LIVES
Swiss National Centre of Competence in Research

FNSNF
Swiss National Science Foundation

Han, Y., A. C. Liefbroer & C. H. Elzinga (2016)

Understanding social-class differences in the transition to adulthood using Markov chain models

in G. Ritschard & M. Studer (eds), Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016, pp 155-177.

UNIVERSITÉ DE GENÈVE

FONDATION POUR L'UNIVERSITÉ DE LAUSANNE

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

UNIL | Université de Lausanne
Institut des sciences sociales

Mechanisms of the transition to adulthood: an application of Hidden Markov Models

Y. (Sapphire) Han, A. C. (Aart) Liefbroer and C. H. (Cees) Elzinga

Abstract An increasing number of studies focuses on understanding the processes underlying the transition to adulthood. However, the transition to adulthood is a complex process of a series of events that are often interlinked. Even though life courses are greatly varying sequences of roughly the same life course events, the complexity is caused by the fact that these sequences consist of correlated events and spells and these correlations depend on gender, social class, cohort and cohort-related macro events. Our previous work demonstrated that the application of stochastic models like the Latent-Class model helps to describe the variation in life courses and its correlation with gender and social class. But the Latent-Class model cannot account for correlated events within life courses nor can it account for switches between latent types during the life course. We argue that (Hidden) Markov models, as a simple generalization of the Latent-Class model, has the ability to account for correlations between events and spells and also allows for switches between latent types or model life courses. Therefore, this study will use (Hidden) Markov models to produce a typology of trajectories of the transition to adulthood. Furthermore, we will test hypotheses on social class- and gender differences in observed life courses and latent types or model-life courses, using data from the Gender and Generation Programme (GGP), which provides full monthly life course sequence data between age 15 to 40.

Y. (Sapphire) Han
Netherlands Interdisciplinary Demographic Institute, The Hague, University of Groningen, e-mail: han@nidi.nl

A. C. (Aart) Liefbroer
Netherlands Interdisciplinary Demographic Institute, The Hague, University Medical Centre Groningen, Vrije Universiteit Amsterdam, the Netherlands, e-mail: liefbroer@nidi.nl

C.H. (Cees) Elzinga
Netherlands Interdisciplinary Demographic Institute, The Hague, Vrije Universiteit Amsterdam, the Netherlands e-mail: c.h.elzinga@vu.nl

1 Introduction

The ultimate goal of life course research is to understand how life courses come about and what variables affect their shape. Essentially, this is a holistic question: to answer it requires the postulation of a mechanism of the generation of the complete life course. Holistic life course models must satisfy a few simple properties. First, these models must have a memory, a sense of the past, as it is generally assumed that events in the early stages of the life course may affect stages or outcomes later on [Mayer, 2009]. A second requirement is that the process that generates the life course, is affected by variables that that are supposed to influence the life course: gender, religion, parental education, etc. Finally, the model should be formulated in terms of a process that is not directly observable: since the life course is generated through mental, not directly observable, processes that are conscious or unconscious, and decisions that are voluntary or involuntary. Of course, such models should be testable and amenable to causal analysis.

Over the past decades, life course research has been dominated by two different paradigms: Event History analysis (EH) [Blossfeld et al., 2007] and Sequence Analysis (SA) [Cornwell, 2015]. EH-models are not holistic: they try to explain the waiting times for certain life course events to occur and the SA-approach leads to finding most frequent patterns in the wide variety of observed life courses, but it does not account for this variety. Recently, we have seen other methods and models being applied as well, for example Latent Class analysis [Barban and Billari, 2012] and Structural Equation Models [Pakpahan et al., 2015] but neither of these methods satisfies all of the three requirements as formulated above.

However, there is a broad class of models that does satisfy the above requirements: the class of so called Hidden Markov Models. These models have a memory in the sense as intended, they allow for time-constant and time-varying covariates and are formulated on the basis of a latent, hidden, random process over a finite set of states, a Markov chain. The models are testable in the sense that their parameters can be estimated [Bartolucci et al., 2012, Rabiner, 1989] and easily allow for causal analysis once formulated as a log-linear regression model [Paas et al., 2007]. Hidden Markov Models belong to a larger family of latent structure models that has been amply described by [Langeheine and Van de Pol, 1990, Vermunt, 1997].

This paper aims to model the life course, confined to the relatively turbulent transition to adulthood, through using Hidden Markov Models. The transition to adulthood is usually described by a collection of events [Elder Jr, 1985] from which two correlated processes can be distilled [Buchmann and Kriesi, 2011]: the school-to-work transition and the process of family formation. Here, we provide an example that solely focusses on the transition into adulthood in the family domain, as there is a large body of literature on the processes involved. More specifically, we example the family-life trajectories of French men and women born between 1956 and 1965, using data from the French Generations and Gender Survey (GGS).

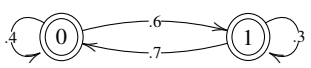
The paper is structured as follows: in this lengthy introduction we discuss the main concepts of Hidden Markov Models and make some general remarks on their application to life course research, Section 2 discusses our data and methods used,

Section 3 discusses our results and Section 4 summarizes, concludes and suggests further research.

1.1 Hidden Markov Models

Hidden Markov Models generalize the much simpler idea of a Markov chain. A Markov-model or Markov-chain is a random process over a set of states such that the probability of being in a particular state at the next observation only depends on the state-history of the process. If the relevant state history just consists of the present state, such a chain is called "first-order". Figure 1 shows a graphical representation of a first-order 2-state Markov-chain and its matrix of transition probabilities.

Fig. 1 A graph showing a first-order, 2-state Markov chain and its transition probability matrix **A**. The states are labeled as "0" and "1" and the arrows represent the transition probabilities.



$$\mathbf{A} = \{a_{ij}\} = \begin{pmatrix} .4 & .6 \\ .7 & .3 \end{pmatrix}$$

Let us denote the k distinct states of a Markov chain as $Q = \{q_1, \ldots, q_k\}$ and let S_t denote the state that the system is in at time t, i.e. S_t could have any of the "values" or labels from the set Q. Then we say that a random process over Q is a first-order Markov-chain, precisely when

$$Prob(S_t = q_j | S_0 \ldots S_{t-1}) = Prob(S_t = q_j | S_{t-1} = q_i) = a_{ij}, \tag{1}$$

and we denote this probability by a_{ij}. If we now define the initial state-probabilities as $Prob(S_0 = q_i) = \pi_i$, the Markov-chain λ is fully defined by the k-vector $\pi = (\pi_1, \ldots, \pi_k)$ of initial state probabilities and the $k \times k$-matrix of transition probabilities $\mathbf{A} = \{a_{ij}\}$: $\lambda = (\pi, \mathbf{A})$.

