UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 157

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 157

STATISTICAL ANALYSIS OF LIFE SEQUENCE DATA

SATU HELSKE

JYVÄSKYLÄ
2016

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 157

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 157

STATISTICAL ANALYSIS OF LIFE SEQUENCE DATA

SATU HELSKE

To be presented, with permission of the Faculty of Mathematics and Science
of the University of Jyväskylä, for public criticism in Auditorium S212
on October 1st, 2016, at 12 o'clock noon.

JYVÄSKYLÄ
2016

Editor: Pekka Koskela
Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FI–40014 University of Jyväskylä
Finland

ISBN 978-951-39-6757-4 (print)
ISBN 978-951-39-6758-1 (pdf)
ISSN 1457-8905

University Printing House
Jyväskylä 2016

Abstract

Life courses are studied across disciplines for understanding the implications of life transitions on different aspects of life. Life course trajectories include, e.g., family trajectories, residential histories, and occupational careers. Trajectories embed events and transitions that may be singular or repetitive. Links between events and choices in different life domains form an interdependent system, often requiring joint analysis of different dimensions.

This thesis considers and compares different statistical approaches – event history analysis (EHA), hidden Markov models (HMMs), and sequence analysis (SA) – in the analysis of complex life sequence data. EHA is the traditional method for analysing the effects of time-constant and time-varying covariates on the timing and duration of events and transitions. In hidden Markov modelling we assume a latent or hidden level, i.e., one or more unobservable statuses that may be constant or time-varying. Observed states are regarded as being generated by a hidden or latent Markov chain. SA is a more recent model-free data-mining type of approach where the focus is on the comparison of whole trajectories. It is a descriptive tool, typically used for finding and visualizing groups of individuals with similar trajectories.

These methods are described and tested with empirical analyses, e.g., to study which types of joint family and career trajectories are typical and which atypical, to find associations between individuals' childhood characteristics and their future partnership trajectories, and to compress information across various life domains into more general life stages. This thesis also presents new software for the analysis and visualization of complex sequence data.

The three approaches provide versatile information on the phenomena of interest, as the methods capture time in different ways. The choice of the method(s) depends on the type of the data and the aims of the study. Applying model-free and modelling approaches or even combining them is often beneficial as they are not substitutes but complete each other in the analysis of life course data.

Acknowledgements

The person starting this work was very different from the person finishing it. It has been a journey of discovery – there have been moments of joy and success as well as moments of frustration and despair. During these years I've grown a lot, both in scientific as well as personal life. The encouragement and support of many people has made this possible.

First, I am very grateful for Professor Mervi Eerola for farsightedly directing me to and on this path. Thank you for showing me new opportunities, for pushing me past my comfort zone, and for giving me space to grow as a scientist. Thank you for helping me find my passion.

I wish to thank Professor Fiona Steele for collaboration and for a memorable visit to the University of Bristol. Thank you for sharing your knowledge and experience with me. During a confusing time, you helped me find new joy in doing research and gave me confidence as a scientist.

I would also like to express my gratitude to Dr Katja Kokko and Dr Eija Räikkönen, whose expertise in psychology and life course research has been invaluable. Big thanks to Eija for being my "big sister in science".

My gratitude also goes to Professor Jukka Nyblom and Professor Antti Penttinen for their helpful suggestions, and to Professor Kari Auranen and Professor Jaakko Nevalainen for their reviews and valuable comments.

A number of you also helped in other ways. Professor Anneli Kauppinen gave me my first university job as a research assistant – thank you for inspiring me towards research. I am grateful to Johanna Ärje for her support as a fellow PhD student, especially during the last few months of finishing our dissertations. More importantly, big thanks for being such a good friend and for taking my mind out of everyday issues. I also want to thank all other colleagues at our department for sharing their experiences in doing research, surviving through PhD studies, and combining work and family life.

For financial support I am indebted to the Jyväskylä Doctoral Program in Computing and Mathematical Sciences (COMAS), to the Finnish Cultural Foundation, and to the Department of Mathematics and Statistics. Special thanks to the Department for being so flexible with practical issues – it made combining family and work much easier.

I warmly thank my family and friends for their support and love, and for giving me necessary breaks from thinking about work. Special thanks to my father Olli and uncle Heikki for pushing me to apply for PhD studies (even though it might have been only for the fancy hat and sword). My mother Helka for her endless faith and support. My darling Aini for helping me find joy in the simplest things, for teaching me that I can survive through anything, and for letting me play with her Legos.

My deepest gratitude goes to Jouni; my colleague, co-author, spouse, and best friend.

Without you, I do not know how I could have survived through the hassle of these past few years. Us working together has been so much fun that often it felt more like a hobby. We have had a wonderful journey together and surely will continue to have, in years to come.

Oxford, August 2016

Satu Helske

List of original publications

This thesis consists of an introductory part and the publications listed below.

I Eerola, M. and Helske, S. (2016) Statistical analysis of life history calendar data. *Statistical Methods in Medical Research.* 25(2), 571–597. doi:10.1177/0962280212461205

II Helske, S., Steele, F., Kokko, K., Räikkönen E., and 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. doi:http://dx.doi.org/10.14301/llcs.v6i1.290

III Helske, S. ja Helske, J. (2015) Mixture hidden Markov models for sequence data: the seqHMM package in R. (Submitted.)

IV Helske, S., Helske, J. and Eerola, M. (2016) Analysing complex life sequence data with hidden Markov modelling. In G. Ritschard & M. Studer (eds), *Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016*, pp 209–240.

The author of this dissertation has constructed data and performed all analyses in all of the included articles. She had the main responsibility in writing Articles II–IV and the sequence analysis part of Article I. Some of the research questions and statistical models have been formulated in collaboration with the co-authors.

