Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data

Satu Helske, Jouni Helske and Mervi Eerola

Book Chapter

Available at: Linköping University Institutional Repository (DiVA)
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-152155

Tweet

LiU LINKÖPING UNIVERSITY

Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data

Satu Helske, Jouni Helske, and Mervi Eerola

1 Introduction

Longitudinal data often consists of multiple parallel sequences that ought to be analyzed jointly. For example, life course data may contain sequences of employment, family formation, and residence. Such data is often referred to as multichannel or multidimensional sequence data. A multichannel approach often gives a simpler representation of the data as opposed to combining states across life domains (the extended alphabet approach); the latter approach rapidly grows the state space as the number of channels and/or states grows. If some data is only partially observed, the multichannel approach also allows for handling data as it is instead of having to make difficult decisions on how to combine observed and unobserved states (Helske and Helske 2018).

Joint analysis of complex multidimensional data poses several challenges. Multichannel sequence analysis (Gauthier et al. 2010) has been the standard tool for the analysis of multichannel sequence data (for empirical applications see, e.g., Eerola and Helske 2016; Müller et al. 2012; Spallek et al. 2014). This approach is

S. Helske (✉)
Institute for Analytical Sociology, Linköping University, Linköping, Sweden

Department of Sociology, University of Oxford, Oxford, UK

Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland
e-mail: satu.helske@liu.se

J. Helske
Department of Science and Technology, Linköping University, Linköping, Sweden

Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland

M. Eerola
Centre of Statistics, University of Turku, Turku, Finland

simple and fast in computing dissimilarities between sequences, and cluster analysis is often used for grouping similar sequences. Describing and visualizing results is, however, often challenging.

We propose an approach for compressing the information within multichannel sequences and for facilitating the interpretation of such data by finding (1) groups of similar trajectories and (2) similar phases within trajectories belonging to the same group. For the first task we use the standard multichannel sequence analysis approach and for the second task we propose using hidden Markov models (HMMs). With the help of HMMs the data can then be illustrated with a graph showing typical phases within trajectories and the transitions between them and/or shown as simplified (single-channel) trajectories consisting of these typical phases. We illustrate this approach with an empirical application to complex longitudinal life course data but such an approach, and HMMs in general, are useful in various longitudinal problems across disciplines.

Hidden Markov models have been widely used in economics, bioinformatics, and engineering (see, e.g., MacDonald and Zucchini 1997; Durbin et al. 1998; Rabiner 1989), often to study single long sequences such as time series. In social sciences, such models are commonly referred to as latent Markov (chain) models (Wiggins 1955, 1973; Van de Pol and De Leeuw 1986); typically they have been used for analysing panel data with a few measurement points. In the social science framework, Vermunt et al. (1999) extended the HMM to include individual covariates and Bartolucci et al. (2007) further developed it for multichannel observations. See also Taushanov and Berchtold (2018) in this bundle.

Hidden Markov modelling have been applied in various longitudinal settings; for accounting for measurement error and unobserved heterogeneity (e.g., Van de Pol and Langeheine 1990; Poulsen 1990; Breen and Moisio 2004; Vermunt et al. 2008; Pavlopoulos and Vermunt 2015), for finding latent sub-populations (e.g., Van de Pol and Langeheine 1990; McDonough et al. 2010; Bassi 2014), and for detecting true unobservable states (e.g., various periods of the bipolar disorder in Lopez 2008).

To the best of our knowledge, few papers apply HMMs to multichannel social sequence data and they all consider binary observations. Bartolucci et al. (2007) studied criminal trajectories of 11,400 offenders, applying HMMs to ten-channel data with six time points. Ip et al. (2015) analysed and classified 18-item profiles of food security among 248 Latino farm worker households in the USA for eight time points. Rijmen et al. (2008) studied 12 parallel trajectories of emotions at 63 time points among 32 anorectic patients. Our analysis extends this framework into multichannel data with much longer and multinomial sequences.

The rest of the paper is structured as follows. We start by giving an introduction to HMMs (we assume that the reader is familiar with sequence analysis and refer to the introduction chapter in this book for the less experienced). We then proceed to framing our goals in the context of complex life course data. We continue by describing the data and the empirical analysis and show the results. We conclude with discussing the usefulness of the method, the challenges it poses, and mention some future directions.

2 Hidden Markov Model

In hidden Markov models, observations are related to a hidden process following a Markov chain. Hidden states can only be detected through the observed sequence(s), as they generate or "emit" observations on varying probabilities.

