Hidden Markov Models: An Approach to Sequence Analysis in Population Studies

Danilo Bolano

National Center of Competence in Research LIVES
Institute for Demographic and Life Course Studies
University of Geneva, Switzerland
Danilo.Bolano@unige.ch

PPA 2014 Annual Meeting

Abstract. In this paper we provide an extensive overview of Hidden Markov models for longitudinal data. It is a stochastic model used to describe the evolution of observable events that depends on internal factors which are not directly observable. We will illustrate the general version of the model and the estimation procedures and the more interesting extensions for social sciences like the inclusion of covariates, the Mixture Transition Distribution model for high-order Markov Models, the Double Chain Markov Model and so on. Some empirical examples from life course perspective have been provided.

1 Introduction

Over the last decades we have observed a widespread diffusion of retrospective population surveys and large scale panel studies like the Panel Study of Income Dynamics (36 waves until 2013), the Health Retirement Study in the US (20 waves), or administrative panel data like the Swiss Household Panel (13 waves), the US National Longitudinal Survey and so on.

The increasing complexity of the data hence the need for specific advanced statistical methods in social sciences. In longitudinal life course methodology, Billari (2001) distinguishes between two main approaches, the most popular event-based approach (event history) and what it calls the *holistic approach*. The first case is a generalization of the life table which aims at discovering a casual relationship with the focus on one given event and each individual is represented by a collection of time stamped events. The holistic approach however mainly relies on sequence analysis and the life trajectories are seen as a whole unit of interest.

Abbott, who introduced sequence analysis in social sciences in 1995, makes a distinction between the approaches that consider the entire sequence as a whole unit and "step-by-step" methods. In the former case the objective is to identify typical patterns using measures of dissimilarity or distance between individual trajectories (like the Optimal Matching distance). Markov chain models are instead classified as step-by-step method where "the central interest is a fairly deep and complex dependence of an -interval-measured sequence upon its own past" (p.104 Abbott, 1995). A Markovian process is an appropriate way to model a life course. Life trajectories can be seen as the result of a stochastic process in which the sequence of states (like employment or civic status, health condition etcetera) are linked by transition probabilities to move from one to another with some time dependences. Being in a certain state (e.g. condition) today, influence the probability of being in an other state tomorrow. In other words, in the Markovian framework life trajectories are seen as the result of a stochastic process in which the probability of occurrence of a particular state, or event, depends on its past.

Moreover as pointed out by Sutton (2006) the stochasticity plays a central role in population and life course studies. Individual events are stochastic by definitions, they are subject to random influences and unpredictable a priori. For that reasons, we are not interested in the point estimation for a particular individual but on the general tendency of the expected distributions of individual outcomes. According to Sutton, "the proper goal of sociological research, [...], is not to make bad predictions about individuals, but strong predictions about distributions" (Sutton, 2006, p.10). It is worth to note in fact that Markovian models are distributional stochastic models to the contrary of other well-known and widespread approaches that are essentially point processes (e.g. the ARMA and ARIMA models in time series

literature). Trying to predict the next value of a series, the advantage of the point approach is that the model will provide a clear answer consisting in one numerical value. The drawback is that in most cases the answer will be either inaccurate or totally wrong. Furthermore, in population studies the point prediction is quite useless. An adequate probabilistic model will not provide a single value but will lead to a complete representation of the possible futures through a (maybe multi-modal) distribution. In that sense, the answer given by this approach will generally have a higher probability to help taking the right decisions, because it shows all possibilities rather than only a (probably) wrong one.

Despite this is a quite intuitive way to analyze a life course and the great majority of social processes can be conducted to this framework Markovian models are still only sparsely used in population studies. This is unfortunate, since the current trend in social surveys is clearly to switch from cross-sectional to longitudinal surveys, hence the need for advanced modeling methods of such data.

According to Abbott (1995), the main reason for a limited uses in social sciences of such methods is related to the underlying hypothesis of stationary, that is quite rare in reality, and the involving in the analysis of just one previous time period. But Abbott in his discussion considers only the simplest case of a Markovian process: a stationary process with one time lag. It doest not take into account the opportunity to model non-stationary Markov process as well as the introduction of higher order Markov chain.

In this article we will focus on Hidden Markov Model pointing out the relevance of this approach for life course studies illustrating several examples. The Hidden Markov Model is a Markovian based model used when the observed data are influenced by an underlying latent process. Like other latent based models, HMM is particular suitable for analyzing life trajectories. The evolution of many aspects of a life course depends on internal factors or theoretical constructs that are not directly observable and can evolve overtime. For instance, health trajectories of an aged person may depend on his/her current unobserved level of frailty. HMM is a *parameter-driven model* (Cox, 1981) in which the observed outcomes are independent conditional on an underlying changing parameter process which follows a Markov chain. And the distribution of the observed outcomes is determined by the current state of the hidden process.

The use of hidden processes is also a way to relax the homogeneity assumption in markovian processes trying to capture the complexity of social behaviors and the latent variable may also have the role of accounting for the unobserved heterogeneity between individuals and it can be used both to explain the observed trajectories and for (probabilistic) clustering. We will illustrate how simply we can switch from the simplest case to more interesting extensions for social sciences like including covariates, modeling of higher order Markov process and its approximation (Berchtold and Raftery,

2002), Double Chain Markov Model (Berchtold, 2002) and so on. We will also illustrate how for a given research problem, the state space can be defined in different ways and different specifications of the model can be tested in order to find the ones that 'best' fit the data.