Contents

Chapter 1

Introduction

Longitudinal life courses are studied across disciplines – e.g., sociology, developmental psychology, demography, economics, and epidemiology – for understanding the implications of life transitions on different aspects of life. Social, physical, or environmental exposures and choices during earlier life stages can influence, e.g., educational choices, family formation, health, and well-being later in life.

Life course data are longitudinal in nature, but data structures vary depending on data collection and statistical methods. An *event history* is a longitudinal record of the timing(s) of one or more types of events. Time can be measured as continuous or discrete. *Sequence data* may consist of individual-level time series or be collected as panel data, with continuous or categorical observations. Sometimes the timings and durations of the states are omitted altogether and the focus is merely on the order of the states or transitions. Transitions from one data type into another are often possible, although not necessarily reciprocally.

This thesis considers methods for analysing categorical life sequence data with fixed time intervals (annual or monthly observations), mainly following the social science paradigm.

1.1 Life course context

Life courses can be formally described with four concepts: trajectory, stage, event, and transition (Levy, 2005). A *trajectory* describes all that happens between two boundaries, e.g. the whole lifespan from birth to death, and can be seen as "long-term patterns of stability and change" (George, 1993). On a formal level, a trajectory is composed by a sequence of transitions and stages (Levy, 2005). An individual's life course is naturally composed of not only one but multiple interdependent trajectories describing different life domains such as education, work, family, and residence.

A *life stage* typically refers to a life period of relative stability between two transitions (e.g., marriage or unemployment). A *transition* is a relatively short period of change from one stage to another (e.g., transition into parenthood). *Events* are less clearly qualified and the definitions also vary between scientific disciplines. They are momentary occasions that can be singular or repetitive. Transitions are often referred to as events, but there are also events that do not state an apparent transition (e.g., committing a crime). From a statistical point of view, the difference between a transition and an event is often (but not always) negligible.

1.2 Life sequence data

In social science applications, the term *sequence* typically refers to successions of categorical states or events that describe a trajectory. The states come from a finite state space often referred to as the *alphabet*. The order of the states must be explicit; it need not be temporal, but usually events unfold over a period of time. Life sequences may include, e.g., family trajectories, residential histories, and occupational careers. Links between events and choices in different life domains form an interdependent system, often requiring joint analysis of different dimensions. In multidimensional or *multichannel* life sequence data, the life course of an individual is described with multiple parallel sequences representing the states in different life domains. In formal notation, y_{itc} represents the observation of individual $i,\ i = 1, \ldots, N$, at time $t,\ t = 1, \ldots, T$ in channel $c,\ c = 1, \ldots, C,$. See Chapter 5 for illustrations of multichannel sequence data.

The composition of sequences requires careful consideration and is dependent on, e.g., research questions and methods. Article II gives a more thorough discussion on the definition of states. Joint analysis of multiple life domains poses even more challenges. With methods focusing on whole sequences, a simple option is to combine states in different channels time point by time point, i.e., to grow the alphabet. If the number of combined states is moderate, this can be an informative approach. However, the number of combined states grows rapidly as the number of channels or states grows, easily resulting in difficulties in the analysis and visualization. Also, if data are not completely observed, combining missing and non-missing information into one observation is usually problematic. One would have to decide whether such observations are coded completely missing, which is simple but loses information, or whether all possible combinations of missing and non-missing states are included, which grows the alphabet even larger and makes interpretation more difficult.

Life course data are usually collected retrospectively, with respondents asked to recall the timing of events of interest. Follow-up studies register repeated observations over time and are typically regarded as producing more reliable data compared to retrospective data collection. However, such studies can be expensive and difficult to perform. The *life history calendar* (LHC; also called the event history calendar) is a data-collection tool for obtaining reliable retrospective data about life events. The advantage of an LHC is that the order and proximity of important transitions in multiple life domains can be studied at the same time. It encourages respondents to incorporate temporal changes as cues in the reporting of events; see Figure 1.1 for an illustration. The LHC approach has shown the ability to provide data of remarkably high quality (Belli, Stafford, & Alwin, 2008). Article I discusses and demonstrates statistical analysis of LHC data.

1.3 Overview on methods for categorical sequence data

This thesis considers three statistical methods that can be used for analysing categorical sequence data, namely event history analysis, sequence analysis, and hidden Markov modelling.

Traditionally, life course data have been studied with *event history analysis* (EHA), where the focus is on the timing of events or transitions. Typical examples in life course settings include the age at the first marriage, the duration of unemployment, or birth intervals of mothers.

Sequence analysis (SA) is a more recent but steadily developing approach where the focus is

Figure 1.1: A part of an artificial life history calendar between the ages 16–30. The individual moved in with her first partner (*C1*) at the age of 20 and separated after two years. At the age of 23 she moved in with a second partner (*C2*), got married to him (*M2*) the next year, and had her first and only child at the age of 24. She went to school (*SC*) until starting at the university (*U*) at the age of 19, graduating after three years. She had three part-time jobs between ages 18–21, then started at a full-time job in which she stayed until the end of the follow-up.

Marriage/cohab.	**Age**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	...	**30**
Partner(s)						*C1*	*C1*		*C2*	*M2*	*M2*		*M2*
Children	**Age**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	...	**30**
Children										*1*	*1*		*1*
Education	**Age**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	...	**30**
Type of education		*SC*	*SC*	*SC*	*U*	*U*	*U*						
Work	**Age**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	...	**30**
Full-time work							*1*	*1*	*1*	*1*	*1*		*1*
Part-time work				*1*	*2*	*3*	*3*						
⋮													

on complete sequences. A typical goal is to find groups of individuals with similar trajectories. SA has become central to the life course perspective where it has been used to understand various trajectories and crucial transitions (Gauthier, Bühlmann, & Blanchard, 2014), e.g., for analysing careers or partnership histories during the whole life course.