Let us assume we have multichannel sequence data with N individuals, T timepoints, and C channels and a hidden Markov model with S hidden states. Now $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iT})$ represents the hidden state sequence for individual $i = 1, \ldots, N$ from time 1 to time t and y_{itc} denotes the observation of individual i at time $t = 1, \ldots, T$ in channel $c = 1, \ldots, C$.

Figure 1 illustrates the structure of an HMM for two-channel data. The first order Markov assumption states that the probability of transitioning to the hidden state at time t only depends on the hidden state at the previous time point $t-1$. Here we also assume the same latent structure applies to all channels, i.e., hidden state z_{it} emits observed states y_{itc} in all channels c and observations $y_{it1}, \ldots, y_{itC}$ are assumed conditionally independent given the hidden state z_{it}.

The following probabilities characterize a discrete first-order hidden Markov model for multichannel data:

- *Initial probability* vector $\pi = \{\pi_s\}$ of length S, where π_s is the probability of starting from the hidden state s:

$$\pi_s = P(z_{i1} = s); \quad s \in \{1, \ldots, S\}.$$

- *Transition probability* matrix $A = \{a_{sr}\}$ of size $S \times S$, where a_{sr} is the probability of moving from the hidden state s at time $t-1$ to the hidden state r at time t:

$$a_{sr} = P(z_{it} = r | z_{i(t-1)} = s); \quad s, r \in \{1, \ldots, S\}.$$

- *C emission probability* matrices $B_c = \{b_s(m_c)\}$ of size $S \times M_c$, where $b_s(m_c)$ is the probability of the hidden state s emitting the observed state m_c in channel c and M_c is the number of observed states in channel c:

$$b_s(m_c) = P(y_{itc} = m_c | z_{it} = s); \quad s \in \{1, \ldots, S\}, m_c \in \{1, \ldots, M_c\}.$$
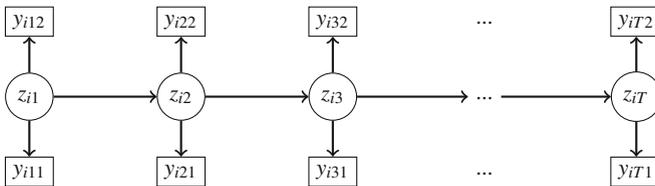

Fig. 1 Illustration of hidden and observed state sequences in a hidden Markov model for two-channel data of individual i. The hidden state at time t is illustrated with z_{it} inside a circle and the observed state at time t in channel c with y_{itc} inside a rectangle. Arrows indicate dependencies between states

Typically, the maximum likelihood estimates of these probabilities are calculated with the Baum–Welch algorithm, i.e., the expectation–maximization (EM) algorithm for HMMs (Baum and Petrie 1966; Rabiner 1989). The most probable path of hidden states for each subject given their observations and the model can be computed using the Viterbi algorithm (Viterbi 1967; Rabiner 1989). Missing observations are handled straightforwardly. When observation y_{itc} is missing, it does not contribute to the estimation of model parameters nor hidden states. See Helske and Helske (2018) for a more extensive presentation on HMMs for multichannel data.

3 Combining Sequence Analysis and Hidden Markov Models for Complex Life Sequences

For analysing complex life sequence data, we aim to compress the information into two types of components:

1. groups with similar life course patterns and
2. typical life stages within each group.

The first component corresponds to finding clusters or latent classes of individuals who have experienced similar life events in similar order and timing. The other, time-varying components should correspond to life stages during which individuals are more likely to have similar experiences, e.g., observed states within the sequences. These life stages could be either stable episodes between two transitions (e.g., employed and married without children) or characterized by transitions in some of the life domains (e.g., moving between unemployment and short-term jobs). Individuals may, and typically do, go through several different life stages during their life course.

SA followed by cluster analysis is a typical strategy for grouping life trajectories. Hidden Markov models, in turn, may be used for finding time-varying latent structures and transitions between them. At first, we use multichannel SA to compute pairwise dissimilarities and then group individuals into clusters. Separate HMMs are then fitted for each cluster. The number and nature of the hidden states are determined independently for each group.