In an ordinary Markov chain, transition probabilities depend on the present state only (see Equation (1)). However, it is easy to extend this model to account for the effect of covariates. Let $\mathbf{v}$ and $\mathbf{w}_t$ denote vectors of time-constant and time-varying covariates. Then a direct extension of the Markov-chain model is formulated as

$$Prob(S_0 = q_i | \mathbf{v}, \mathbf{w}_t) = \pi_i(\mathbf{v}, \mathbf{w}_t), \tag{2}$$

$$Prob(S_t = q_j | S_0 \ldots S_{t-1}, \mathbf{v}, \mathbf{w}_t) = Prob(S_t = q_j | S_{t-1}, \mathbf{v}, \mathbf{w}_t) = q_i) = a_{ij|\mathbf{v},\mathbf{w}_t}. \tag{3}$$

Clearly, this formulation implies a separate Markov-chain for each point in the $(\mathbf{v}, \mathbf{w}_t)$-space.

A Markov chain could be used to model a set of observed life course sequences by simply identifying each of the observed states as a model state and estimating the transition probabilities from the relative transition frequencies of the observed sequences. The result of that would be a more or less accurate summary of the

observed transition frequencies. However, it would not lead to a credible model for the way these sequences were generated. Therefore we now turn our attention to an extension of the Markov chain: the Hidden Markov Model.

In a Hidden Markov Model (HMM), the Markov chain is defined over a set of latent, unobservable states. So, the stochastic process as such is not observable. Furthermore, it is supposed that, at each state, the process 'emits' an observable (an observable can be univariate or multivariate) according to a state-specific probability distribution over the full set of observables, in the present context the observable states of a life course. Thus, in a k-state HMM with a set of observables $Y = \{y_1, \ldots, y_n\}$, there must be a set B of k state-specific probability distributions $\mathbf{b}_j = (b_{j1}, \ldots, b_{jn})$, each satisfying $\sum_i b_{ji} = 1$:

$$b_{ji} = Prob(o_t = y_i | s_t = q_j). \tag{4}$$

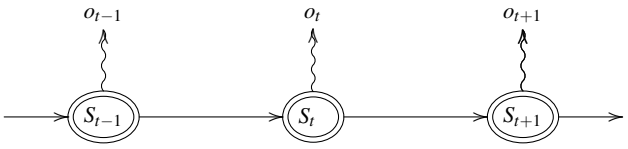
This allows us to represent the set B as a $(k \times n)$-matrix

$$\mathbf{B} = \begin{pmatrix} b_{11} & \ldots & b_{n1} \\ \vdots & \ddots & \vdots \\ b_{1k} & \ldots & b_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_k \end{pmatrix} \tag{5}$$

whereof each row is a distinct probability distribution over the observables and the complete HMM $\lambda = (\pi, \mathbf{A}, \mathbf{B})$ is specified by the initial state distribution π, the $(k \times k)$-matrix $\mathbf{A}$ of transition probabilities and the $(n \times k)$-matrix $\mathbf{B}$ of emission probabilities.

In Fig. 2, we show a graph of the HMM-generated events in a time-window $(t-1, t+1)$: at $t-1$, the system arrives in state S_{t-1} and emits observable o_{t-1} (governed by $\mathbf{B}$) and then switches to state S_t (governed by $\mathbf{A}$) and again emits an observable, etc.. The reader should be aware that the system may, depending on the

Fig. 2 A graph showing the time-window $(t-1, t+1)$ of a Hidden Markov process. At each time t, the system is at some latent state S_t and emits an observable. Note that the hidden state S_t is not necessarily different from S_{t+1}. The observable is a random sample from the set of observables, according to a probability distribution that is specific for each state q_i, $i = 1, \ldots, k$.



probability a_{jj}, actually stay in the same state j for quite a while and during that time emit various different observables. Similarly, the observables may remain the same for quite a while, at the same time but "below the surface", the system actually switches state several times. In practice, if we observe that people stay in the same

observable state for many years, it is to be expected that the diagonal elements of **A** are relatively big, i.e. close to 1.

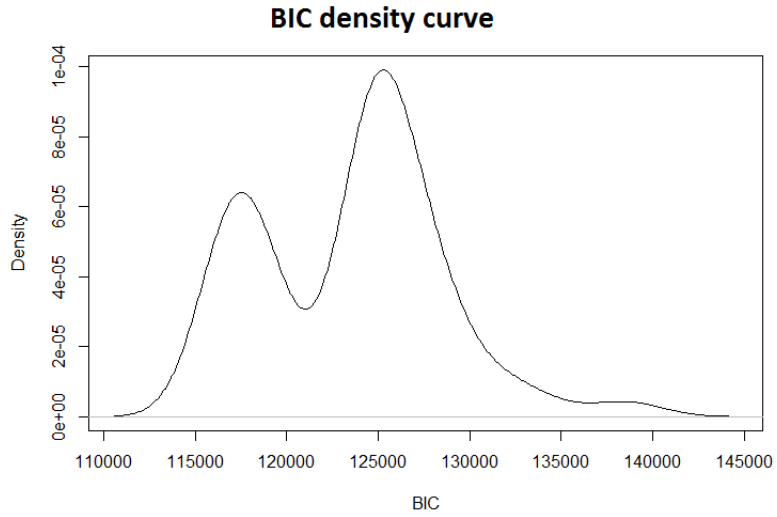
1.2 Modelling with HMM's: Some practical considerations

Let $O_i = o_{i1} \dots o_{iT}$ denote an observed sequence from a set $O = \{O_1, \dots, O_N\}$ of such sequences and let $Prob(O_i|\lambda)$ denote the likelihood of that sequence, given the model. Furthermore, let $Q_i = q_{i1} \dots q_{iT}$ denote the path along the latent states that maximizes $Prob(Q_i|O_i, \lambda)$, i.e. the latent sequence that "best accounts" for the observations, given the model.
Being able to calculate the likelihood of the observations given the model is a precondition for EM-estimation of the parameters of the model and calculating Q_i, the most probable latent sequence, is a precondition for a substantive interpretation of the model. Both problems, evaluating $Prob(O|\lambda)$ [Baum et al., 1970], and calculating Q_i [Viterbi, 1967] were already solved in the sixties of the previous century and have been amply described in many sources [Rabiner, 1989, Zucchini and MacDonald, 2009, Bartolucci et al., 2012]. Here, we will not deal with the intricacies of these methods. Instead, we will discuss some practical issues that are related to these methods and their output.