In *hidden Markov models* (HMMs), observed states in different life domains are regarded as being generated by unobservable hidden or latent states. Transition probabilities between hidden states follow the Markov property; the simplest and most common models are first order models, where the transition to the subsequent state depends on the current state only. The *mixture hidden Markov model* (MHMM) is a generalization of the simple hidden Markov model where multiple HMMs with different parameterizations are modelled jointly for different subpopulations.

In addition to the methods addressed in this thesis, there are also other approaches that are suitable for categorical sequence data. *Latent class analysis* (LCA; e.g. Vermunt, Tran, & Magidson, 2008) is most commonly used in cross-sectional studies but has also been applied in various longitudinal settings for identifying subpopulations in multidimensional data. It can be presented as a restricted variant of the MHMM where there is no change in the hidden process within a subpopulation (i.e., one hidden state per group). LCA does not take into account the interdependence between observations measured in different time periods; it uses each time point as a separate variable.

Semi-parametric group-based trajectory modelling (Nagin, 1999) is another method for finding groups of individuals with similar trajectories. The method can be used for studying binary trajectories (as well as count data and continuous observations) but is not well suited for categorical trajectories with more than two unordered categories.

The *Markov model* (MM; e.g. Vermunt et al., 2008) considers transition probabilities between observed states. It can also be seen as a special case of the HMM without the hidden structure (hidden states correspond to the observed states perfectly, thus sometimes also called the manifest Markov model), or as a special case in EHA with the Markov assumption. As

with HMMs, also the MM can be expanded to a mixture model with differing subpopulations. The *mover–stayer model* (Goodman, 1961) is a well-known special case of the mixture Markov model. Applying MMs to multichannel sequence data is not straightforward.

Markovian models have been extended and modified in numerous ways. E.g., the *hierarchical HMM* (Rijmen, Vansteelandt, & De Boeck, 2008) allows for hidden states and transitions at different levels. The *mixture transition distribution model* (Berchtold & Raftery, 2002; Raftery, 1985) can be used to approximate higher-order MMs. It involves a much lower number of parameters, creating models that are easier to interpret. The *double chain Markov model* (DCMM; Berchtold, 1999) includes a direct relation between observed states, i.e., the model is a combination of two Markov chains of which one is observed and the other one hidden. The main purpose of the DCMM is the modelling of non-homogeneous time series data, i.e., data where the transition probabilities are dependent on time.

Chapter 2

Event history analysis

An event history is a longitudinal record of the timing(s) of one or more types of events. Event history analysis is a model-based probabilistic approach for the study of how individual time-invariant and time-varying characteristics influence the timings of transitions or events. Event history analysis is presented and discussed from various viewpoints in several books by, e.g., Andersen and Keiding (2002), Blossfeld and Rohwer (2001), Hougaard (2012), Lee and Wang (2003), and Mills (2011).

The term *event history analysis* is used primarily in social sciences, but in other fields of application these methods go under different names. In the initial studies in biostatistics, the event of interest was death, resulting in the term *survival analysis*. Other naming conventions include, e.g., *duration analysis* in economics and *reliability analysis* in engineering.

The time before an event occurs is typically referred to as episode, spell, survival time, risk period, or waiting time. In the simplest case, only a single event of interest occurs to each subject. Usually we cannot observe the exact timing of the event for each subject; event times can be *censored* due to, e.g., death, moving away, or the end of the follow-up. Censoring can occur in various ways. Right-censoring is the most common type; in such case the event of interest is not observed by the end of the follow-up.

Most processes operate in continuous time, but in practice time is measured in discrete units. If the units are small, we can use continuous-time models, but often the exact event times are unknown within larger time units such as years. Here the focus lies on the discrete-time model which can be used as an approximation to a continuous-time model (Allison, 1982). An event happening "at time t" occurs during an interval $[t, t+1)$. For presentations of discrete-time event history models, see Allison (1982, 1984), Mills (2011), and Steele (2011), among others.

A *risk set* is the set of individuals who are being followed in the study and "at risk" of experiencing the event of interest at each time point. In the simplest single-event case it is the set of individuals who have not yet experienced the event; individuals experiencing the event are removed from the risk set. In the case of recurrent events, individuals return to the risk set after becoming at risk again (e.g., "at risk" of a new marriage after divorce).

Failure, survival, and hazard are basic concepts in the analysis of event times. We assume that Y is a positive random variable that represents the occurrence time of the event (such as *becoming a parent*); it is recorded in discrete time points, typically time intervals, $t = 1, \ldots, T$.[1] The unconditional probability that the event occurs at time t is given by the

[1]The common convention in EHA is to denote the random variable with T. Due to conflicts with other

failure function

$$f_t = P(Y = t). \tag{2.1}$$

The *survival function* is the probability that the event has not occurred before time t. It is defined as

$$S_t = P(Y \geq t). \tag{2.2}$$

The *hazard* is the probability that an event occurs at time t given that the event has not occurred prior to t. It can be expressed as a ratio of the failure and the survival functions:

$$h_t = P(Y = t|Y \geq t) = \frac{f_t}{S_t}. \tag{2.3}$$

Likewise, we can express the probability of failure at time t through the hazard and the survival:

$$f_t = h_t S_t = h_t \prod_{i=1}^{t-1}(1 - h_t). \tag{2.4}$$

Suppose that at time t (or at the beginning of the interval t) the size of the risk set is N_t and the number of events at t is d_t. The general likelihood for the hazards of right-censored data can be written

$$L = \prod_{t=1}^{T} h_t^{d_t}(1 - h_t)^{N_t - d_t}. \tag{2.5}$$

Cases experiencing the event contribute information on the probability of a failure (an event happening) and censored cases only on the probability of survival.