We estimate left-to-right HMMs where transitions to previous hidden states are not allowed. We had several reasons to do this. First, left-to-right models are simpler to estimate since some of the parameters are restricted to zero. Second, due to the nature of the life trajectories, also the observed states tend to show a left-to-right behaviour and many of the HMMs would end up being estimated close to left-to-right models anyway. Third, we find that left-to-right models are often easier to interpret in the context of life course: individuals go through different life stages but even if they return to have a similar life stage compared to a previous one – say re-marriage after a divorce – this second life stage comes with a different history compared to the first time.

4 Data

We illustrate the analysis of complex life sequence data using a subsample of the German National Educational Panel Survey (NEPS) (Blossfeld et al. 2011). We restricted the analysis to the life courses of an age cohort born in 1955–1959. Only individuals who were born in Germany or moved there before the age of 14 are included.

The data consisted of monthly life statuses of 1731 individuals in three life domains (labour market participation, partnerships, and parenthood) from age 15 to age 50. For each individual, there were three parallel sequences of length 434, which made altogether 2,253,762 data points (of which 2,232,730 were observed and 21,032 were missing). Using the monthly time scale also allowed for the detection of smaller fluctuations in life courses, e.g. recurrent transitions between short-term unemployment and employment.

4.1 Sequences

The sequences in three life domains were constructed as follows:

Labour market participation with 4 states:

- Studying (in school, vocational training, or vocational preparation)
- Employed (full-time or part-time)
- Unemployed
- Out of the labour market (for other reason than studies, e.g., parental leave, taking care of children or other family members, military or non-military service, voluntary work, or other gap in the employment history)

Partnerships with 4 states:

- Single (never lived with a partner)
- Cohabiting
- Married/in a registered partnership
- Divorced/separated/widowed

Parenthood with 2 states:

- No children
- Has (had) children (biological, adopted, or foster children)

The coding for parenthood was very simple. A practical reason was that this record was available for most individuals, whereas more detailed information was often missing. On the other hand, we can argue that specifically the experience of becoming a parent is relevant as one step in the developmental process into adulthood.

For the latter two life domains, the status of each month was usually determined from the latest event. An exception was made for the rare partnerships that lasted for less than a month; there separation was coded from the following month onward. In a case of multiple records per month in the career domain, the final status was given according to assumed importance: school and vocational training came before employment, which in turn dominated over vocational preparation, unemployment, and other non-employment statuses.

Altogether 306 individuals (17.7%) had some missing information in one or two life domains. Thus, at each time point we had at least some information from each individual.

5 Analysis

We have little prior knowledge on the structure of the model; hence, how many clusters to choose and how many hidden states to include in each cluster? Since the complexity of these types of life course trajectories varies a lot (e.g., some individuals have no family-related transitions while others have many), we expected the groups to have varying numbers of hidden states.

5.1 Sequence Analysis and Clustering

We started by applying multichannel sequence analysis and computed the dissimilarities between the sequences. These were then used in cluster analysis.

The dissimilarities between sequences were determined according to the generalized Hamming distance with user-defined substitution costs (see Table 1). We set the highest cost to be the same in all life domains to give them equal weight. We gave no cost for substituting missing states since we wanted to determine dissimilarity based on the observed trajectories. Regarding the costs within different life domains, our choices were mainly based on how far the states are regarded on the pathway to adulthood and, in terms of labour market participation, also on how close the other states can be regarded to employment which is often the favourable state. The metric compares observed states time point by time point and gives a cost for mismatches. It generally works well in a multichannel problem where timing is important (Studer and Ritschard 2016) and resulted in meaningful clusters with high goodness-of-fit.

We used Ward's clustering method for the Hamming dissimilarities and chose six clustering solutions with 7–12 clusters for further examination. The choice was based on goodness-of-fit statistics, the dendrogram, and the interpretability of the clusters. Ward's method was chosen because it typically produces usable and relatively even-sized clusters compared to most of the other clustering methods (Aassve et al. 2007; Helske et al. 2015). Also, the method is hierarchical (agglomerative), so when two smaller clusters are merged, all other clusters remain the same. This

Table 1 Substitution costs for Hamming distances in three life domains: labour market participation, partnerships, and parenthood

Labour market participation

	→ ST	→ EM	→ UN	→ OU	→ *
Studying (ST) →	0	3	2	1	0
Employed (EM)→	3	0	2	2	0
Unempl. (UN) →	2	2	0	1	0
Out of LM (OU)→	1	2	1	0	0
Missing (*)→	0	0	0	0	0

Partnerships	→ S	→ C	→ M	→ D	→ *
Single (S)→	0	2	2	3	0
Cohab. (C)→	2	0	1	2	0
Married (M)→	2	1	0	2	0
Div./sep. (D)→	3	2	2	0	0
Missing (*)→	0	0	0	0	0

Parenthood	→ NC	→ CH	→ *
No child (NC)→	0	3	0
Has child (CH)→	3	0	0
Missing (*)→	0	0	0

means that among the $7 + 8 + 9 + 10 + 11 + 12 = 57$ clusters in the six sets of clustering results, only $7 + 2 + 2 + 2 + 2 + 2 = 17$ were unique, resulting in significant decrease in the number of models to be estimated compared to non-hierarchical clustering.