First, one should be aware that evaluating a HMM involves the estimation of quite some parameters: with k postulated latent states, we have to estimate $k-1$ parameters $\hat{\pi}_i$; $k-1$ since we must have that $\sum_i^k \hat{\pi}_i = 1$. Likewise, we have to estimate $k(k-1)$ parameters to obtain $\hat{\mathbf{A}}$ and $k(n-1)$ parameters to get $\hat{\mathbf{B}}$. So, the surface of the likelihood function $Prob(O|\lambda)$ is quite irregular and therefore, attempts to find its maximum, be it through EM [Dempster et al., 1977] or through any other method like simulated annealing [Andrieu and Doucet, 2000], will most often converge to a local instead of the global maximum. Extending the HMM to incorporate covariates will only aggravate this problem. Therefore, the estimation of a HMM should be repeated quite some times to find a configuration $(\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ that (probably) comes close to the maximum sought for. For example, peeking around the corner of our modelling life courses with HMM's, we display the density of the BIC-values as obtained over 1000 repetitions of estimating a 4-state model. Clearly, these BIC-values are quite different, as are the underlying configurations $(\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$. Obtaining this curve took almost three hours of computation time and quite some memory. Increasing the number of states and the number of trials soon requires unfeasible computation times and memory for this exercise.

However, we do not consider this to be a serious problem for applying HMM's to model life courses. Normally, the size of the observation alphabet will be small and the number of postulated states k will be rather small too. The latter number should reflect the number of stages or states in which the subjects will take demographic decisions and these decisions are small in number; they pertain to leaving the parental home, partnering, reproducing and, eventually, returning to the parental home or breaking up a partnership. Therefore, in practice, k should be small. Con-

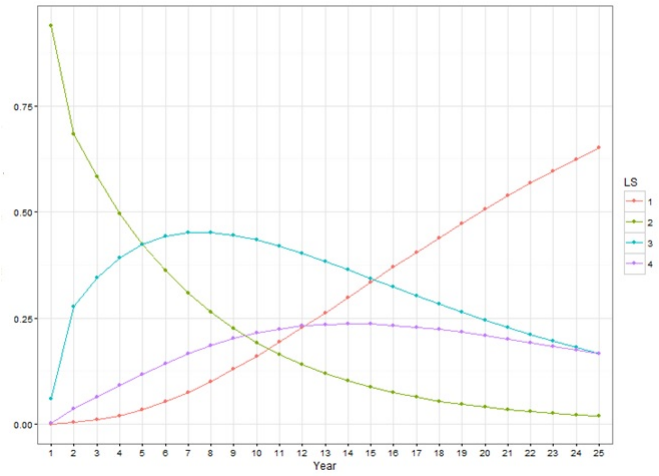
Fig. 3 BIC-density plot as obtained from repeating the estimation of a 4-state HMM 1000 times with random initial values.



sidering the big data sets that social demographers use today, it is to be expected that the optimal value of the information criteria that we use to evaluate the fit of a HMM - variants of minimum-χ^2, AIC or BIC - cannot be good indicators of the optimal number of latent states as these indicators will drive us to accept large numbers of latent states while substantive interpretation is problematic. Rather, the size of k is to be fixed on a theoretical basis: can we assign a credible interpretation to these states in view of the latent trajectories and the way these trajectories and the emission distributions, i.e. the probability of picking particular behavioral, demographic alternatives, change as a result of covariates that pertain to societal pressure on rather than social capital of the subjects studied.

Then how do we start interpreting the latent states? A first clue to this interpretation is provided by studying the latent state sequences over time and evaluating the probabilities $Prob(S_t = q_j|t,\lambda),\ t = 1,\ldots,T,\ j = 1,\ldots,k$, i.e. the relative frequencies of state occupancy, aggregated over the sample studied. Again peeking around the corner of our analysis yet to be presented, we show a plot of these relative frequencies for a 4-state HMM in Figure 4: This plot shows something that is not evident from the estimated transition probabilities: most subjects start in the latent state labeled as LS1 so it should be associated with a decision about leaving the parental home and this should be reflected in the emission probability distribution over the observables: it should be characterised by a relatively high probability of emitting the event “leaving the parental home”. So, the marginal state occupancies

Fig. 4 Plot of the marginal probabilities over time of latent state-occupancy in a 4-state HMM.



over time and the emission distributions will help us to interpret the latent states. However, these considerations do not suffice for a credible interpretation.

A credible interpretation can only arise in the light of the way covariates affect the parameters of the model: do the estimated effects of covariates corroborate, or at least are not at variance with, the knowledge that we already have about the effects of these covariates on the occurrence and timing of life course events. For example, we may expect that low-educated will enter parenthood earlier than high-educated and this should be reflected in the differences between transition probabilities and/or the emission probabilities of lower- and higher-educated. Therefore, it is not enough to only evaluate a HMM as such: we need to enrich the model with relevant covariates in order to decide on the credibility of the interpreted model.

How to incorporate covariates into a HMM? Thereto, we consider the likelihood $Prob(O_i|\lambda,\mathbf{v})$ of a particular sequence O_i from a set $O = \{O_1,\ldots,O_N\}$ of such sequences, wherein $\mathbf{v}$ denotes a vector of covariates. Since, according to the model, the observed sequences result from the latent sequences Q_i, we can decompose this likelihood as follows:

$$Prob(O_i|\lambda,\mathbf{v}) = Prob(O_i,Q_i|\lambda,\mathbf{v}) \tag{6}$$

$$= Prob(O_i|Q_i,\lambda,\mathbf{v})Prob(Q_i|\lambda,\mathbf{v}), \tag{7}$$

and thus, the likelihood of our data given the model can be decomposed as

$$Prob(O|\lambda,\mathbf{v}) = \underbrace{\prod_i Prob(O_i|Q_i,\lambda,\mathbf{v})}_{\text{“operational”}}\underbrace{\prod_i Prob(Q_i|\lambda,\mathbf{v})}_{\text{“structural”}}, \tag{8}$$