Explanatory covariates can be included in the hazard model by conditioning not only on the survival, but also on the covariates. For individual i with (possible time-varying) covariates $\mathbf{x}_{it}$, the hazard is now of the form

$$h_{it} = P(Y_i = t|Y_i \geq t, \mathbf{x}_{it}). \tag{2.6}$$

There are a variety of modelling possibilities available for categorical sequence data. For binary dependent variables common choices are the logit model, the probit model, and the complementary log-log model; the multinomial logit model can be used with more than one state. Brown (1975) and Allison (1982) showed that such models can be estimated with standard methods for binary data.

A general discrete-time model for the dependence of h_t on time t and a vector of covariates $\mathbf{x}_{it}$ for individual i can be written (similarly to generalized linear models)

$$g(h_{it}) = \alpha(t) + \beta' \mathbf{x}_{it}, \tag{2.7}$$

where g is a link function (such as the logit function); $\alpha(t)$ is a function of time, defining the baseline hazard; and β are the regression coefficients related to the covariates. Typical choices for the baseline hazard include polynomials and piecewise-constant functions. The latter are used throughout this work, as these are simple to model and present, and typically produce valid approximations.

---

conventions and to keep consistent throughout this work, here Y points to the observations and T is reserved for the last time point.

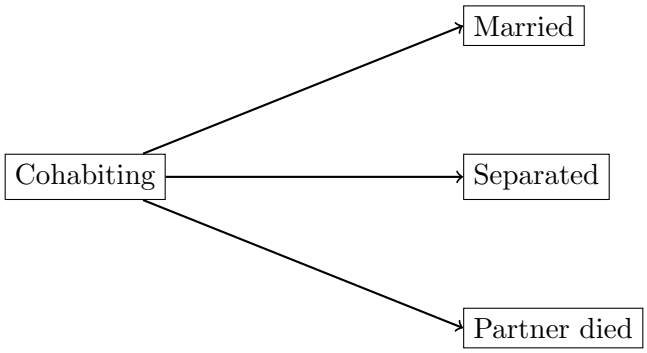


Figure 2.1: Illustration of the competing risks event history model for the outcomes of a cohabitational union.

Covariates are used to capture variation in the hazard between individuals. Often we do not include all important variables, either because they are not available or because their importance is unknown. Variability between individuals in their risk of experiencing the event of interest that is due to unmeasured characteristics is called *unmeasured heterogeneity* or *frailty* (Vaupel, Manton, & Stallard, 1979). The standard approach for allowing for frailty is to include a random effect in the model:

$$g(h_{it}) = \alpha(t) + \beta' \mathbf{x}_{it} + u_i. \tag{2.8}$$

Now the random effect u_i presents the unobserved variability between individuals, which is typically assumed to follow the normal distribution $N(0, \sigma_u^2)$. Another, although much less common option to control for unobserved heterogeneity, is to use fixed-effects models (Allison, 2009). The simple idea is to use each individual as their own control, comparing rates at different levels of covariates. Naturally, this approach requires multiple measurements per individual.

Most often, the basic model is not relevant in life course applications where events can be repeatable, individuals may move between different states, or they can make choices between competing events. Often, the interest lies in the whole history of events. The basic event history model has been extended in several ways; many of the most important developments are based on the counting process theory (e.g., Andersen, Borgan, Gill, & Keiding, 1993).

Competing risk models (or multiple destination models) refer to cases where subjects may experience an event due to a number of reasons, say, ending cohabitation due to marriage, separation, or the death of the partner (see Figure 2.1 for an illustration). There are different techniques for the analysis of such cases. In the *cause-specific model* (or the latent approach), EHA is conducted separately for each event type, while the other event types are treated as right-censored cases.

In the competing risks model we have an initial state s_0 (corresponding to *cohabiting* in the example), and S absorbing states, state $s, s = s_1, \ldots, s_S$ corresponding to "event of type s" (*married*, *separated*, and *partner died* in the example). The cause-specific hazard for event of type s is defined

$$h_t^{(s)} = P(Y = t, \text{event of type } s | Y \geq t). \tag{2.9}$$

Competing risk models are a special case of the *multistate model*. Without going into

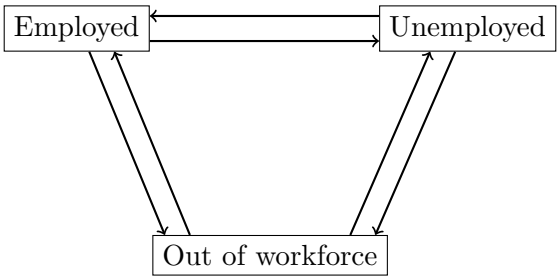


Figure 2.2: Illustration of the multistate event history model for employment.

details here, the idea is that instead of only one transition, we study (possibly) multiple transitions between two or more states, say, employment, unemployment, and out of workforce. Figure 2.2 illustrates a multistate model where all transitions between the employment states are possible.

Chapter 3

Sequence analysis

Instead of the transition-oriented EHA, in sequence analysis (SA) the focus lies on whole trajectories. The aim is to identify patterns that account for all states of interest during the entire observation period.

SA is a model-free data-mining type of method for the statistical analysis of sequences. It was originally developed in bioinformatics to organize, classify, and parse protein and DNA sequence data (see, e.g., Durbin, Eddy, Krogh, & Mitchison, 1998). SA was introduced in social sciences in the mid-1980s by Andrew Abbott (Abbott, 1983) and since then it has developed and spread to many disciplines. Recently, Cornwell (2015) gave a comprehensive overview of social sequence analysis.

The idea of SA is to measure the distance or (dis)similarity between each pair of sequences consisting of categorical states. Typical steps in SA include the following:

1. creating sequences using a finite set of states,
2. assessing the dissimilarities between sequences,
3. analysing the dissimilarities, and
4. visualizing sequence data.