5.2 Hidden Markov Models for Clusters

At the next step, we estimated five HMMs with 4–8 or 5–9 hidden states separately for each of the 17 unique clusters—fewer hidden states for clusters with simpler observed trajectories, more for the more complex ones. Since the goal was to find general life stages between adolescence and middle age in a given group, having too few or too many hidden states was not plausible nor interpretable. When increasing the number of hidden states, at some point they lost their distinctive nature (consecutive states had very similar emission probabilities) and/or they were rarely "visited" in the most probable paths of hidden states.

A well-known problem with the HMM estimation is that most of the optimization methods are sensitive to initial estimates of the parameters. In order to reduce the risk of being trapped in a poor local optimum, we estimated the models numerous times with random starting values. We continued re-estimation until we had found the same optimum for at least 100 times (which turned out to be much more than necessary).

For each cluster, we compared the HMMs with a different number of hidden states to find the best model. Bayesian information criterion (BIC) and other information criteria are common choices for comparison of HMMs with different numbers of hidden states. Another common option for model selection is cross-validation.

We chose to use BIC as it generally selects parsimonious models. Unfortunately, here BIC kept suggesting models with more and more states. We did, however, use BIC as one source of information for choosing the number of hidden states by looking for turning points in BIC after which additional hidden states offered little improvement. In addition to BIC, the choice of the number of hidden states was based on the interpretability of the model and the prevalence of the hidden states in the individual trajectories.

5.3 Software

Analyses were conducted with the R software (R Core Team 2015) by using the packages TraMineR (Gabadinho et al. 2011) for sequence analysis, cluster (Maechler et al. 2015) for cluster analysis, and seqHMM (Helske and Helske 2018) for hidden Markov modelling. For the estimation of HMMs we used the automatic re-estimation routine for the EM algorithm provided in the model estimation function.

6 Results

The number of hidden states per cluster varied between six and eight. The model of eight clusters resulted in the smallest BIC (even the highest likelihood) and was chosen as the best model. We present a few different ways to describe the results: a table showing the most typical transitions in each cluster, a figure illustrating the structure of the HMMs, and a figure of the most probable hidden states, i.e., the trajectories of general life stages for each individual.

Table 2 describes each cluster in terms of some important transitions and states: typical labour market participation (showing the timing of completing education and the type of employment after that), partnership histories (age at first partnership, the type and number of partnerships), and parenthood (the timing of the first child). It also shows the number of individuals in each cluster and the proportion of the sample, as well as the hidden states described with the most important transitions.

Figure 2 illustrates the HMM structure for each of the eight clusters. It shows the HMMs as directed graphs where the pies represent hidden states and the slices show the emission probabilities of observed states within each hidden state (to draw attention to the most prevalent observations, we only show probabilities that are greater than 0.05). The arrows indicate transition probabilities between the hidden states—the thicker the arrow, the higher the probability.

Figure 3 illustrates the most probable hidden state paths. We have assigned similar colours to similar hidden states across clusters.

As an example of how to interpret these figures, let us look at the smallest of the clusters titled Single parents (cluster H). All individuals start from the first hidden

Table 2 Description of clusters by typical timing of the completion of education, type of employment, the timing, number, and types of typical partnerships, and the timing of parenthood. Hidden states are described with changes in the most probable observations (ordered by prevalence) (out = out of the labour market (not studying), div. = divorced or separated). For all clusters, the first two hidden states are omitted as they are approximately the same: hidden state 1 is *studies, single, no children* and hidden state 2 is *employed, single, no children*