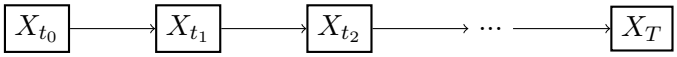
2 The general framework: Markovian models

A Markov chain is a stochastic process that models the serial dependence between adjacent periods (like the rings of a chain). Let consider a random variable X_t observed over time, $t = 0, 1, 2, \cdots, T$ [1]. Markovian models are used to model the probability of observing a certain modality of X_t given the modalities observed in the previous periods.

In particular, in its traditional formulation, a Markov chain is a memoryless process: the next modality, or states, depends only on the current one and not on the previous sequence of states. This is called *Markov property* and such models are also known as first-order Markov chain (Figure 1):

$$\begin{aligned} &Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots) \\ &\qquad = Pr(X_t = x_t | X_{t-1} = x_{t-1}) \quad t = 1, ..., T \end{aligned} \tag{2.1}$$

Figure 1: A graphical representation of a first order Markov chain



This assumption is often too simplistic and it can be replaced by an higher-order Markov process where the current state depends on multiple previous states. For instance, for an hidden chain of order two:

$$\begin{aligned} &Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots) \\ &\qquad = Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}) \quad t = 1, ..., T \end{aligned} \tag{2.2}$$

We will discuss of high-order Markov chains and their approximation in Section (6.3).

The probabilities $Pr(X_t|X_{t-1})$ (Equation 2.1) that govern the transitions between states, called *transition probability*, are denoted by $q(t)$ and they are represented in matrix form by the transition matrix $Q(t)$.

In literature using Markovian processes, the transition probabilities are often considered independent of t. It means that at any moment of the series the probability of switching from a given state to another is the same. In this case we have a **time-homogeneous markov process**.

[1] In this article we will consider only discrete-time Markov chain

For convenience, we consider X_t be a categorical variable taking value in the finite set $\{1, 2, \ldots, k\}$. In case of homogeneous Markov process the transition matrix, $Q(t)$, can be simply written as (k x k) matrix Q.

$$\begin{aligned} q_{ij}(t) &= Pr(X_t = j | X_{t-1} = i) \\ &= Pr(X_{t-1} = j | X_{t-2} = i) = q_{ij} \quad t = 1, ..., T \quad i, j = 1, ..., k \end{aligned} \tag{2.3}$$

The homogeneity assumption is justified in many applications but it has been often criticized in social sciences where, in particular analyzing long series like a life course of an individual, the hypothesis of time independence may be not realistic. For instance, the probability of losing a job during a working career: the transition probability between having a job and being unemployed may differ if the person is at beginning of his career or at an later stage of his working file. Even if this assumption is quite common used in behavioral and social studies, different non-homogeneous models have been developed. For instance, using Markov processes both at hidden and visible level as in the Double Chain Markov Model (Berchtold, 2002) briefly presented in Section (6.5)

Another relevant set of parameters that govern the transition probability between states, is the prior or initial probabilities $\boldsymbol{\pi}$:

$$\pi_i = Pr(X_{t_0} = x_{t_0}) \tag{2.4}$$

It indicates probability of having a given state i at first time point.

2.1 An example of how to use the transition probabilities with homogeneous Markov process

A way to show the flexibility of the markovian models and possible applications to population studies, is looking at the transition matrix.

We can consider for example a first-order transition matrix with three states. In the general formulation, the transition matrix is written as follows:

$$\mathbf{Q} = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} \\ q_{2,1} & q_{2,2} & q_{2,3} \\ q_{3,1} & q_{3,2} & q_{3,3} \end{pmatrix}$$

Each cell represents the probability to move from a specific state to the others. For instance, $q_{1,2}$ is the probability to move from state 1 to state 2 and so on. The cells on the main diagonal represent the probability of being in the same state for two consecutive periods. Being a series of probabilities the sum for each row must be equal to one $\sum_{j=1}^{k} q_{i,j} = 1$.

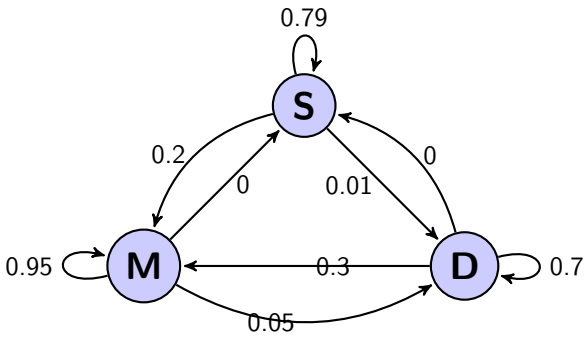
Suppose we want to analyze the changing in marital status into 2 consecutive year for a certain population. We have three possible modalities,

"single" (S), "married" (M) and "divorced/widow" (D). The transition matrix of a Markov chain of oder one, presented below, represents the "marital status transitions" between any two successive states[2]:

$$Q = \begin{array}{c|ccc} & & X_t & \\ X_{t-1} & \text{S} & \text{M} & \text{D} \\ \hline \text{S} & 0.79 & 0.20 & 0.01 \\ \text{M} & 0 & 0.95 & 0.05 \\ \text{D} & 0 & 0.30 & 0.70 \end{array}$$

Using a path diagram, the transition matrix can be represented as follows

Figure 2: A graphical representation of a first order transition matrix with three states



According to the research question and previous investigations, the researcher may be interested to put some constraints. If we expect that the different states are hierarchical, meaning that when a subject enters in a state s_i, he can only stay in this state or go to a state with number $s_j > s_i$. This might be the case when the phenomenon under study evolves in time with the age of the subjects or if we are analyzing "linear" events. A good example in health studies is hearing capabilities: young people are supposed to have maximal capabilities, and then these capabilities will decline with age. Such situations can be represented by the following matrix:

$$\mathbf{Q} = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} \\ 0 & q_{2,2} & q_{2,3} \\ 0 & 0 & 1 \end{pmatrix}$$

So, according to the specific research question, the research can define the model in an appropriate manner. From a practical point of view, this is

[2]This is just an illustrative example with simulated data

particularly easy. To impose the different constraints discussed above, it is sufficient to set to zero the required parameters.