the multiplications arising from the assumption that the sequences have arisen independently from each other. From the latter Equation 8, we observe that the total likelihood of our data given the model can be decomposed into two separate parts. The second part is called the "structural" part of the model since it pertains to the unobservable structure (the transition probability matrix) of the stochastic process over the latent states, the components of the postulated model. The first part of the multiplicative structure (additive if we consider log-likelihood) we called "operational" for two reasons. The first reason is that we may consider the emission probabilities as choice-options given the latent state the person is in. For example, if the subject is in a state where partnering is the key issue, it may chose between different options to operationalize the positive choice for partnering: marriage, cohabitation or a relational agreement without living together. The second reason for calling this part "operational" is the fact that, given our belief in the validity of the number and structure of the latent states, this part of the model is affected by our way of operationalizing the life course in observational labels: choosing a different alphabet will affect the model fit without altering the structural validity[1]. Unfortunately, we cannot separate these parts of the model when assessing the model's fit as the observations are our only entry to the latent, structural part of the model. Therefore, it is not wise to have the same covariates play a role both in the structural and in the operational part of the model as it would severely hinder the substantive interpretation of the model. Here, we prefer to assume that covariates do not have a role in the operational part of the model, i.e. we assume that

$$Prob(O_i|\lambda,\mathbf{v}) = Prob(O_i|Q_i,\lambda)Prob(Q_i|\lambda,\mathbf{v}), \tag{9}$$

or, equivalently, that only the initial and state transition probabilities are affected by covariates. The reason for this preference is that we know that most life courses in developed countries only differ in the timing and duration of the various stages on the rout to adulthood. This implies that, in most countries, the behavioral alternatives and the order in which they are expressed are roughly the same for most people.

1.3 Applications of HMM

HMM's have been successfully applied in several fields [Bartolucci et al., 2012], including but not limited to psychological and educational measurement [Vermunt et al., 1999], medicine and health [Cook et al., 2000], criminology [Bijleveld and Mooijaart, 2003], marketing and related fields [Paas et al., 2007], interactions during survey-interviews [Elzinga et al., 2007] and labor market research [Richardson et al., 2011]. However, we have not yet seen the application of HMM's to life course research. One reason

[1] [Bartolucci et al., 2012] call this second part of the model the 'measurement model'. This is adequate when the observables contain a measurement error as is often the case in responses to psychological test or survey items; in life course encoding such errors are rare (but see [Manzoni et al., 2010]).

for that could be that only recently, big demographical data sets, cheap software and fast PC's have become widely available.

2 Data and Method

2.1 Data

The Gender and Generation Programme (GGP) is a Longitudinal Survey of 18-79 year olds in 19 countries that that examines the relationships between generations and genders, by collecting nationally representative data in all participating countries. [Fokkema et al., 2016] provide extensive information on design and representativeness of the GGS. In this study, we select respondents (males and females, in total 1900) in the France GGP data of a cohort between birth year 1956 and 1965. In the selected dataset, full annual fertility, partnership and leaving parental home information between age 15 and 40 are available and background information such as gender, education level, parental education, parental divorce.

To demonstrate the application of HMM, we construct a multi-channel sequence dataset of respondents' fertility history (4 categories: no child, 1 child, 2 children, 3 and more children), partnership history (3 categories: single, cohabitation, marriage), and leaving parental home (2 categories: yes or no). To investigate the link between background variables, we include gender (2 categories: female and male) and education level (2 categories: high and low). To visualize the multi-channel sequence dataset, four sequence index plots separated by gender and education level are shown in Figure 5 (a) and (b). Take Figure 5 (a) left panel for example, it contains the fertility (4 categories), partnership (3 categories) and leaving home annual information throughout young adulthood (age 15 - 40) of high educated male respondents. The complexity of the dataset is obvious: it contains 24 categories ($4 \times 3 \times 2$) of 25 repeated annual measures.

Insert Figure 5

2.2 Method

During the analysis, the selected multi-channel sequence dataset was fitted to hidden Markov model (HMM) with hidden states equals 3-6, each of a time-homogeneous model with 1000 random starting values. This paper selectively shows the result of HMM 4 state solution as proof of concept of the application of HMM in the transition to adulthood research. Two types of covariates, i.e., gender (female vs. male) and education (low vs. high) were introduced in the latent model of HMM 4 state solution (also a time-homogeneous model with 1000 random starting values). The

reason of using 1000 random starting values for HMMs is to reduce the influence of local maximum. The estimation procedure of HMM relies on Baum-Welch or EM (Expectation-modification) algorithm [Rabiner, 1989]. With the increment of hidden states, the number of parameters to be fitted also increase drastically. During the analysis, it is found out that for the given dataset, HMM 6 state solution is unstable. It took more than 1 GB in the RAM, and 1000 random starting value repentances were not enough to generate stable solution. It might be possible to perform HMM with high number of hidden states on HPC (high performance computing) environment. HMMs with hidden states from 3 to 5 generate stable solution. Choosing the HMM with 4 hidden states is due to the fact that it largely reduce the data complexity at the same time providing substantively interesting interpretation.

All analyses were performed in R environment for statistical computing and graphics in a 64 bit PC with 32 GB RAM. R packages LMest [Bartolucci et al., 2015] and markovchain [Spedicato et al.,] were utilized for Hidden Markov models. Sequence visualization and related techniques were performed by R package TraMineR [Gabadinho et al., 2011].

3 Result

In this section, results of fitting HMM 4 state solution (without and with covariates in the latent model) to the multi-channel France GGP sequence data are presented. In each model, the time cost, the model fit parameter (BIC), the output parameters (initial probability distribution π, transition probability distribution a_{ij} and emission probability distribution b_j), the visualization and interpretation of these output parameters and the mechanisms revealed by HMM are presented.

3.1 Hidden Markov model 4 hidden state solution

The HMM with 4 hidden states were performed with 1000 random stating values to reduce the influence of local maximum. It took 2.6 hours and achieved a minimum BIC of 115961. The fitted HMM with the lowest BIC was chosen as the HMM 4 solution. The interpretation of the HMM is based on its output parameters, i.e., initial probability distribution, transition probability distribution and emission probability distribution. As described in Introduction Section, initial probability distribution reveals the proportion of hidden states that respondents occupy in the beginning of their life courses; transition probability distribution reveals the transition rate (per year in this study) to other hidden states once respondents arrive at a certain hidden state; emission probability distribution links the hidden states to the observed life course.