| Clusters	Education	Employment	1st partnership	Partnerships	Parenthood	N	%	Women (%)	Hidden states
Short education & early family (A)	Before 20	Mostly employed	Early 20s	1 or 2 marriages / marriage + cohab.	Early 20s	461	27	59	3. Empl./studies, married/cohab. → 4. Empl., married, child → 5. Div./cohab. → 6. Unempl./out/studies, married → 7. Empl.
Short education & later family (B)	Early 20s	Steady employment / many out of empl.	Mid-20s	1 long marriage	Around 30	403	23	46	3. Cohab./div. → 4. Married → 5. Empl./out, child → 6. Out → 7. Empl.
Long education & later family (C)	Mid-20s	Mostly empl. / some out of empl.	Varying	Long cohab., 1 long marriage	30s	266	15	32	3. Empl./studies, cohab./div. → 4. Married → 5. Out/empl./unempl./studies, child → 6. Empl., cohab./div. → 7. Empl., married
Career break & early family (D)	Before 20	Out of empl., some employed after 35	Early 20s	1 long marriage	Early 20s	159	9	96	3. Married/cohab. → 4. Empl./out, child → 5. Out, married → 6. Unempl./studies → 7. Empl. → 8. Div.
Partnership(s) & no child (E)	Varying	Varying, but mostly employed	Early 20s	1 long marriage / multiple partners	No child	177	10	49	3. Empl./studies, cohab. → 4. Empl., married → 5. Div. → 6. Cohab. → 7. Empl./out/unempl., married/div.
No or late family (F)	Varying	Mostly employed / some long unempl. after 35	Never / After 35	0 / 1 cohab. or marriage	Never / After 40	116	7	41	3. Unempl./studies/out → 4. Empl. → 5. Cohab./div. → 6. Married, no child/child