When we consider a higher order Markov chain, the transition matrix can be more complex to represent including several elements set to zero corresponding to transitions that cannot occur (the so-called *structural zero*). For this reason it is possible to write the transition matrix in a compact way called *reduced form* and denoted by *RA*. See Berchtold and Raftery (2002) for more details. The reduced form of a transition matrix of a second order Markov chain with three possible states is

$$\text{RA} = \begin{array}{cc|ccc} & & X_t & & \\ X_{t-2} & X_{t-1} & 1 & 2 & 3 \\ \hline 1 & 1 & q_{111} & q_{112} & q_{113} \\ 2 & 1 & q_{211} & q_{212} & q_{213} \\ 3 & 1 & q_{311} & q_{312} & q_{313} \\ 1 & 2 & q_{121} & q_{122} & q_{123} \\ 2 & 2 & q_{221} & q_{222} & q_{223} \\ 3 & 2 & q_{321} & q_{322} & q_{323} \\ 1 & 3 & q_{131} & q_{132} & q_{133} \\ 2 & 3 & q_{231} & q_{232} & q_{233} \\ 3 & 3 & q_{331} & q_{332} & q_{333} \end{array}$$

Before to analyze in details the Hidden Markov Models, we will briefly recall some theoretical aspects of the finite mixture models. The HMM, in fact, can be also considered as a generalization of the mixture model. As we will show, in the mixture model the unobserved variables which control the mixture component are assumed independent to each other. If there is a dependency and its assumed to follow a Markov process, the mixture model become an Hidden Markov Model.

3 Hidden Markov Models

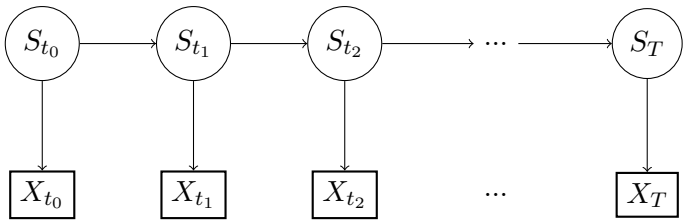
The Hidden Markov Model (HMM) is a Markovian process used to describe the evolution of observed events that are influenced by an underlying internal factor which is not directly observable that follows a Markov chain. Life course data can be exactly represented by this framework. They are longitudinal in their essence and life events of an individual can be represented by the sequence of symbols of the variable of interest with an underlying hidden construct. For instance, we can analyze the evolution of the health condition of an individual taking into account the evolution of its unobserved level of frailty.

Despite the limited use in demography and sociology, latent Markov models are widely used in biosciences and in some fields of social sciences

like behavioral and criminal studies. As in psychology to model learning process; in economics and finance where they are knows as regime switching models; for EEG analysis; in behavioral sciences, in genetics, to study biological sequences, DNA and protein modeling. In particular, there is a extensive literature in speech recognition since Baum and Petrie (1966). The application of HMM in several field is also due to its numerical and statistical properties: availability of all moments (mean, variance and so on), the likelihood is linear and easy to compute, it is possible to account for outliers and it can be used for forecasting and probabilistic clustering.

An HMM consists of two stochastic processes: an invisible process of hidden states and a visible process of observable symbols. The hidden process is assumed to follow a Markov chain and the observed sequence is considered as independent on the hidden states. In the hidden Markov model in fact (see for example Rabiner, 1989), at each time the state of the latent chain is unknown and there isn't a full identification between the state of the chain and the corresponding observed output [3]. The successive outputs of the observed variable are defined as conditionally independent because they are linked only indirectly through a latent Markov chain and the relationship between an unobserved state and the actual observations derives form a probability distribution. On other words, the observations are independent conditional on some unobserved parameter process (Cox, 1981) with distribution determined by the current state of the parameter process. Figure 3 is a path diagram used to represent a first order Hidden Markov chain.

Figure 3: A graphical representation of a first order Hidden Markov chain. S_t is the hidden state at time t, X_i is the observed random variable[5]



Then, the HMM consists of two main parts: a measurement model and a dynamic one. The measurement part models the relationship between the

[3]Please notice the terminology used. We call 'symbol' or 'event' the single observation and 'state' the invisible factor underlying the observation. Differently from the terminology used in classical sequence analysis where the event is time stamped and it defined by state changes.

[5]By convention, circles represent unobserved latent variable. The squares the observed variable

states of the hidden chain and the observations. The dynamic model, however, explains the dynamics (i.e. the transitions) between states overtime. If the states represent the construct of interest, an unobserved entity, the transition dynamics represent the changes in the construct.