The output parameters are shown in Table 1 (the hidden states are ordered as 'A', 'B', 'C' and 'D'). It is difficult to interpret these number without graphic illustra-

tion. Latent probability distribution graph (shown in Figure 6) is useful as a first step to interpret Table 1. This graph is based on the initial probability distribution and transition probability distribution shown in Table 1. There are four curves representing the dynamics of the four hidden states during respondents' 25 year young adult life course. Curve A is the state where almost every respondent begins with, and the proportion of respondents in this state has been dropping ever since. Curve B shows that, between age 20 and 30, overall majority of respondents take this hidden state. Proportion of respondents in state B starts increasing since the age 15 until the age 22. Curve C also shows a 'first increase then decrease' pattern as state B, however, proportion of respondents in this state is always lower than state B and the timing of decreasing (age 30) is later than that of state B. Curve D indicates that the proportion of respondents in state D keeps increasing throughout the whole young adulthood and becomes the majority after age 30.

Insert Table 1
Insert Figure 6

To understand the mechanisms behind the dynamic transition between these latent states, one can plot the state transition graph. As shown in Figure 6, from starting state A, one is 14 times more probable to transit to state B than to state C, given the transition probabilities of 0.14 (A to B) and 0.01 (A to C) in Table 1. Combined with emission probability distribution, state A is featured as being single (probability = 0.97), no child (probability = 0.99) and living with parents. State B is featured as being single (probability = 0.53), cohabiting (probability = 0.27) or married (probability = 0.20), no child and left parental home, whereas state C is featured as being single (probability = 0.17), cohabiting (probability = 0.25) or married (probability = 0.59), 1 child and left parental home. State C can be reached also from state B, which is 9 times more probable from State A. From state C, one can transit to state D. State D is an absorbing state, which means once one arrives at this state, transition to other states is not possible any more. State D is featured as low probability of being married (probability = 0.76), having 2 (probability = 0.68) or more (probability = 0.33) children. Summarizing the information given by Table 1, Figure 6 and Figure 7, one can interpret these four hidden states as inclinations of transition, reflecting the respondents' tendency to act during the stay of a certain state. Respondents in state A as the beginning of young adulthood: they are single, living with their parents, and having no child. They are also probably in school or training for future employment, which are not observable in the current dataset. Their life course activities are preparing them to leave parental home and start an independent life. Therefore, state A can be interpreted as inclining 'Leaving home'. Respondents are mainly in state B between age 20 and 30: they have different partnership status (mainly being single probability = 0.53), left parental home, and having no child. In this state, respondents' behaviors in are preparing themselves in 'Family formation'. In state C, the probability of partnership status shows high proportion of being married (0.59) and respondents already have one child. This state can be interpreted as 'Family extension'. State D, the absorbing state, is the 'Family completion' state. The young adulthood life course end at age 40, where the observed life course stops

in this study. Note that, the above-mentioned transition pattern applies to the whole sample, but respondents from different background (gender or education) may have different transition rate between states. The hidden Markov model with covariates are useful in studying the differences between social classes.

Insert Figure 7

After interpreting the hidden states as inclinations leading to young adult demographic transitions, it is necessary to visualize the hidden states paths throughout the whole young adulthood of respondents. The necessity comes on the one hand from need to check whether the hidden states paths fit substantive expectation and on the other hand from the three basic problems in any HMM application (describe in Introduction Section). The hidden states paths throughout the whole young adulthood of respondents (sort from end) are shown as sequence index plot in Figure 8 (a). Sequence index plots of longitudinal data use stacked bars or line segments to show how individuals move between a set of conditions or states over time [**?**]. Compared with the multi-channel sequence life course shown in Figure 5, the complicated partnership, fertility and leaving home trajectories are reduced into four category of inclinations. To better understand the heterogeneity in the transition into adulthood among these inclinations, sequence analysis with OM [**?**] was performed on the hidden state paths. Four typologies of the hidden states (chosen by cluster quality statistics [**?**]) were presented in Figure 8 (b) and Figure 8 (c). Figure 8 (b) are the sequence index plots of each typology and Figure 8 (c) is the sequence medoid plot [**?**], which is the most representing existing sequence in each typology. With the help of Figure 8 (b) and (c), four types of transition into adulthood have been identified, namely, 1: Late fertility or no fertility, 2: Traditional pathway (Leaving home at age 21, Family formation at age 26, Family extension at age 29, and followed by Family completion), 3: Small family (remaining in Family extension state until end of observation age 40), 4: Early transition (Leaving home at age 18, Family formation at age 21, Family extension at age 24, and followed by Family completion).

Insert Figure 8

3.2 HMM 4 with covariates in its latent model

As discussed in Introduction Section, HMM can allow covariates in its latent model to explain the heterogeneity in the population. One of the most intuitive way to include covariates in latent model is to allow for different transition probability distribution for different groups of respondents. It take 7.2 hours to perform HMM with 4 hidden states, 1000 random starting values, and including 2 variables, namely, education (high vs. low) and gender (female vs. males). The lowest BIC among these 1000 repentances is 113005, which is lower than the BIC of HMM 4 without covariates. The output model parameters are initial probability distribution, emission probability distribution (shown in Table 2) and 4 different transition probability

(shown in Table 3) distributions. Compared with the initial probability distribution, transition probability distribution and emission probability distribution of HMM 4 without covariate (shown in Table 1), there are some unnoticeable change in some probabilities, and the interpretation of these 4 hidden states remains the same.

Insert Table 2
Insert Table 3

The information of Table 3 reveals the difference in the transition to adulthood in gender and education level. From state A: 'Leaving home', high educated females are 0.01 faster (transition probability 0.16 vs. 0.15) to state B: 'Family formation' than the low educated males and similar pattern can be found in high educated males against low educated males. Compared with males, females are faster in transition from 'Leaving home' to 'Family formation'. From state B: 'Family formation', high educated females are 0.05 slower (transition probability 0.08 vs. 0.13) to state C: 'Family extension' than low educated females, and similar pattern can be found in high educated males against low educated males. Besides, males moves slower from 'Family formation' to 'Family extension' than females. Transition from state 'Family extension' to state D: 'Family completion' shows different pattern than from state 'Family formation' to state 'Family extension'. From State 'Family extension' to state 'Family extension', high educated females are 0.05 (0.19 vs. 0.14)than low educated females, and similar pattern can be found in high educated males against low educated males. For this transition, males are faster than females. To summarize, (1) high educated move out of parental home faster than low educated, females faster than males; (2) high educated start having child and change their partnership status slower than low educated, females faster than males; (3) high educated are faster having more child once they have one child, males faster than females.

4 Conclusion and Discussion

Most peoples life courses are made up of a multitude of changes in multiple life domains. A key challenge of life course research is to make sense of this complexity by searching for fundamental processes that drive these observable transitions and by examining which factors influence them. In this paper, we claim that Hidden Markov modeling (HMM) holds great promise in unraveling these processes, and we provide a relatively simple example of its potential by applying it to the family transition into adulthood among French men and women born between 1956 and 1965.