Divorced / separated parents (G)	Varying, mostly before 20	Mostly employed, some out of empl.	Varying, mostly early 20s	1 cohab. or marr., div. / sep. during 30s	Varying, typically late 20s	102	6	61	3. Empl./studies, cohab./married/div. → 4. Married, child → 5. Empl./unempl./out, div. → 6. Empl., cohab./married → 7. Div.
Single parents (H)	Before / early 20s	Mostly employed, some out of empl.	Never / After 35	0 / 1 cohabitation	Varying, typically late 20s	47	3	72	3. Out/empl./studies, child → 4. Empl. → 5. Unempl./empl./studies → 6. Empl., cohab./married/div.
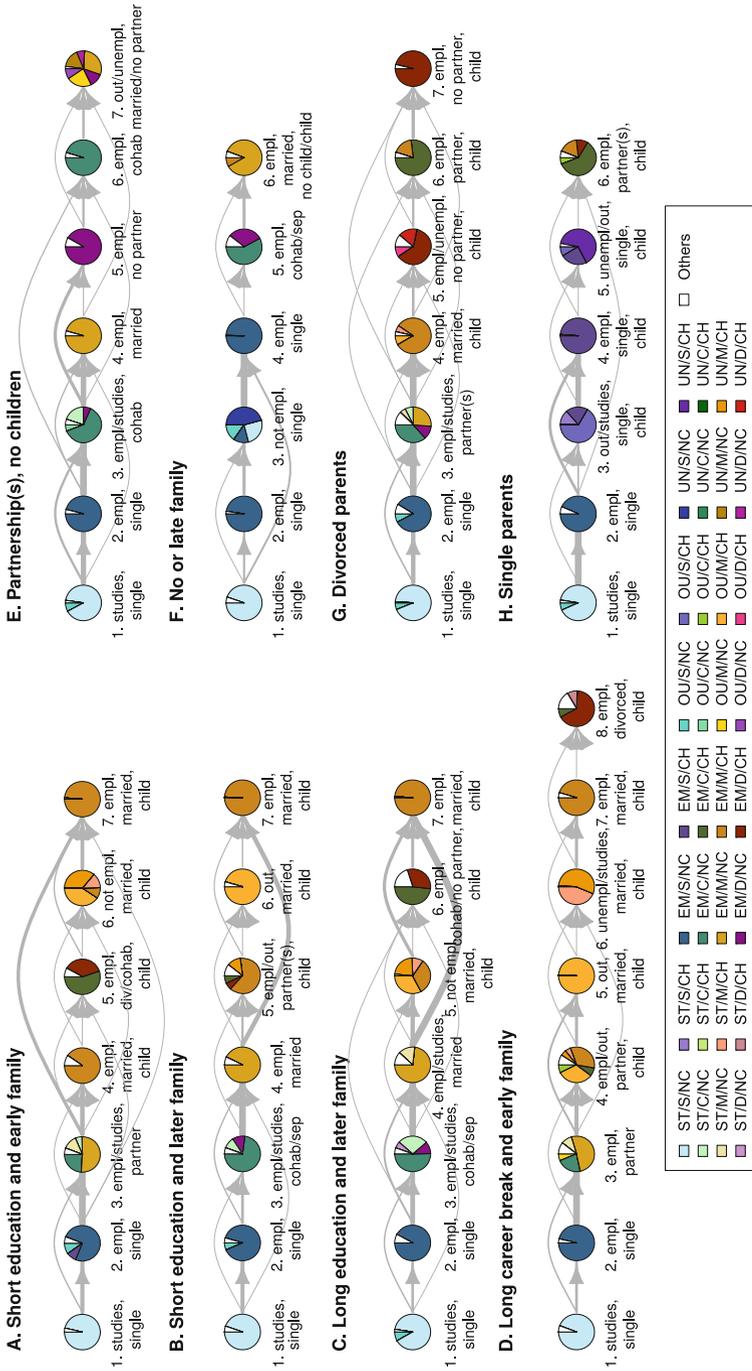
Fig. 2 HMM graphs for the eight clusters A–H. State abbreviations show labour market/partnership/parenthood statuses: ST = studying, EM = employed, OU = Out of the labour market, UN = unemployed; S = single, C = cohabiting, M = married, D = divorced/separated; NC = no children, CH = has child(ren). Hidden states are described by the most probable observations
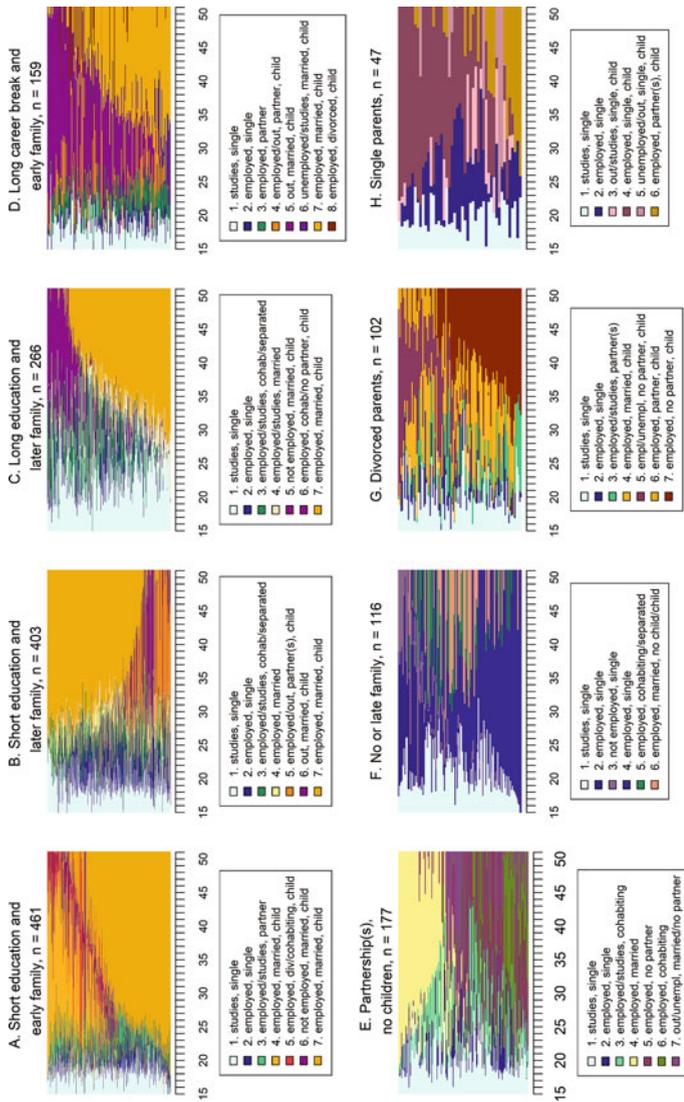
Fig. 3 Most probable hidden state paths between ages 15 and 50 for individuals in eight clusters. Hidden states are described with the most probable observed states showing labour market participation (studying/employed/unemployed/out of the labour market), partnership statuses (single/co-habiting/married/divorced/separated; also partner = cohabiting or married, no partner = divorced/separated from marriage or after cohabitation), and parenthood status (if has had children). Multiple relevant observed states within a life domain are ordered by emission probabilities. See Fig. 2 for visualizations of the hidden states in each cluster