Formally, a discrete time HMM consists of five elements:

1. a set $\mathbf{S}(t)$ of hidden states s_i, $i = \{1, \ldots, k\}$
2. a set $\mathbf{X}(t)$ of observed symbols
3. a matrix $Q(t)$ of transition probabilities to move from one state to another at each time point t, $t = 0, \ldots, T$. It is the analogous of the transition matrix presented in Section (2.1) for a Markov chain
4. matrix B of probabilities $p_i(x)$ of having the observation x being in the hidden state i
5. a vector π of initial probabilities

The first three points mean that the unobserved factor is a categorical variable which have k possible levels or states (Point 1). $X(t)$ is a random observed variable. And we have a certain probability $q_{ij}(t)$ of moving to state i to state j at time t (Point 3).

The simplest HMM can be summarized as follows:

$$Pr(S_t = i|S_0^{t-1} = s_0^{t-1}) = Pr(S_t = i|S_{t-1} = j) = q_{ij} \quad t = 1, ..., T \tag{3.1a}$$

$$Pr(X_t = x_t|X_0^{t-1} = x_0^{t-1}, S_0^t = s_0^t) = Pr(X_t = x_t|S_t = i) = p_i(x_t) \quad s = 1, ..., k \tag{3.1b}$$

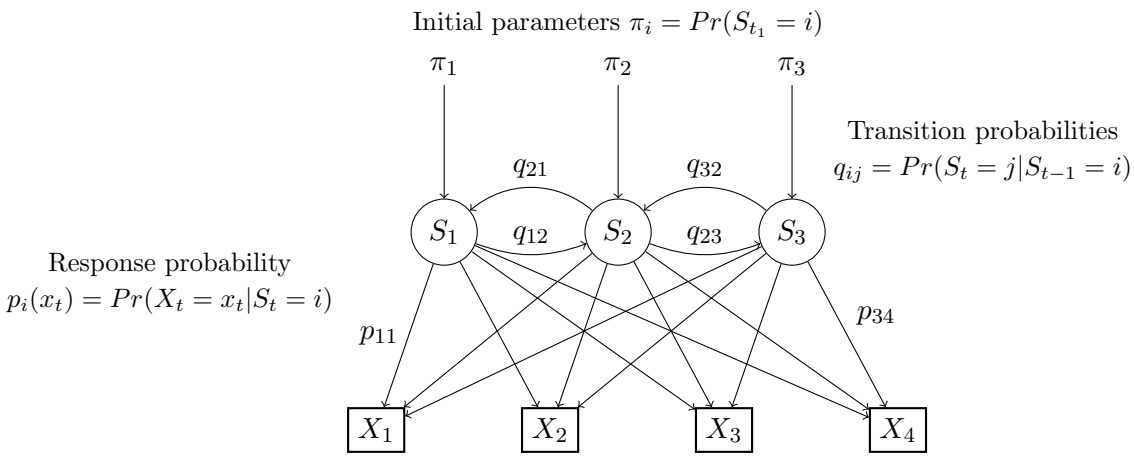
Equation (3.1a) represents the latent part of the model where an unobserved variable S_t follows a Markov property. So, the current state S_t depends only on the previous state S_{t-1} and not on the earlier periods. The second equation however, refers to the measurement part also known as *state-dependent process*. At time t when the hidden state is known ($S_t = s_t$), the probability distribution of X_t depends only on the current latent state and not on previous observations or on the previous states. For this reason, the observed process is called *conditionally independent*.

When the observed variable X_t is a categorical variable, $p_i(x)$ represented in Equation (3.1b) is the probability *mass function* of X_t if the latent chain is in state i in time t. The continuous case it is similar but we have to consider $p_i(x)$ as the probability *density* function of X_t when we have the state i at time t.

The k distributions p_i are called *state dependent distributions* of the model and they are represented in the vector π.

All set of parameters presented in this section are usually presented in the compact form $v = (Q, B, \pi)$. Graphically, the parameters of a HMM with hidden variable S_i with three possible states $i = \{1, 2, 3\}$ and an observed variable X_t, $t = \{1, 2, 3, 4\}$ is

Figure 4: A graphical representation of HMM parameters: 3 states



3.1 Marginal distributions and the likelihood

In the section, we will derive the distribution of X_t with an homogeneous Markov chain. For convenience, we will derive in case of discrete values but for continuous variables the derivation is similar.

Let X_t a discrete variable with $t = 1, \ldots, T$, defining $u_i(t) = Pr(S_t = i)$ the probability of being in state i at time t,

$$\begin{aligned} Pr(X_t = x_t) &= \sum_{j=1}^{k} Pr(S_t = i) Pr(X_t = x_t | S_t = j) \\ &= \sum_{i=1}^{k} u_i(t) p_j(x_t) \end{aligned} \tag{3.2}$$

This equation is straightforward from Figure 3.

The probability of a given observation x depends on the state of the hidden process. So, to calculate this probability we have to multiply the probability of being in a certain state j for the conditional probability of that particular state has determined the observation x (Equation 3.1b). Then, we have to sum over the all possible k hidden states.