The following sections focus on steps 2. and 3. Definition of states has already been discussed more thoroughly in Section 1.2 and further in Article II, and visualization of sequence data is addressed in Chapter 5 and in Articles I and III.

3.1 Assessing sequence dissimilarities

The most crucial decision in SA concerns the definition of sequence dissimilarity. Studer and Ritschard (2016) identified five sequence aspects that are important when assessing socially meaningful differences in sequences (see also Billari, Fürnkranz, & Prskawetz, 2006; Settersten & Mayer, 1997):

1. *experienced states* (distinct elements in the sequence),
2. *distribution of states* (total time in each state),
3. *timing* (appearance of states),
4. *duration* (episode lengths), and
5. *sequencing* (order of states).

Aspects 1 and 2 are related to the prevalence of states while the rest are associated to their appearance in time. The first aspect refers to the list of distinct states in a sequence. Sequences with similar states refer to individuals with shared experiences. Accordingly, if all states are regarded as equally dissimilar from all other states, two sequences with no common states (e.g., *studying–employed* and *unemployed–out of employment*) are maximally dissimilar (Dijkstra & Taris, 1995). The distribution of the states within a sequence tells about the total time spent in each distinct state. These reveal the total exposure times to each state, e.g., the total years spent in education.

In the life course framework, the meaning and influence of the occurrence of a state are typically related to age and often the interest is in the timings of transitions between states. For instance, the transition to parenthood may have very different effects on an individual's life course at ages 15 and 35, as is being in or out of education at the age of 16. Episode duration, i.e., the continuous time spent in the same state is related to the distribution of states but allows us to differentiate between long-term and short-term exposure to, say, unemployment.

Sequencing describes the order in which states appear in a sequence; e.g., the social implications of becoming a parent have been and in many cultures still are very different depending on whether it happened before or after marriage.

3.2 Dissimilarity measures

In practice, the (dis)similarity between a pair of sequences is assessed via a dissimilarity measure. Many different dissimilarity measures have been proposed; they have diverse characteristics and varying sensitivity to sequence aspects such as timing and sequencing. The choice of the measure depends on which aspects are weighted.

Optimal matching (OM; Abbott & Forrest, 1986; McVicar & Anyadike-Danes, 2002) has received the most attention so far. In OM the goal is to find the best alignment for each pair of sequences. Their dissimilarity is computed from the operations needed for editing or transforming one sequence into the other using insertions, deletions, and substitutions of states. OM is a generalization of the Levenshtein distance (Levenshtein, 1966) where the dissimilarity is the number of operations needed for the transformation. In OM, operations can be given different costs reflecting the amount of dissimilarity between the states.

The *Hamming distance* (Hamming, 1950; Lesnard, 2010) can be seen as a special case of OM where only substitutions are used; the cost of an alignment is the number of substitutions required to change one sequence into the other. In a generalization of the Hamming distance, different substitutions are weighted differently. Another special case, the *Levenshtein II distance* (Lesnard, 2010) uses only insertions and deletions; it corresponds to finding the *length of the longest common subsequence* (LCS) between a pair of sequences (Kruskal & Liberman, 1983).

Several researchers have criticised the OM approach (e.g., Elzinga, 2003; Halpin, 2010; Hollister, 2009; Lesnard, 2010; Levine, 2000; Wu, 2000). They have expressed concern over arbitrary and symmetrical transformation costs, lack of sociological meaning of edit operations, and for the representation of order and timing of states. Several efforts have been made to address these issues. Other modified "edit distances" include, e.g., *time-warp edit distance* (Halpin, 2014; Marteau, 2008), *localized OM* (Hollister, 2009), *duration-adjusted OM* (Halpin, 2010), and *dynamic Hamming distance* (Lesnard, 2010). Some methods use OM but modify the sequences. In the *transition SA* method (Biemann, 2011), sequences are constructed from the transitions between states. Another method measures the distance between *sequences of spells* (Studer & Ritschard, 2016), considering spells of different lengths as different elements.

More fundamentally different approaches that are not based on sequence alignment have also been developed. *DT coefficients* (Dijkstra & Taris, 1995) rest upon common pairs of ordered states, discarding repetitions and unshared states. Elzinga (2003) and Elzinga and Studer (2015) have proposed methods based on counting common attributes such as the *number of matching subsequences* (NMS). A generalization of the NMS, the *subsequence vector representation metric* (SVR; Elzinga & Studer, 2015), weights matching subsequences according to their length and accounts for the duration of the spells. *Euclidean distance* and χ^2*-distance* have been used for finding differences in state distributions (Deville & Saporta, 1983; Grelet, 2002).

Studer and Ritschard (2016) have compared dissimilarity measures in regard to three sequence aspects: duration, timing, and sequencing. Methods sensitive to duration include LCS (Levenshtein II) and OM as well as Euclidean and χ^2-distances. With some tuning, the latter two can also be used when the interest is in timing. Naturally, Hamming distances are also very sensitive to timing. Transition SA, OM of spells, and SVR metrics are useful for finding differences in sequencing of states.

SA has also been applied in more complex settings. *Multichannel SA* (Gauthier, Widmer, Bucher, & Notredame, 2010) and *globally interdependent multiple SA* (Robette, Bry, & Lelièvre, 2015) have been proposed for the analysis of multichannel sequence data. *Two-stage OM* (2SOM; Lesnard & Kan, 2011) is a method for analysing nested sequence data (e.g., 24-hour days within 7-day weeks in time use data).

In Article I we discuss probabilistic SA which is commonly used in bioinformatics but to our knowledge has not been applied for social sequence analysis. It can be used for assigning substitution costs in edit distances; the cost for aligning a given pair of states is defined as an odds ratio of two models, of which one assumes that the sequences are related and the other that the states occur randomly.