From a substantive point of view, what the HMM results reveal is that two fundamental viewpoints on the transition to adulthood can be distinguished. The HMM solution with four hidden states views the family transition to adulthood as a process that leads to the intergenerational reproduction of family life. The first challenge that young adults face is about leaving the parental home and finding a suitable partner relationship. The next steps in this intergenerational reproductive process are about

the initiation of a family (entry into parenthood), followed by successive phases of family expansion and family completion. Thus, the 4-HMM solution suggests a model of the full family cycle starting as a child in a family of origin and ending up as an adult in a next generation family. Our analysis also reveals clear differences in the speed and likelihood of making this transition between men and women and between the higher and lower educated. For instance, higher educated women make the first fundamental transition (out of the parental home) at earlier ages than lower educated women, but postpone the establishment of a family of their own. However, once they decide to establish a family, higher educated women are faster in making the family expansion step. Thus, our analysis shows that the pace and rhythm of this fundamental family succession model differs strongly between low and high educated women.

The 5-HMM solution (not presented) provides another interesting view on the family transition into adulthood. Rather than viewing this transition as a unilinear trajectory where young adults only differ in the likelihood and speed of moving to successive stages as is central to the 4-HMM solution, the 5-HMM solution distinguishes between two alternative family pathways into adulthood. As in the 4-HMM solution, the first challenge every young adult faces is when to leave the parental home. One pathway strongly resembles the traditional pathway where young adults first establish a traditional family, characterized by marriage and possibly a child, followed by a subsequent stage of family expansion. However, a second pathway is distinguished as well, where young adults opt for a more autonomous lifestyle, characterized by single living and/or unmarried cohabitation. After this stage, these young adults are confronted by another fundamental choice, either to continue this alternative lifestyle track and opt for children outside marriage, or to align themselves into the traditional pattern by moving back into the traditional family pathway. As with the 4-HMM solution, linking covariates to the 5-HMM structural model offers interesting insights. For instance, highly educated women are more likely to start off on the alternative track than low educated women, but once they enter this track, they are also more likely to revert to the traditional pattern than low educated women who start off on the alternative track.

Whether one interprets the data on the basis of the 4-HMM or 5-HMM solution at least partly depends on ones theoretical interests. The 4-HMM solution offers a succinct interpretation of the traditional family life pattern, pointing at three major decisions to be taken in the course of the family-life cycle [Glick, 1955]. The 5-HMM solution incorporates more heterogeneity into this family life cycle [Glick, 1989], and offers interesting opportunities to study the process of family change that is often captured under the heading of the Second Demographic Transition [Lesthaeghe, 1995].

A major advantage of both of these models is that they greatly limit the complexity of the process of transition into adulthood, by reducing the large number of transitions between observable states to a small number of transitions between unobservable, latent states. This property could be even more useful if the number of potential states and transitions becomes even larger, for instance if one wants to study both family transitions and career-related transitions in one model.

The models introduced in this paper have clear merit for life course research. Several extensions of the Hidden Markov model could be envisaged, for example, constrained HMM. Constrained HMM is useful when one has a clear idea about the structure of the transition pattern, and want to test the hypothesized transition probability distribution. This paper did not elaborate on this type of topic yet, but it can be of great interest for future research. Generally, in applying these models to life course data, researchers have to be aware of both theoretical and practical restrictions on the analyses. Models should not become too complex in order for them to be mathematically feasible to estimate and to be theoretically interpretable. Our paper suggests a number of guidelines in this respect that may prove useful to future users.

Acknowledgement

The research leading to these results has received funding from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 324178 (Project: Contexts of Opportunity. PI: Aart C. Liefbroer).

References

[Andrieu and Doucet, 2000] Andrieu, C. and Doucet, A. (2000). Simulated annealing for maximum a posteriori parameter estimation of Hidden Markov Models. *IEEE Transactions on Information Theory*, 46(3):994–1004.

[Barban and Billari, 2012] Barban, N. and Billari, F. C. (2012). Classifying life course trajectories: a comparison of Latent Class and Sequence Analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(5):765–784.

[Bartolucci et al., 2015] Bartolucci, F., Farcomeni, A., Pandolfi, S., and Pennoni, F. (2015). LMest: an R package for latent Markov models for categorical longitudinal data. *arXiv preprint:1501.04448*.

[Bartolucci et al., 2012] Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov models for longitudinal data*. CRC Press.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41:164-71*, 41(1):164–171.

[Bijleveld and Mooijaart, 2003] Bijleveld, C. C. and Mooijaart, A. (2003). Latent Markov modelling of recidivism data. *Statistica Neerlandica*, 57(3):305–320.

[Blossfeld et al., 2007] Blossfeld, H.-P., Golsch, K., and Rohwer, G. (2007). *Event History Analysis with STATA*. Lawrence Erlbaum.

[Buchmann and Kriesi, 2011] Buchmann, M. C. and Kriesi, I. (2011). Transition to adulthood in Europe. *Annual Review of Sociology*, 37:481–503.

[Cook et al., 2000] Cook, R. J., Ng, E., and Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent markov models. *Biometrics*, 56(4):1109–1117.

[Cornwell, 2015] Cornwell, B. (2015). *Social Sequence Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, New York.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelyhood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

[Elder Jr, 1985] Elder Jr, G. H. (1985). Perspectives on the life course.

[Elzinga et al., 2007] Elzinga, C. H., Hoogendoorn, A. W., and Dijkstra, W. (2007). Linked Markov sources: Modeling outcome-dependent social processess. *Sociological Methods & Research*, 36:26–47.

[Fokkema et al., 2016] Fokkema, T., Kveder, A., Hiekel, N., Emery, T., and Liefbroer, A. C. (2016). Generations and gender programme wave 1 data collection: An overview and assessment of sampling and fieldwork methods, weighting procedures, and cross-sectional representativeness. *Demographic Research*, 34:499.

[Gabadinho et al., 2011] Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37.

[Glick, 1955] Glick, P. C. (1955). The life cycle of the family. *Marriage and Family Living*, 17(1):3–9.

[Glick, 1989] Glick, P. C. (1989). The family life cycle and social change. *Family relations*, pages 123–129.

[Langeheine and Van de Pol, 1990] Langeheine, R. and Van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research*, 18(4):416–441.

[Lesthaeghe, 1995] Lesthaeghe, R. (1995). The Second Demographic Transition in Western countries: An interpretation. *Gender and family change in industrialized countries*, pages 17–62.