Suppose as in the previous example to consider a latent process with three possible latent levels of frailty, $S = \{1, 2, 3\}$ and X the self reported health using a Likert Scale with five modalities ("Very bad", "Bad", "Average", "Good", "Very good"). The probability of observing: "be in good health" (G) can be calculated as follows

$$
\begin{aligned}
&Pr(X = G) = \\
&= Pr(S = 1)Pr(X = G|S = 1) + \\
&+ Pr(S = 2)Pr(X = G|S = 2) + Pr(S = 3)Pr(X = G|S = 3)
\end{aligned} \tag{3.3}
$$

Equation(3.2) can be easily extended to a sequence of observations $\{x_0, x_1, \ldots, T\}$. Considering that[6]

$$
Pr(X_t, X_{t+m}, S_t, S_{t+m}) = Pr(S_t)Pr(X_t|S_t)Pr(S_{t+m}|S_t)Pr(X_{t+m}|S_{t+m}) \quad m = 1, \ldots, T-t \tag{3.4}
$$

And then using also Equation (3.1b),

$$
\begin{aligned}
Pr(X_t = x_t, X_{t+m} = x_{t+m}) = \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} Pr(X_t = x_t, X_{t+m} = x_{t+m}, S_t = i, S_{t+m} = j) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} Pr(S_t = i)p_i(x_t)Pr(X_{t+m}|S_t = i)p_j(x_{t+m}) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} u_i(t)p_i(x_t)q_{ij}(m)p_j(x_{t+m}) \quad m = 1, \ldots, T-t
\end{aligned} \tag{3.5}
$$

In matrix form Equation(3.5) becomes

$$
Pr(X_t = x_t, X_{t+m} = x_{t+m}) = \mathbf{u}(t)\mathbf{P}(x_t)Q^k\mathbf{P}(x_{t+m})\mathbf{1'} \tag{3.6}
$$

Where $\mathbf{u}$(t) is a vector of the k probabilities $u_i(t)$, $\mathbf{P}$(t) a (k X k)diagonal matrix with diagonal elements the state-dependent probability (density) functions $p_i(x_t)$. Q is the transition probability matrix. See Zucchini and MacDonald (2009) for the proofs.

Then, the likelihood L_T of observing $\{x_0, \ldots, x_T\}$ with a k-states HMM and *initial distribution* π can be written as follows

$$
\begin{aligned}
L_T = Pr(X_0^T = x_0^T) &= \sum_{s=1}^{k} Pr(X_0^T = x_0^T, S_0^T = s_0^T) \\
&= \pi\mathbf{P}(x_1)Q\mathbf{P}(x_2)Q \ldots Q\mathbf{P}(x_T)\mathbf{1'}
\end{aligned} \tag{3.7}
$$

[6]For convenience we use a compact notation.

4 Computational methods: The estimation procedure

In practical situations, given the flexible but complex structure of the HMM, there are three fundamental issues to address:

- **Evaluation problem**. How well a given model describe the observed sequence of data? Or in other words, given a sequence of observation $X_0^T = \{x_0, x_1, x_2, \ldots, x_T\}$ and the model v, how do we efficiently compute $L(X_0^T|v)$.
- **Optimal state sequence**. Given the data and a model, how can we search for the optimal sequence of hidden states?
- **Parameter estimation**. How to optimize the model parameters $\{\pi, Q\}$ given the data.

The solutions of these problem are well known in literature and they will briefly shown in the final version of this paper.

5 Model selection and assessment

Because of its flexibility, it is possible to fit a large number of different models just increasing the number of states in order to find the model that best fit of the data. As shown before, the EM algorithm estimates the parameters of a full specified model. It means that it is up to the researcher to set or to identify the most relevant number of states comparing a series of model. For instance, if we are completely data driven with a series of 1,000 observations we might compute and compare 1,000 different first-order HMM. And if we consider that we can also include higher order dependence, then, the number of alternative models easily explodes. In the paper Bolano and Berchtold (2013) we have proposed a hierarchical model selection procedure for HMM with continuous variables.

A classical approach in model assessment like the likelihood ratio, it is not applicable. The models are not nested and the distribution of the likelihood ratio is unknown and it is generally not asymptotically distributed as a chi-square McLachlan and Peel (2000). Alternatives for non-nested models includes the Akaike Information Criterion (AIC Akaike, 1974), the Bayesian Information Criterion (BIC Schwarz, 1978) and different variants and corrections of them. For a given model M:

$$\begin{aligned} AIC &= -2logL_M + 2p_M \\ BIC &= -2logL_M + p_M log(N) \end{aligned} \tag{5.1}$$

Where L_M is the likelihood of the model M, p_M the number of freely estimated parameters, N the number of observations used in fitting the model. So the difference between the two criteria lies in the penalized terms (the second term in Equations (5.1)). In literature two variants have been proposed: the adjusted AIC (A-AIC) and the adjusted BIC (A-BIC). The adjustment refers to the number of parameters to include in p_M. The idea is not to count the parameters that have been estimated to be zero due to they not explain any part of the data. Some scholars show that in mixture modeling AIC tends to identify too many states and on the other hand, in same case BIC seems to underestimate the true number of components. Nevertheless, the BIC is still the standard procedure in model comparison in mixture modeling and HMM literature.

However, it is important to notice that in practice the selection of the best model based only on formal criteria like the BIC or likelihood ratio test may lead to model that might be not interesting or complex to use and analyze. Often, increasing the number of states or components may lead to an improve in the goodness of fit of the model bringing to the inclusion of negligible and rare states. Thus, the choice of the best model specification cannot be completely driven by formal mathematical procedure but it has to depend on the specific research question and on the type of data available.

6 Extensions of Hidden Markov Model

In its traditional formulation, Markovian processes have been used to describe univariate time series. So for analyzing a single sequence of a random variable. But, given the flexibility of the HMM, several modification and generalization have been introduced. In the this section we will discuss some interesting extensions for more complex types of observation. As modeling multivariate series, including covariates (these two extensions are particularly relevant in social sciences with the diffusion of panel data), how to consider higher-order Markov chain and a Markovian model-called Double Chain Markov model to include serial dependence directly in the observed sequence.