See Aisenbrey and Fasang (2010), Robette and Bry (2012), Elzinga and Studer (2015), and Studer and Ritschard (2016) for more detailed comparisons between different dissimilarity criteria.

3.3 Analysing sequence dissimilarities

Regardless of the chosen dissimilarity method, the result is a matrix of pairwise dissimilarities. Information in dissimilarities is always compressed in some way, typically with *cluster analysis*. The idea is to find typologies of sequences such as typical life courses. Ward's agglomerative clustering method (Batagelj, 1988; Ward, 1963) is a common choice since it usually works well with sequence dissimilarities and creates meaningful and relatively even-sized clusters

compared to other methods (Aassve, Billari, & Piccarreta, 2007, and Article II). At each step, the algorithm combines the two clusters (at the first step, sequences) that minimize within-cluster discrepancy (variance for dissimilarities; Studer, Ritschard, Gabadinho, & Müller, 2011) and maximize inter-cluster discrepancy. Another useful clustering algorithm is the partitioning around medoids (PAM; Kaufman & Rousseeuw, 2009; Studer, 2013) which seeks to minimize the sum of distances from the medoid (an observation whose average dissimilarity to all objects in the cluster is minimal).

Choosing the best number of clusters is not a straightforward task. Studer (2013) discusses and compares several measures of the quality of clustering for sequence data, e.g., the pseudo-R^2 value (Studer et al., 2011) which can be interpreted as the share of the total discrepancy of the sequences that is accounted for by the clustering (or by any covariate). Piccarreta (2015) proposes criteria for assessing the clustering of multichannel data.

Also other methods for investigating sequence dissimilarities have been considered. *Multi-dimensional scaling* (MDS; see e.g. Halpin & Chan, 1998; Piccarreta & Lior, 2010, and Article I) is a technique for visualizing the (dis)similarity of sequences in a low-dimensional space. The first MDS dimension (when rotated according to principal components) describes the most distinctive characteristic of the sequences, the second the next most distinctive characteristic, and so on. MDS is useful in ordering sequences meaningfully and for assessing the quality of clustering.

Another clustering method for sequence dissimilarities is a *self-organizing map* (SOM, also called a Kohonen map; Kohonen, 2001). See Massoni, Olteanu, and Rousset (2009) and Rousset, Giret, and Grelet (2012) for applications to sequence data. The idea behind SOM is similar to MDS: by using the dissimilarity matrix, sequences are projected on a low-dimensional discretized representation. The result is a grid where similar sequences are in the same or neighbouring clusters.

External information can be taken into account after clustering (e.g., as predictors in a regression model) or during the clustering phase. *Regression trees* (Breiman, Friedman, Olshen, & Stone, 1984) have been used for discovering the most significant discriminant covariates (Studer et al., 2011, and Article II). The idea of regression trees is to recursively partition data into clusters using the values of a predictor. Binary splits for the values of a variable are created so that the highest proportion of variation is explained (measured with pseudo-R^2).

Also the ANOVA-like *discrepancy analysis* framework (Studer et al., 2011) can be used for studying the relationship between sequence dissimilarities and a set of covariates (a similar approach called *analysis of dissimilarity* was also published by Bonetti, Piccarreta, & Salford, 2013). This approach gives information on which covariates are significant in explaining differences between sequences, but unlike regression trees, it is not very useful in showing what theirs effects are.

Chapter 4

Hidden Markov modelling

Hidden Markov models (HMMs) have been widely used in economics, bioinformatics, and engineering (see, e.g., Durbin et al., 1998; MacDonald & Zucchini, 1997; Rabiner, 1989) to study time series or other types of single sequences. In social sciences, such models are commonly referred to as latent Markov (chain) models (van de Pol & De Leeuw, 1986; Wiggins, 1955, 1973); typically they have been used for analysing panel data with a few measurement points.

HMMs are strongly connected to EHA; they are generalizations of Markov models, which in turn are special cases of event history models with the Markov assumption. The main characteristic of the HMM is that observed sequences are regarded as being generated by an unobservable stochastic process, a hidden (or latent) Markov chain.

In the social science framework, Vermunt, Langeheine, and Bockenholt (1999) extended the HMM to include individual covariates and Bartolucci, Pennoni, and Francis (2007) further developed it for multichannel observations. The basic model was generalized to the mixture hidden Markov model (MHMM) by van de Pol and Langeheine (1990) (who called it the mixed Markov latent class model) and further extended to include time-constant and time-varying covariates by Vermunt et al. (2008) (who named the resulting model as the mixture latent Markov model). In a mixture model, we assume that the data consists of latent subpopulations with differing model structures.

Hidden Markov modelling can be applied in various longitudinal settings; for accounting for measurement error and unobserved heterogeneity (e.g., Breen & Moisio, 2004; Pavlopoulos & Vermunt, 2015; Poulsen, 1990; van de Pol & Langeheine, 1990; Vermunt et al., 2008), for finding latent subpopulations (e.g., Bassi, 2014; McDonough, Worts, & Sacker, 2010; van de Pol & Langeheine, 1990, and Article IV), for detecting true unobservable states (e.g., various periods of the bipolar disorder in Lopez, 2008), and for compressing information across multichannel sequences (e.g., for finding more general life stages as in Article IV).

To the best of my knowledge, there are only few applications of the (M)HMM approach to multichannel social sequence data. Bartolucci et al. (2007) studied criminal trajectories using HMMs with multiple binary sequences per subject. Ip et al. (2015) analysed and classified binary profiles of food security for US farmworker households. Rijmen et al. (2008) studied 12 parallel trajectories of emotions among anorectic patients.

