, in the next simulation we will use a model with a much smaller overlapping. The overlapping of the two density functions from which the data was simulated are show in the figure below.



Conclusion

We saw that using different starting values we reached different maximum points of the likelihood function. We discarded all suggested models which had a smaller value than any other. From our investigation it seemed the best model came when assigning plausible starting values with respect to the observed data as supposed to values that does not reflect the data. The viterbi algorithm found the 'correct' state in 85.5 % of the cases.

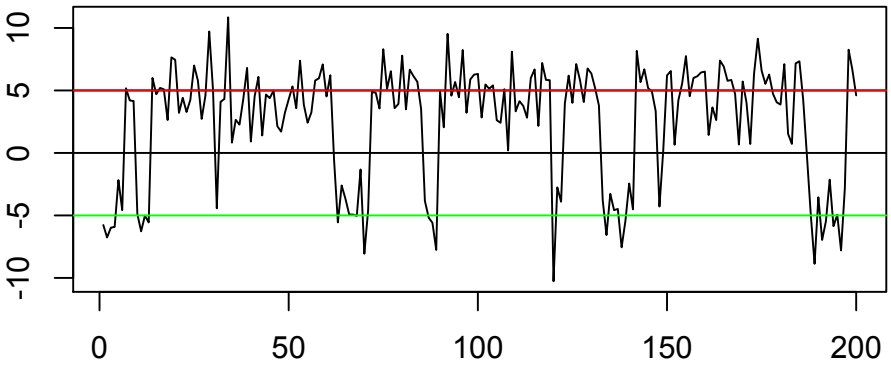
5.2.2 Model 2

We will here make the same assumption of 2 states as in the previous simulation. The data is simulated from the following model and the graph shows the plotted data.

$$\mathbf{\Gamma} = \begin{pmatrix} 0.95 & 0.05 \\ 0.10 & 0.90 \end{pmatrix}, p_1(x) \in N(5,2), p_2(x) \in N(-5,2), \boldsymbol{\delta}=(0.10,0.90)$$

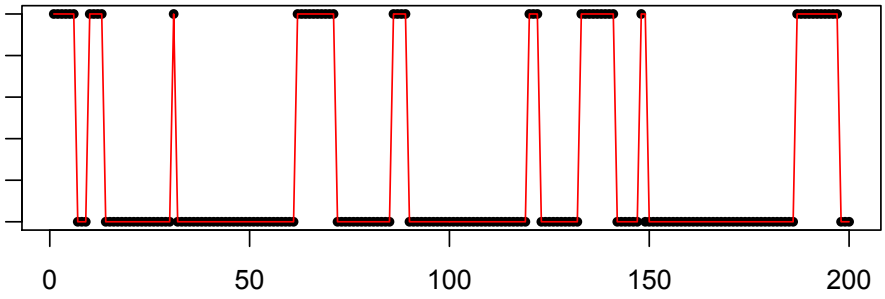
After careful investigation to find and discard local maxima we end up with the following parameter estimations

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.9467 & 0.0533 \\ 0.1850 & 0.8150 \end{pmatrix}, \hat{\boldsymbol{\delta}} = (0,1)$$

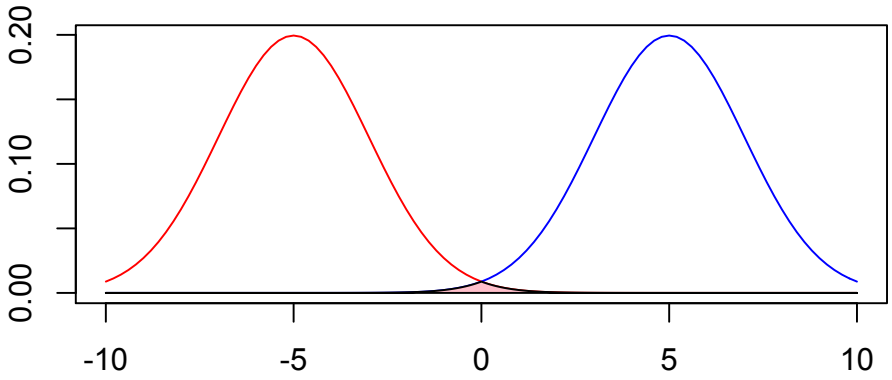


$$\hat{\boldsymbol{\mu}} = (4.9145, -4.8870), \hat{\boldsymbol{\sigma}} = (2.0391, 1.9829)$$

We run the Viterbi algorithm with these parameters and we the results are illustrated in the graph below with the red line again illustrating the actual state.



The Viterbi algorithm found the actual state in 199 out of 200 times. Observation 149 was 0.1906369, the actual state was 2 but the Viterbi algorithm suggested state 1.
The figure below shows the overlapping of the density functions from which this data was simulated. The overlapping is far smaller than in the case in Model 1.