6.1 Multivariate series

One generalization of HMM is to consider multivariate time series. In many applications, especially in human science, researchers does not face a single series of observation but longitudinal, when multiple individuals are followed overtime and then the data consist in a series of individual sequences.

Considering N time series $\{X_{1,t}, X_{2,t}, \ldots, X_{n,t}\}$, we assume that $X_{i,t}$, $t = 0, \ldots, T$ and $i = 1, \ldots, N$ are mutually independent conditionally on the hidden state of individual i at time t, $S_{i,t}$. But, as pointed out by Zucchini and MacDonald (2009) it does not mean that the individual component

series are serially independent or that the series are mutually independent. Serial dependence and cross-dependence between series are in fact induced by the hidden Markov process.

The conditional independence assumption in case of longitudinal data becomes:

$$\begin{aligned} f(X_{i,t} = x_{i,t} | X_{i,0} = x_{i,0}, \ldots, X_{i,t-1} = x_{i,t-1}, S_{i,0} = s_{i,0}, \ldots, S_{i,t} = s_{i,t}) = \\ = f(X_{i,t} = x_{i,t} | S_{i,t} = s_{i,t}) \end{aligned} \tag{6.1}$$

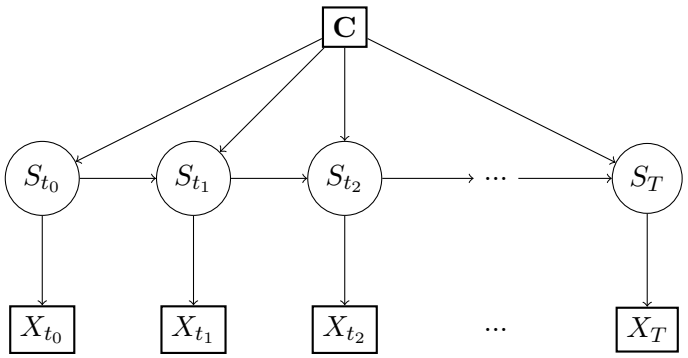
Then, considering the Markov property and the independence assumption (6.1), the likelihood function for longitudinal data can be derived as the product of the likelihood of the N independent sequences. So, the estimation procedures presented are still valid.

6.2 Using covariates

Analyzing longitudinal data, it is natural to study the evolution of the key variable according to the effect of external factors. In HMM individual covariates may be included both in the latent process and in the measurement model. They can be time varying (e.g. age, income, parental status) or fixed in time (e.g gender, date of birth), categorical or continuous.

In Figure (5) we consider a set of l covariates $\mathbf{C} = \{C_1, \ldots, C_l\}$ which may effect the latent process.

Figure 5: Graphical representation of a first order Hidden Markov chain with a set of covariates $\mathbf{C}$ on the hidden level



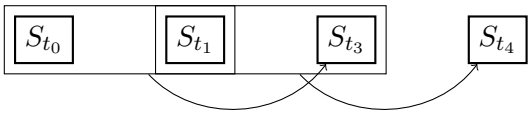
6.3 High Order Homogeneous Markov Chain and the MTD

In the first order (Hidden) Markov chain, the state of a model at time t depends only on the state of the model at time $t-1$. No prior states are

relevant. But in many situations, the present observation depends not only on the first lag but on the last f observations ($f > 1$). So, let S_t a random variable who follow an homogeneous Markov chain of order f we have,

$$\begin{aligned} &P(S_t = s_t | S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, \ldots, S_{t_0} = s_{t_0}) = \\ &P(S_t = s_t | S_{t-1} = s_{t-1}, \ldots, S_{t-f} = s_{t-f}) \end{aligned} \tag{6.2}$$

Figure 6: A second order Markov Chain



Unfortunately, the number of independent parameters increases exponentially with the order f and it becomes complicated to estimated them in an efficient way or even we might have problem of identifiability if the amount of data is small.

Be k the number of values taken by the variable S, the total number of parameters to estimate in a Markov chain of order f is equal to $k^f(k-1)$. For example, if S is a discrete variable with three categories in a first order Markov chain we have 6 independent parameters, for a second order Markov chain we have 18 independent parameters, for $f = 3$, the number of parameters already exposes to 54.

A parsimonious way to approximate high-order Markov chains is the Mixture Transition Distribution Model (MTD). In the MTD, introduced by Raftery (1985), the idea is to consider separately the effect of each lag instead of considering the effect of the f previous states on the current one. The conditional probability becomes

$$P(S_t = s_t | S_{t-1} = s_{t-1}, \ldots, S_{t-f} = s_{t-f}) = \sum_{g=1}^{f} \lambda_g q_{i_g i_0} \tag{6.3}$$

where λ_g is the weight parameter associated to lag g and $q_{i_g i_0}$ are the probabilities in a (k x k) transition matrix Q. Each row of Q represents a probability distribution and therefore sums to one. So, we have only one transition matrix Q with $k(k-1)$ independent parameters and a vector of lag parameters. Therefore, the total number of independent parameters to estimate becomes $k(k-1)+(f-1)$. For instance, increasing the time dependence of one unit, we will have only one additional parameter to estimate.

Using the example mentioned before, in a third order Markov chain with three states, instead of having 54 independent parameter, with MTD will have only 8 parameters to estimate. Table 1 show the number of independent

parameters in case of (hidden) MM and MTD for different dependence orders and values taken by the random variable S.