A few more studies propose extended or modified versions of the HMM or the MHMM. In their study of emotions of anorectic patients, Rijmen et al. (2008) extended their analysis to the *hierarchical HMM*. Zhang, Jones, Rijmen, and Ip (2010) studied children's measure-

ments of cognition and behaviour using trichotomized multichannel sequences. For that, they proposed the *extended multivariate discrete HMM*. Crayen, Eid, Lischetzke, Courvoisier, and Vermunt (2012) used a *hierarchical mixture latent Markov model* for two-channel categorical sequences to model dynamics of mood regulation of university students during one week. The hierarchical model had two parallel latent structures; one between the days and the other within the days. Ip, Zhang, Rejeski, Harris, and Kritchevsky (2013) proposed the *partially ordered mixed hidden Markov model* and applied it to binary disability sequences of older adults. Lu, Pan, Zhang, Dubé, and Ip (2015) propose the *reciprocal Markov model*, an extension of the HMM with intertwined hidden and observed Markov chains, to analyse trajectories of emotional and food intake statuses.

4.1 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.

For simplicity, let us start with an example of one hidden state sequence $\mathbf{z} = (z_1, z_2, \ldots, z_T)$ which generates one observed sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. A discrete first-order hidden Markov model is characterized by the following parameters:

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

$$\pi_s = P(z_1 = s); \quad s \in \{1, \ldots, S\}.$$

- *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:

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

- *Emission probability* matrix $B = \{b_s(m)\}$ of size $S \times M$, where $b_s(m)$ is the probability of the hidden state s emitting the observed state m:

$$b_s(m) = P(y_t = m | z_t = s); \quad s \in \{1, \ldots, S\}, m \in \{1, \ldots, M\}.$$

The observed state y_t at time t is independent of all other observations and hidden states given the current hidden state z_t. The first order Markov assumption states that the hidden state transition probability at time t only depends on the hidden state at the previous time point $t-1$:

$$P(z_t | z_{t-1}, \ldots, z_1) = P(z_t | z_{t-1}). \tag{4.1}$$

The transition matrix A is thus enough for defining the model for the transitions of a first-order Markov chain.

It is possible to extend the model to account for a longer history by using second- or higher-order Markov chains. In *homogeneous* HMMs, transition probabilities a_{sr} are constant over time. Modelling non-homogeneous HMMs is more complicated as such models need a

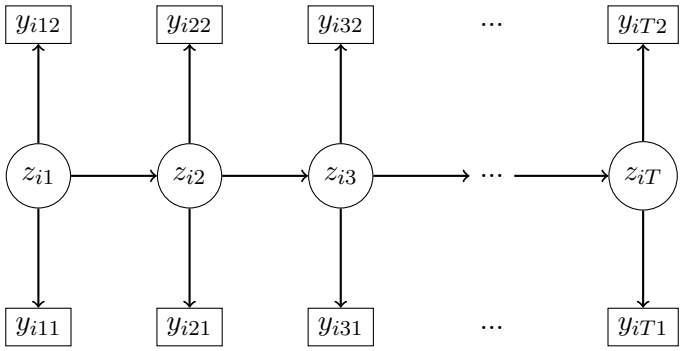


Figure 4.1: Illustration of hidden and observed state sequences in a hidden Markov model for two-channel data of individual i. The hidden state at time t is illustrated with z_{it} inside a circle and the observed state at time t in channel c with y_{itc} inside a rectangle. Arrows indicate dependencies between states.

set of transition matrices; see, e.g., Paliwal (1993) and Berchtold (1999) for extensions to non-homogeneous models.

Let us now extend to multichannel sequence data with N individuals, T timepoints, and C channels (naturally, the following applies for single-channel data, i.e., subjects with one sequence only, by setting $C = 1$). Now $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iT})$ represents the hidden state sequence for individual $i, i = 1, \ldots, N$ and y_{itc} denotes the observation of individual i at time $t, t = 1, \ldots, T$ in channel $c, c = 1, \ldots, C$. Here we assume the same latent structure applies for all channels, i.e., the hidden state z_{it} emits the observed states y_{itc} in all channels c. Observations $y_{it1}, \ldots, y_{itC}$ are assumed conditionally independent given the hidden state z_{it}, i.e., $P(\mathbf{y}_{it}|z_{it}) = P(y_{it1}|z_{it}) \cdots P(y_{itC}|z_{it})$ (see Figure 4.1 for an illustration for two-channel data). Naturally, the assumption of conditional independence of observations across channels is not unambiguously valid and must be evaluated in each case. In Article III we discuss the matter with more detail. If conditional independence cannot be assumed, data must be converted into single-channel representation – Section 1.2 already discussed issues relating to this.

For multichannel data, instead of only one emission probability matrix B we now have multiple matrices B_c, one for each channel $c, c = 1, \ldots, C$. The log-likelihood of the parameters $\mathcal{M} = \{\pi, A, B_1, \ldots, B_C\}$ of the HMM is written as

$$\log L = \sum_{i=1}^{N} \log P\left(Y_i|\mathcal{M}\right), \tag{4.2}$$

where Y_i are the observed sequences for subject i. The probability of the observation sequence

of subject i given the model parameters is

$$\begin{aligned} P(Y_i|\mathcal{M}) &= \sum_{\text{all } z} P\left(Y_i|z,\mathcal{M}\right) P\left(z|\mathcal{M}\right) \\ &= \sum_{\text{all } z} P(z_1|\mathcal{M}) P(\mathbf{y}_{i1}|z_1,\mathcal{M}) \prod_{t=2}^{T} P(z_t|z_{t-1},\mathcal{M}) P(\mathbf{y}_{it}|z_t,\mathcal{M}) \\ &= \sum_{\text{all } z} \pi_{z_1} b_{z_1}(y_{i11}) \cdots b_{z_1}(y_{i1C}) \prod_{t=2}^{T} \left[a_{z_{t-1}z_t} b_{z_t}(y_{it1}) \cdots b_{z_t}(y_{itC})\right], \end{aligned} \tag{4.3}$$

where the hidden state sequences $z = (z_1, \ldots, z_T)$ take all possible combinations of values in the hidden state space $\{1, \ldots, S\}$ and where $\mathbf{y}_{it}$ are the observations of subject i at t in channels $1, \ldots, C$; π_{z_1} is the initial probability of the hidden state at time $t = 1$ in sequence z; $a_{z_{t-1}z_t}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t; and $b_{z_t}(y_{itc})$ is the probability that the hidden state of subject i at time t emits the observed state at t in channel c.