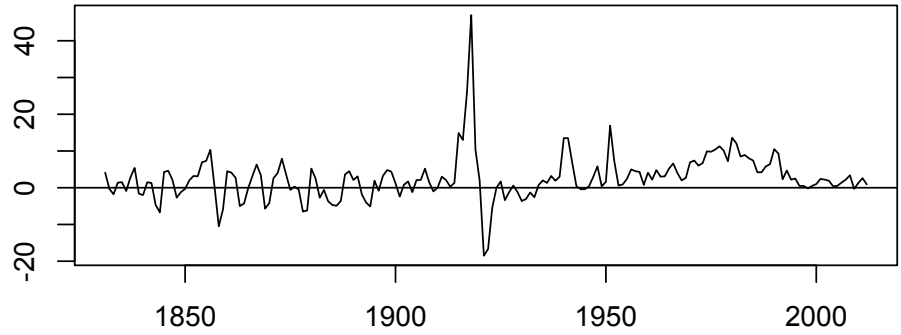
Conclusion

This data was simulated from a model where the overlapping of the density functions of the state dependent variables was far smaller than in the previous. As expected the Viterbi algorithm found the actual state more often than in Model 1.

6 Modeling Inflation

6.1 The Model and Analysis

We will here use historical data of the yearly inflation for Sweden from the period 1831 - 2012. The data was collected from SCB. The goal is to make a forecast of the inflation for the year 2013. For this we assume that the inflation each year is normally distributed and further in accordance with the assumptions for Hidden Markov models that the inflation the next year is only dependent on the inflation the current year. The graph below shows the inflation against time.



Models with 2,3,4,5 and 6 states were tried and the AIC selected the model with 5 states. After running the Baum-Welch algorithm we got the estimates shown below. The weights are calculated using the output values from R where the elements with ϵ are their respective real value (i.e as before very close to 0) but to make the reading lucid we replace these small values with ϵ below. We also round off to 4 decimals in the text for the same reason.

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.4818 & 0.1586 & 0.3596 & \epsilon & \epsilon \\ 0.0669 & 0.7307 & 0.1701 & \epsilon & 0.0323 \\ 0.0317 & 0.2151 & 0.6642 & 0.0890 & \epsilon \\ \epsilon & 0.1199 & 0.0481 & 0.8320 & \epsilon \\ 0.2310 & \epsilon & \epsilon & \epsilon & 0.7690 \end{pmatrix}$$

$\hat{\boldsymbol{\mu}}$=$(-4.9851, 0.1966, 3.3749, 8.8025, 7.2510)$

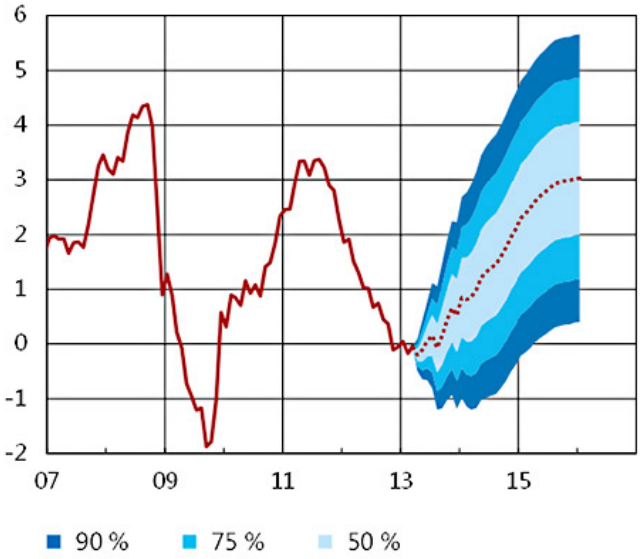
$\hat{\boldsymbol{\sigma}}$=$(0.9761, 1.6408, 1.5138, 2.9868, 19.7776)$

We calculate
$\boldsymbol{\phi}_T$ =(ϵ, 0.7777, 0.2180, 0.0014, 0.0029)

We can now calculate the weights for any h by $\boldsymbol{\phi}_T\mathbf{\Gamma}^h$. For $h = 1$ we get the vector of weights (0.0596, 0.6153, 0.2772, 0.0206, 0.0273). This is also the probabilities of the Markov chain being in the different states at 2013

and we see that the most probable state is state 2 followed by 3 and a low probability for the other 3.