state (State 1, indicated with light blue in the hidden state paths), a life stage where they are childless singles and mostly studying. For almost all, the next transition is to State 2 (dark blue), moving to employment. A few make a straight transition to State 3 (light pink), a life stage of becoming parents and being out of the workforce. State 4 (darker purple) describes a life stage during which individuals are singles, have children, and are employed. This is the most prevalent life stage for the members of this cluster and many stay there until the end of the follow-up. A few move out of employment, mostly to unemployment (State 5, light purple). During the last life stage, experienced by almost half of the members, individuals form their first partnerships (State 6, yellow).

In general, the clusters were well separated from each other by the timing and occurrence of labour market participation and family states. The two largest clusters composing of half of the respondents were characterized by (mostly) short education and family. The biggest difference was in the timing of partnership and parenthood transitions which occurred either earlier in life (cluster A) or later (cluster B). The third largest cluster (cluster C) mostly consisted of individuals, more often men, who had long education and later family transitions. Another cluster with early family transitions (cluster D) consisted of mostly women and was characterized with a long career break for mostly taking care of children.

Two clusters were characterized by no or very late parenthood. They differed in the timing of the partnerships; the larger cluster (cluster E) had earlier first partnerships while in the smaller cluster (cluster F) partnerships were delayed or omitted altogether. The two smallest clusters consisted of parents living divorced or separated (cluster H) or single parents (cluster G).

7 Discussion

When analysing complex sequence data with multiple channels, describing and visualizing the data can be a challenge. By combining sequence analysis and hidden Markov models the information in data can be compressed into hidden states (life stages) and clusters (general patterns in life courses). Hidden states were able to capture general life stages that included not only rather stable episodes such as being employed and married with children (e.g., State 7 for cluster A) but also life stages characterized by change, e.g., moving between unemployment and short-term employment (State 3 for cluster F).

We presented two different ways of HMM-based visualizations that give complementary information but could also be shown alone—it is up to the researcher to decide which one is more informative in each case. The HMM graphs show the structure of the hidden states and the transitions between them; also parameter estimates could be easily included in the graph. The most probable paths of hidden states show individual-level information on the approximate prevalence and timing of different life stages.

Despite its usefulness as a data reduction technique, this approach comes with some challenges. A major one is the estimation of several HMMs when the number of hidden states and clusters is unknown. For these challenges, we used a few approaches. In terms of the number of clusters, we used a hierarchical clustering method which reduced the number of models to be estimated compared to non-hierarchical clustering. We then estimated a single model numerous times with randomized starting values to find the one with the highest likelihood, using parallel computation for improved efficiency.

Another issue is that we take the SA clusters as fixed. In reality, there is, of course, a lot of uncertainty which we do not take into account. Also, we do not discuss other trajectory grouping techniques besides SA. To our knowledge, there are not many methods suitable for multichannel sequence data; we experimented with latent class analysis (Collins and Wugalter 1992) which did not lead to satisfactory results. On the other hand, regarding the parameter uncertainty, in theory it is possible to compute asymptotic standard errors from the Hessian matrix obtained from the numerical optimization algorithms, but in practice the underlying asymptions are typically not met (Zucchini and MacDonald 2009).

The mixture hidden Markov model (MHMM) offers a solution to the problem of uncertainty of clustering. In the MHMM, instead of fixing individuals to the clusters defined during the SA step, we could use all data to estimate a mixture of HMMs where each individual belongs to each cluster with some probability (preferably with a large probability for one cluster and a small probability to all others). In a complex setting, SA can be used to determine the range of potential clustering structures. It can also be of aid when setting initial values for the estimation process, which is often essential when using very large models.

Although in theory the MHMM approach allows even more flexibility to the modelling and potential for more interesting ways of inference, there are some practical computational problems in the MHMM methodology. The parameter estimation of HMMs is often very sensitive to initial values, and the computational costs increase rapidly when the number of hidden states grows. These problems are even more prominent in complex MHMMs, especially when the structure of the model (in terms of the number of hidden states and/or clusters) is not known. For this study, we were not able to find stable solutions for MHMMs despite large computational resources available—the multichannel structure, long sequences, and the relatively large number of individuals in our data was too challenging a combination for parameter estimation. Nevertheless, the MHMM can be useful in other settings. It has been successfully used for simpler problems, e.g., for accounting for measurement error and unobserved heterogeneity.