Table 1: Number of independent parameters to estimate

Number of states, k	Order f	(Hidden) Markov Chain	MTD
3	1	6	6
3	2	18	7
3	3	54	8
3	4	162	9
5	1	20	20
5	2	100	21
5	3	500	22
5	4	2,500	23
10	1	90	90
10	2	900	91
10	3	9,000	92
10	4	90,000	93

For a complete review of MTD model with several comparison and extensions and applications see Berchtold and Raftery (2002)

6.4 Using HMM for probabilistic clustering

Another interesting feature of the HMM is to use the transition probabilities to provide probabilistic clusters. Even if there are equivalent to the Hidden Markov Model, in literature these models are known as Latent Markov Models and there are considered as an extension of latent class models but with repeated measures. The goal of these models is to classify individuals into a finite number of homogeneous distinct groups.

A well known way to proceed is to consider a finite mixture models and to define the transition matrix Q as a (block) diagonal matrix. So, each state is full absorbing defining the membership to a single group. The idea is to consider the sample as drawn by an heterogeneous population where each sub-population is described by a component of the mixture. In other words, each group is represented by a single state of the transition matrix and once a person enter in a particular state ("become member of a group"), he cannot leave that group.

For a three states HMM, the transition matrix is then written as follows

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Using such constraint in a Double Chain Markov Model (Section6.5) allows to account for time dependence on the observed variable and, at same, to determine a probabilistic clustering using the hidden states.

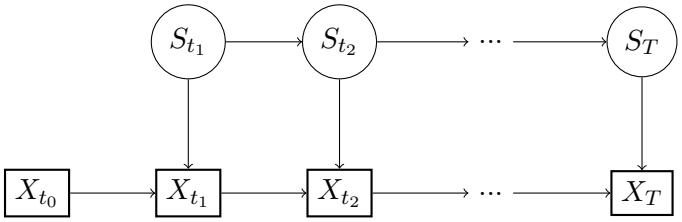
Another easy way, it to identify the latent states as different subpopulations and without including any constraints on the transition matrix, allow the individuals to move between latent classes at each time point. As in general framework for the mixture model, a critical issue is to find the correct number of classes. Procedure based on BIC or other criteria are widely discussed in literature.

See Bicego et al. (2003) for more sophisticated sequential data clustering methods using similarity based approaches.

6.5 Other extensions

A way to combine an HMM and a visible Markov chain governing the relations between observations of the a key variable is the Double Chain Markov Model (DCMM, Berchtold, 2002). It has been designed for the modeling of *non-homogeneous* time-series and the main idea is to decompose the time-series into a set of transition matrices as many as the number of the hidden states. At each time point, a specific transition matrix is selected according to a full homogeneous Markov chain (Figure 7). Then, different transition matrices can be used to model different portions of the observed data.

Figure 7: Path diagram of a Double Chain Markov Model of order one



As in other path models like the SEM, also for Hidden Markov Model is quite simple to extend and "adjust" the model according to the specific research question. It is possible to introduce cross lagged dependence (Figure 8) or autoregressive component of oder two (Figure 9) and so on. But the way to fit the model and interpret the result presented in the previous sections remain the same.

Figure 8: Cross lagged HMM

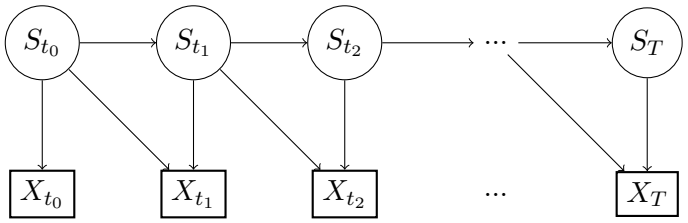
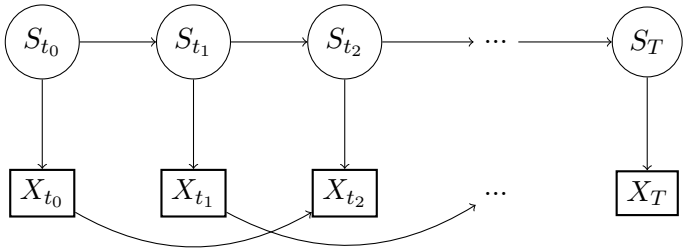


Figure 9: Double Chain MM with AR(2)



Another case in which the HMMs are used is to account for measurement errors. If our variable of interest derives from different factors, for instance the vulnerability of a certain group measured by a series of different indicators, the dependence between these factors can be simplify introduction an unobserved variable. In such way each state will represent the different modalities of our construct of interest.

6.6 An Hidden Markov Model for count data

For convenience in the notation, in the paper we often refer to the observed variable as a categorical one. But HMM can be applied to count data too. In particular Hidden Markov models are particularly well suited for the analysis of data switching between several regimes. Such models can be used to represent and analyze complex time-series in presence of overdispersion like for psychometric or biometrical studies, financial series.

Another very popular and efficient approach is the use of mixture models that are able to describe very complex distributions which do not correspond to any specific statistical family. The general principle of all MTD-like models is to combine different Gaussian distributions (called *components*) through a mixture model where the mean of each distribution is a function of the past observed process.

Both the mixture transition distribution models and the hidden Markov

models include a latent and a visible part and they can include covariates and high-order time dependencies at the visible and hidden level. So we may consider the two approaches as one unique method.

In Bolano and Berchtold (2013) we have introduced a generalization of an HMM. This model, called Hidden Mixture Transition Distribution, is a flexible time dependence mixture mode to represent continuous variables accounting for the observed heterogeneity.