Taking the scalar product between this vector and $\hat{\boldsymbol{\mu}}$ we get the mean for our 2013 forecast distribution. Since the state dependent variables are assumed to be independent the variance of our distribution is the scalar product of the weight vector with its elements squared and $\hat{\boldsymbol{\sigma}}^2$ (the vector of the estimated variances). We get that the forecast distribution for the inflation for year 2013 is $N(1.1388,\ 1.2226)$. The graph below shows the central banks own calculations for the expected inflation (years on the x-axis) and confidence limits in color. The figure below was taken from www.riksbanken.se



Our forecast distribution is based on the yearly avarage and seems to be slightly higher than the graph but still within a reasonably close range.

6.2 Discussion

Looking at the graph we see that one period which clearly distinguishes itself from other periods. The years following World War 1 Sweden experienced a period of very high inflation. This period is likely what caused the AIC to suggest 5 states. The number of observations from that state is likely to be small, causing the variance to be high which we can see in the estimates. Shortly after the outbreak of World War 2 the inflation was again high during a short period of time. During the 1970s Sweden again experienced high

inflation. Periods of deflation occured often in the 19th century and later became more rare. Since 1993 the Swedish central bank has a set out target of holding the inflation at a rate of 2 ± 1 [5]. The policies regarding inflation has thus varied over time as economists have suggested different methods for dealing with it. These changes in policies raises questions about whether modeling such long time series with HMMs where much of the data reflects changes in policies is really suitable. A perhaps better way would be to use data starting at a later period in history. Particularly common feature of the 19th century with rapid changes between deflation and inflation is somewhat exiled to history and it is very questionable if those changes should really have impact on models for modern day inflation. One could also suggest that removing outliers such as the period during World War 1.

The data used is yearly and we would probably benefit from instead using quarterly or even monthly data since inflation often varies within a given year. With yearly data we therefore risk to miss out on a lot of valuable information. With smaller time intervals the observations would be more recent with the same number of observations which also has its benefits. Further the forecasts would be for a more near future instead of as now a yearly average. Our model does not in any way consider other variables that could have possible influence or work as an indicator. We are solely using the rate of infaltion history to predict the future which is not the normal case when predicting future rate of inflation. Further we assume a Gaussian distribution which we have not motivated closer.

We assumed in accordance with the theory of Markov chains that the state of inflation in the next year only depends on the state the current year. This is somewhat of a sloppy assumption and maybe a higher order Markov chain (where the next step depends on several previous steps) would be preferable. But, since this paper is not mainly about modeling inflation but rather cover the theoretical aspects of the models, a full investigation of the different possibilities covered in this discussion will have to wait for another time.

7 Conclusion

In this paper we have discussed the mathematical theory behind ordinary Markov chains and Hidden Markov models. We have provided examples throughout and used simulated data to illustrate the main principle and how the most common inference related issues are solved. Our simulations were mainly focused on two problems. Choosing the best model given several alternatives and the task of detecting local maxima of the likelihood where we gave empirical examples of how these can occur when assigning implausible starting values for the estimation algorithm. We also showed empirically that smaller overlapping of the density functions resulted in higher precision of the Viterbi algorithm. We ended with a real-data example where we modeled the rate of inflation of Sweden to predict future distributions. In our discussion section of that chapter we argued that due to the way the data was constructed as well as the nature of the phenomena our model has obvious flaws and we suggested possible ways to improve the model.

In our quest to detect local maxima we used different starting values of the state dependent variables with the same starting values of the transition probabilities and the initial distribution. Another possible approach would be to vary all paramters simultaneously or the opposite, try different values of the transition probabilities given the same state dependent parameter values.

All our models state dependent variables followed the same distribution but with different parameters. For future work it would be interesting to mix distributions and find applications where that would be preferable. It would also be interesting to further investigate the theoretical basis and extend to higher level Markov chains and Markov chains in continuous time. This did however not fit into the timeframe of this paper.

8 References

[1] Walter Zucchini, Iain L. MacDonlad. (2009). *Hidden Markov Models for Time Series and introduction Using R.* Chapman and Hall/CRC

[2] Sheldon M. Ross. (2010). *Introduction to Probability Models 10th edition* Academic Press.

[3] Henry Stark, John W. Woods. (2002). *Probability and Random Processes with applications to Signal Processing.* Prentice Hall.

[4] Ramaprasad Bahr, Shigeyuki Hamori. (2004). *Hidden Markov Models, Applications to Financial Economics.* Kluwer Academic Publishers.

Internet

[5] http://www.riksbank.se/sv/Penningpolitik/Inflation/Inflationsmalet/. 010513

[6] http://www.riksbank.se/sv/Penningpolitik/Prognoser-och-rantebeslut/Aktuell-prognos-for-reporanta-inflation-och-BNP/. 010513

[7] http://jmlr.csail.mit.edu/papers/volume1/meila00a/html/node12.html 160513