An extention not covered in this paper is the inclusion of external covariates. Personal characteristics and other relevant factors, time-constant as well as time-varying, could be used to predict transition probabilities between life stages. In MHMMs, time-constant covariates may also be used to predict cluster member-ships. See, e.g., Vermunt et al. (2008) for a general presentation of such models.

We are currently studying algorithmic variations which can reduce the computational complexity of the MHMM estimation. Further research is also needed regarding model selection and the goodness-of-fit of left-to-right HMMs and MHMMs. Further theoretical and empirical studies are needed for detecting the reasons for the failure of BIC and for discovering selection criteria that are better suited for finding parsimonious HMMs.

Another topic for future research is the potential of hidden Markov models and Markovian models in general as mechanisms of generating social sequence data.

The aim of our study was to describe complex life sequence data and for that goal, the SA-HMM approach gave satisfactory results in a reasonable time. We were able to find meaningful and well-separating clusters and to visualize their complex life course information by using HMM graphs and the most probable paths of life stages for each individual.

References

Aassve, A., Billari, F. C., & Piccarreta, R. (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population/Revue Européenne de Démographie, 23*(3–4), 369–388.

Bartolucci, F., Pennoni, F., & Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 115–132.

Bassi, F. (2014). Dynamic segmentation of financial markets: A mixture latent class markov approach. In M. Carpita, E. Brentari, & E. M. Qannari (Eds.), *Advances in latent variables* (pp. 61–72). Berlin/Heidelberg: Springer.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 67*(6), 1554–1563.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). *Education as a lifelong process-the German national educational panel study (NEPS)* (Vol. 14) [Special Issue] of *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer.

Breen, R., & Moisio, P. (2004). Poverty dynamics corrected for measurement error. *The Journal of Economic Inequality, 2*(3), 171–191.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*(1), 131–157.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.

Eerola, M., & Helske, S. (2016). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research, 25*(2), 571–597.

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

Gauthier, J.-A., Widmer, E. D., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology, 40*(1), 1–38.

Helske, S., & Helske, J. (2018, forthcoming). Mixture hidden Markov models for sequence data: The seqHMM package in R. *Journal of Statistical Software*.

Helske, S., Steele, F., Kokko, K., Räikkönen, E., & Eerola, M. (2015). Partnership formation and dissolution over the life course: Applying sequence analysis and event history analysis in the study of recurrent events. *Longitudinal and Life Course Studies, 6*(1), 1–25.

Ip, E. H., Saldana, S., Arcury, T. A., Grzywacz, J. G., Trejo, G., & Quandt, S. A. (2015). Profiles of food security for US farmworker households and factors related to dynamic of change. *American Journal of Public Health, 105*(10), e42–e47.

Lopez, A. (2008). *Markov models for longitudinal course of youth bipolar disorder*. Ph.D. thesis, University of Pittsburgh, Ann Arbor, MI.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series* (Vol. 110). Boca Raton: CRC Press.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2015). *Cluster: Cluster analysis basics and extensions*. R package version 2.0.3.

McDonough, P., Worts, D., & Sacker, A. (2010). Socioeconomic inequalities in health dynamics: A comparison of Britain and the United States. *Social Science & Medicine, 70*(2), 251–260.

Müller, N. S., Sapin, M., Gauthier, J.-A., Orita, A., & Widmer, E. D. (2012). Pluralized life courses? An exploration of the life trajectories of individuals with psychiatric disorders. *International Journal of Social Psychiatry, 58*(3), 266–277.

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment: Do survey or register data tell the truth? *Statistics Canada, Catalogue No. 12–001-X, 41*(1), 197–214.

Poulsen, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing, 7*(1), 5–19.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*(2), 167–182.

Spallek, M., Haynes, M., & Jones, A. (2014). Holistic housing pathways for Australian families through the childbearing years. *Longitudinal and Life Course Studies, 5*(2), 205–226.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481–511.

Taushanov, Z., & Berchtold, A. (2018). Markovian-based clustering of internet addiction trajectories. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research, 15*(1–2), 118–141.

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology, 20*, 213–247.

Vermunt, J. K., Langeheine, R., & 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.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 373–385). Burlington: Elsevier.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*(2), 260–269.

Wiggins, L. M. (1955). *Mathematical models for the interpretation of attitude and behavior change: The analysis of multi-wave panel*. Ph.D. thesis, Columbia University, New York.

Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Oxford: Jossey-Bass.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden Markov models for time series: An introduction using R* (Vol. 110). Boca Raton: CRC Press.