The model can be estimated using simultaneously as many independent sequences of observations as wanted. Each sequence will typically correspond to the observation of a separate subject. On the other hand, each hidden state and its associated visible component can be interpreted as a behavior of the subjects under investigation. Multiplying the number of components, we allow the subjects to follow many different behaviors in order to capture both the complexity of the population behavior and the evolution over time of each individual.

The HMTD is a two level model: a visible and observed level are considered and it is useful to model longitudinal data switching between alternative typologies/regimes. The observed heterogeneity is assumed to be induced by one or several latent factors and each level of these factors is related to a different component of the observed process. Individual trajectories are seen as a (weighted) mixture of different patters and the relation between successive components (that represent different states of the hidden variable) is governed by a Markovian latent transition process.

7 Discussion

In this paper we have provided an overview of the Hidden Markov Model. It is a stochastic model used to describe the evolution of observable events that depends on internal factors which are not directly observable.

Modeling observations in these two levels, one visible and a latent one, can be very useful in life course studies, since many aspects of a life trajectory can be represented by this structure. For instance, vulnerability can be considered as a latent variable revealed through observable atypical behaviors, aspects and characteristics. Using HMM we can explore the changes in unobserved and difficult-to-measure aspects and identify the probability of moving from one latent state to another analyzing the vulnerability dynamics of a certain population. Using a latent class approach, we also can recognize similar patterns, directly observable or latent, among the initial population in order to identifying which individual characteristic influence the membership of a specific subgroup.

Hidden Markov Models can be applied to a wide range of longitudinal data, from univariate time series to multivariate and panel data. It can be used to describe the evolution of a process in continuous time or with

discrete-valued time series. It can be applied to categorical or continuous observations, bounded or unbounded counts and so on. It is an extremely flexible model that can be adapted for a wide range of applications. HM models have been extensively applied in the last decade in many areas like speech recognition, behavior analysis, climatology, finance. In sociology and in life course studies, this approach is still sparsely used but we expect that the importance of HMM in social sciences as well as the rage of application will grow further.

This article is intended to illustrate the basic aspects and the powerful flexibility of hidden markov models both formally and proving empirical applications in life course studies. Staring from its traditional formulation, we present several extensions in longitudinal settings as Double Chain Markov model to relax the homogeneity assumption, the introduction of covariates in order to capture the effect of external factors on the transition probabilities, high-order Markov chain to include in the model higher time dependence.

References

Abbott, A., 1995. Sequence Analysis: New Methods for Old Ideas. Annual Review of Sociology 21, 93–113.

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control AC-19 (6).

Baum, L. E., 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3, 1–8.

Baum, L. E., Petrie, 1966. Statistical inference for probabilistic functions of finite state markov chains. The Annals of Mathematical Statistics 37 (6), 1554–1563.

Baum, L. E., Petrie, T., Soules, G., Weiss, N., 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics 41 (1), 164–171.

Berchtold, A., 2002. High-Order Extensions of the Double Chain Markov Model. Stochastic Models 18 (2), 193–227.

Berchtold, A., 2003. Mixture transition distribution (MTD) modeling of heteroscedastic time series. Computational Statistics & Data Analysis 41, 399–411.

Berchtold, A., Raftery, A., 2002. The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. Statistical Science 17 (3), 328–359.

Bicego, M., Murino, V., Figueiredo, A. T., 2003. Similarity-Based Clustering of Sequences using Hidden Markov Models. Machine Learning and Data Mining in Pattern Recognition Lecture No.

Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models. Computational Statistics & Data Analysis 41, 561–575.

Billari, F. C., 2001. Sequence Analysis in Demographic Research. Special Issue on Longitudinal Methodology, Canadian Studies in Population 28 (2), 439–458.

Bolano, D., Berchtold, A., Apr. 2013. Hidden Mixture Transition Distribution (MTD) model. General framework and model selection criteria. In: International Conference of the ERCIM WG on Computational and Methodological Statistics.

Bulla, J., Berzel, A., Jul. 2007. Computational issues in parameter estimation for stationary hidden Markov models. Computational Statistics 23 (1), 1–18.

Cappé, O., Moulines, E., Ryd, T., 2005. Inference in Hidden Markov Models. Springer.

Cox, D. R., 1981. Statistical Analysis of Time Series : Some Recent Developments. Scandinavian Journal of Statistics 8 (2), 93–115.

Dempster, A. P., Lard, N. M., Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B 39 (1), 1–38.

Holland, J., 1975. Adaptation in Natural and Articial Systems. University of Michigan Press, Ann Arbor, MI.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley Series in Probability and Statistics.

Rabiner, L., 1989. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. 77, 257-286.

Raftery, A. E., 1985. A Model for High-order Markov Chains. Journal of the Royal Statistical Society. Series B 47 (3), 528–539.

Schwarz, G. E., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.

Sutton, J. R., 2006. Things you can say when you're tenured: reflections on the relationship of sociology to criminology, and vice-versa. In: American Society of Criminology, Los Angeles, California.

Viterbi, A. J., 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory 16 (2), 260–269.

Welch, L. R., 2003. Baum-Welch Algorithm. IEEE Information Theory Society Newsletter 53 (4), 1–24.

Zucchini, W., Guttorp, P., 1991. A Hidden Markov Model for Space-Time Precipitation. Water Resources Research 27 (8), 1914–1923.

Zucchini, W., MacDonald, I. L., 2009. Hidden Markov models for time series. An introduction Using R. CRC Press.