[Manzoni et al., 2010] Manzoni, A., Vermunt, J. K., Luijkx, R., and Muffels, R. (2010). Memory bias in retrospective collected employment careers: a model-based approach to correct for measurement error. *Sociological Methodology*, 40(1):39–73.

[Mayer, 2009] Mayer, K. U. (2009). New directions in life course research. *Annual Review of Sociology*, 35:413–433.

[Paas et al., 2007] Paas, L. J., Vermunt, J. K., and Bijmolt, T. H. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):955–974.

[Pakpahan et al., 2015] Pakpahan, E., Hoffmann, R., and Kröger, H. (2015). Statistical methods for causal analysis in life course research: an illustration of a cross-lagged structural equation model, a latent growth model, and an autoregressive latent trajectories model. *International Journal of Social Research Methodology*.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Richardson et al., 2011] Richardson, K., Harte, D., and Carter, K. (2011). *Understanding health and labour force transitions: Applying Markov models to SoFIE longitudinal data*. Statistics New Zealand.

[Spedicato et al.,] Spedicato, G. A., Kang, T. S., and Yalamanchi, S. B. The markovchain package: A package for easily handling discrete Markov chains in R.

[Vermunt, 1997] Vermunt, J. K. (1997). *Log-Linear Models for Event Histories*. Advanced quantitatiev Techniques in the Social Sciences 8. Sage, Thousand Oaks.

[Vermunt et al., 1999] Vermunt, J. K., Langeheine, R., and Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2):179–207.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions of Information Theory*, IT-13:260–269.

[Zucchini and MacDonald, 2009] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov models for time series: an introduction using R*. CRC press.

Table 1: HMM model with 4 hidden states output parameter: Transition probability distribution, initial probability distribution and emission probability distribuion (ordered).

Transition probability distribution

State	State			
	A	B	C	D
A	.85	.14	.01	0
B	0	.91	.09	0
C	0	0	.85	.15
D	0	0	0	1

Initial probability distribution

State	A	B	C	D
	.94	.06	0	0

Emission probability distribution

Fertility	State			
category	A	B	C	D
0	.99	1	0	0
1	.01	0	1	0
2	0	0	0	.68
3+	0	0	0	.33
Partnership	State			
category	A	B	C	D
S	.97	.53	.17	.11
U	.02	.27	.25	.14
M	.01	.20	.59	.76
Leaving home	State			
category	A	B	C	D
No	1	0	.02	.01
Yes	0	1	.98	.99

Table 2: HMM model 4 with covariates output parameter: initial probability distribution and emission probability distribuion (ordered).

Initial probability distribution				
State	A	B	C	D
	.94	.06	0	0
Emission probability distribution				
Fertility	State			
category	A	B	C	D
0	.99	1	0	0
1	.01	0	1	0
2	0	0	0	.67
3+	0	0	0	.33
Partnership	State			
category	A	B	C	D
S	.97	.50	.17	.11
U	.02	.29	.25	.14
M	.01	.21	.58	.75
Leaving home	State			
category	A	B	C	D
No	1	0	.02	.01
Yes	0	1	.98	.99

Table 3: HMM model 4 with covariates output parameter: Transition probability distributions of low educated males, high educated males, low educated females and high educated females (ordered).

Transition probability distribution of low educated males

State	State			
	A	B	C	D
A	.86	.13	.01	0
B	0	.91	.09	0
C	0	0	.86	.14
D	0	0	0	1

Transition probability distribution of high educated males

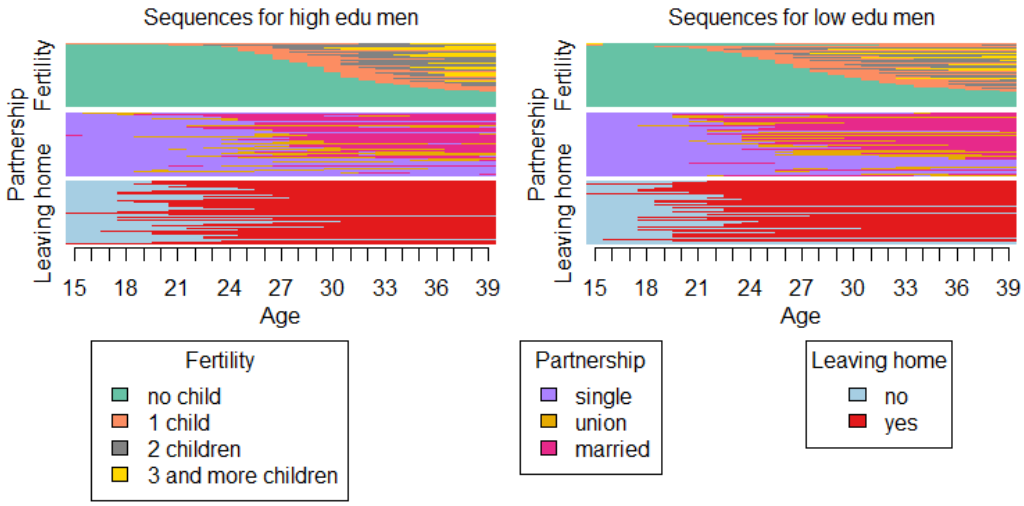
State	State			
	A	B	C	D
A	.85	.14	0	0
B	0	.94	.06	0
C	0	0	.81	.19
D	0	0	0	1

Transition probability distribution of low educated females

State	State			
	A	B	C	D
A	.84	.15	.01	0
B	0	.87	.13	0
C	0	0	.86	.14
D	0	0	0	1

Transition probability distribution of high educated females

State	State			
	A	B	C	D
A	.83	.16	.01	0
B	0	.91	.08	0
C	0	0	.82	.19
D	0	0	0	1



(a)

(b)

Figure 5: Multi-channel sequence presentation of the fertility, partnership and leaving home annual data of respondents between age 15 and 40 in France GGP birth cohort 1956-1965 (a) are sequences for high educated and low educated males. (b) are sequences for high educated and low educated females.

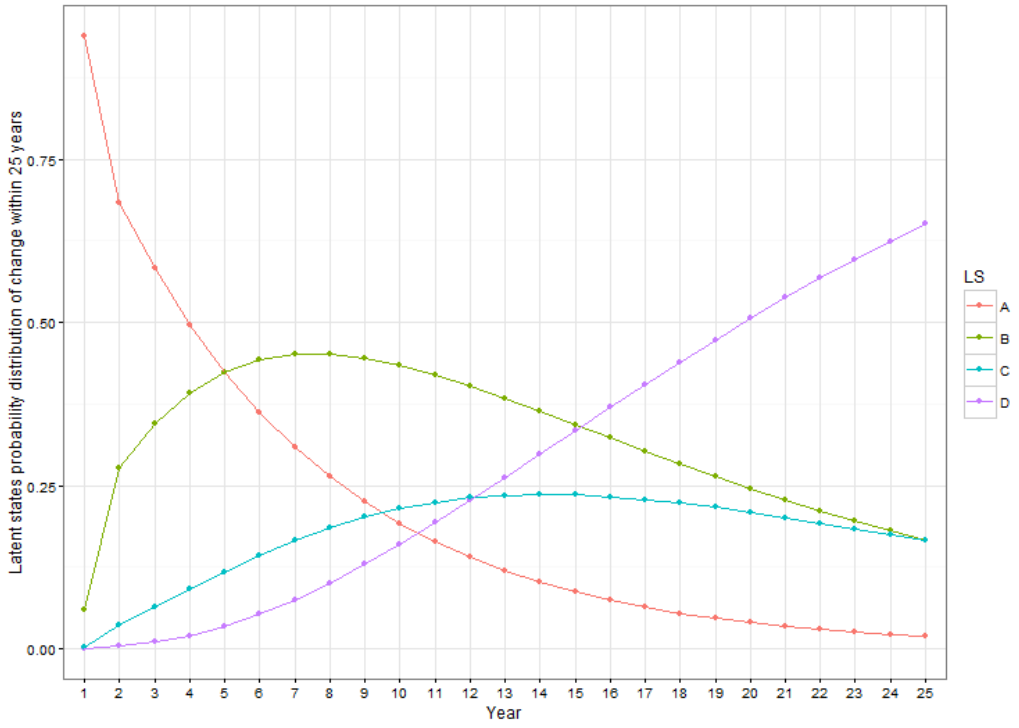


Figure 6: Latent probability distribution of change in the 25 years of HMM 4 based on the initial probability and transition proability distribution of HMM 4. State ordering corresponding to Table 1.

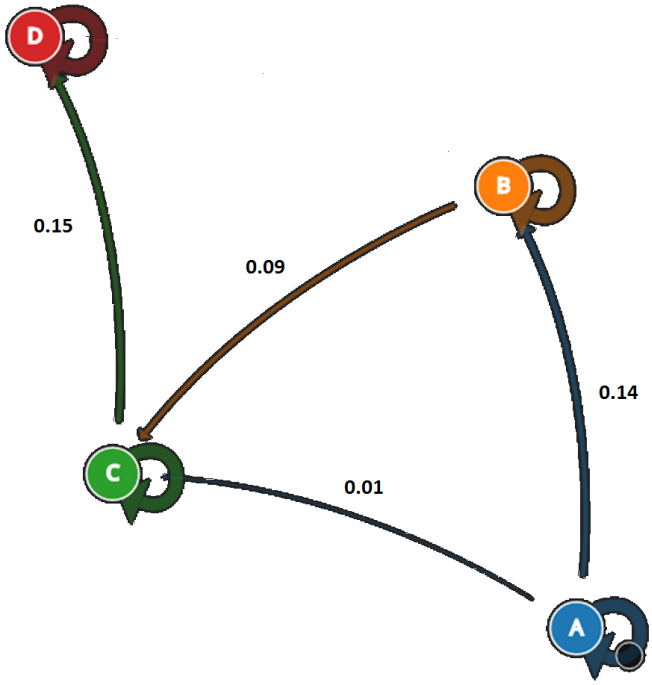
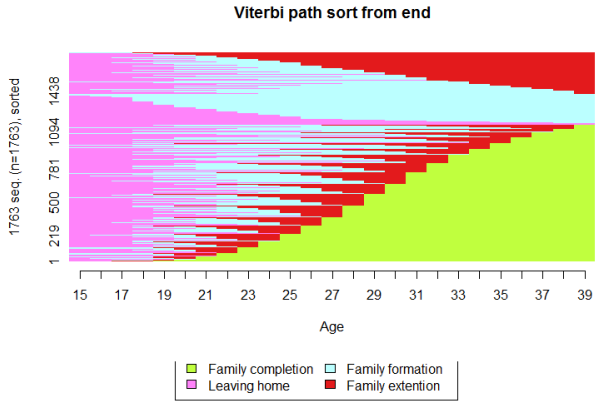
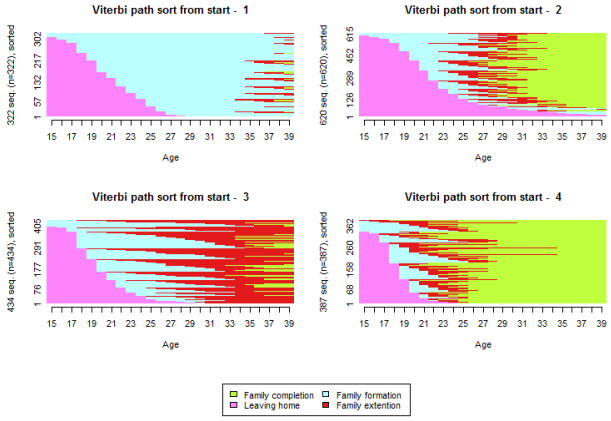


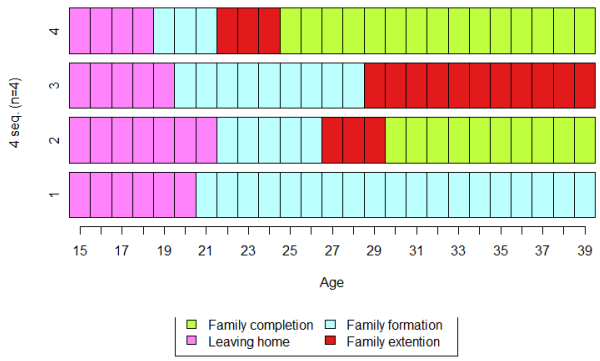
Figure 7: State transition graph based on the transition probability distribution of HMM 4. Thickness of the arrows reflects the transition probabilities (the transition probability not to itself are shown next to the arrow). State ordering corresponding to Table 1.



(a)



(b)



(c)

Figure 8: a: Sequence index plot of Viterbi path of all unique multi-channel sequence in the selected French GGP dataset. b: Sequence index plot of classified (Sequence Analysis) viterbi path. c: Sequence medoid plot of classified (Sequence analysis) viterbi path.