4.2 Mixture hidden Markov model

The mixture hidden Markov model is, by definition, a mixture of simple hidden Markov models. Each cluster (or latent class) is characterized by the parameters of the respective submodel; transitions between submodels are not allowed.

Assume that we have a set of HMMs $\mathcal{M} = \{\mathcal{M}^1, \ldots, \mathcal{M}^K\}$, where $\mathcal{M}^k = \{\pi^k, A^k, B_1^k, \ldots, B_C^k\}$ for submodels $k = 1, \ldots, K$. For each subject Y_i, denote $P(\mathcal{M}^k) = w_k$ as the prior probability that the observation sequences of a subject follow the submodel $\mathcal{M}^k$.

The log-likelihood of the parameters of the MHMM is of the form

$$\begin{aligned} \log L &= \sum_{i=1}^{N} \log P(Y_i|\mathcal{M}) \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(\mathcal{M}^k) \sum_{\text{all } z} P\left(Y_i|z,\mathcal{M}^k\right) P\left(z|\mathcal{M}^k\right)\right] \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \sum_{\text{all } z} \pi_{z_1}^k b_{z_1}^k(y_{i11}) \cdots b_{z_1}^k(y_{i1C}) \prod_{t=2}^{T} \left[a_{z_{t-1}z_t}^k b_{z_t}^k(y_{it1}) \cdots b_{z_t}^k(y_{itC})\right]\right]. \end{aligned} \tag{4.4}$$

4.3 Covariates

Covariates can be added in the model to explain cluster memberships (in mixture models) or initial and transition probabilities. Prior probabilities may be modelled in the usual way with the multinomial distribution. For subject i with time-constant covariates $\mathbf{x}_i$, the prior cluster probabilities are now of the form

$$P(\mathcal{M}^k|\mathbf{x}_i) = w_{ik} = \frac{e^{\beta_k \mathbf{x}_i}}{1 + \sum_{l=2}^{K} e^{\beta_j \mathbf{x}_i}}, \tag{4.5}$$

where β_k is the vector of regression coefficients related to submodel k. The first submodel is set as the reference by fixing $\beta_1 = (0, \ldots, 0)'$.

4.4 Model estimation

The log-likelihoods of (4.2) and (4.4) are efficiently calculated with the *forward–backward algorithm* (Baum & Petrie, 1966; Rabiner, 1989). A common estimation method is the Baum–Welch algorithm, i.e., the expectation–maximization (EM) algorithm in the HMM context. Another option is to use direct numerical maximization.

Most of the estimation methods including the Baum–Welch algorithm require starting values for model parameters – the closer the starting values are to the optimum, the faster it is found. In order to reduce the risk of being trapped in a poor local optimum, a large number of initial values should be tested. Simpler models with few parameters are fast to estimate; therefore, it is possible to fit the model numerous times with varying random starting values for finding the model with the best likelihood. When the model is large, estimation is more time-consuming and good starting values for model parameters are useful or even essential. Articles III and IV discuss the matter of efficient model estimation in more detail.

The most probable path of hidden states for each subject given their observations and the model can be computed using the Viterbi algorithm (Rabiner, 1989; Viterbi, 1967), which maximizes the probability of $P(z|Y_i, \mathcal{M})$. Individual hidden state paths are useful in visualizing complex sequence data in a more parsimonious way.

Chapter 5

Graphical illustrations for sequence data

Visualization is a powerful tool throughout the analysis process from the first glimpses into the data to exploration and finally to presentation of the results. Tufte (1961) referred to graphical excellence as to what "gives the viewer the greatest number of ideas, in the shortest time, with the least ink, in the smallest space, and which tells the truth about data". This chapter discusses methods for visualizing sequence data and hidden Markov models. An emphasis is put on the multichannel case which has been considered less in the literature so far.

5.1 Visualizing sequence data

There are many options for graphical description of sequence data. Most of them either represent sequences or summarize them. *Sequence index plot* is the most commonly used example of the former (see an example applied to multichannel data in Figure 5.1). Such a graph was proposed by Scherer (2001) to show the observations of each subject in the order they appear, illustrating different states with different colours. The horizontal axis shows the time points while individuals are represented on the vertical axis; thus, each horizontal line shows the sequence of one individual.

When the number of subjects is moderate, sequence index plots give an accurate representation of the data, offering an overview on the timing of transitions and on the durations of different episodes. Sequence index plots become more complex to comprehend when the number of individuals and states increases. Sequence analysis with clustering eases interpretation by grouping similar histories together. Piccarreta and Lior (2010) suggested using multidimensional scaling for ordering sequences more meaningfully (similar sequences close to each other). Piccarreta (2012) proposed smoothing techniques that reduce individual noise. Similar sequences are summarized into artificial sequences that are representative to the data. Gabadinho, Ritschard, Müller, and Studer (2011) introduced *representative sequence plots* where only a relatively small number of the most representative sequences (observed or artificial) are shown. A similar approach, *relative frequency sequence plot*, was introduced by Fasang and Liao (2014). The idea is to find representative sequences (the medoids) in equal-sized neighbourhoods to represent the relative frequencies in the data.

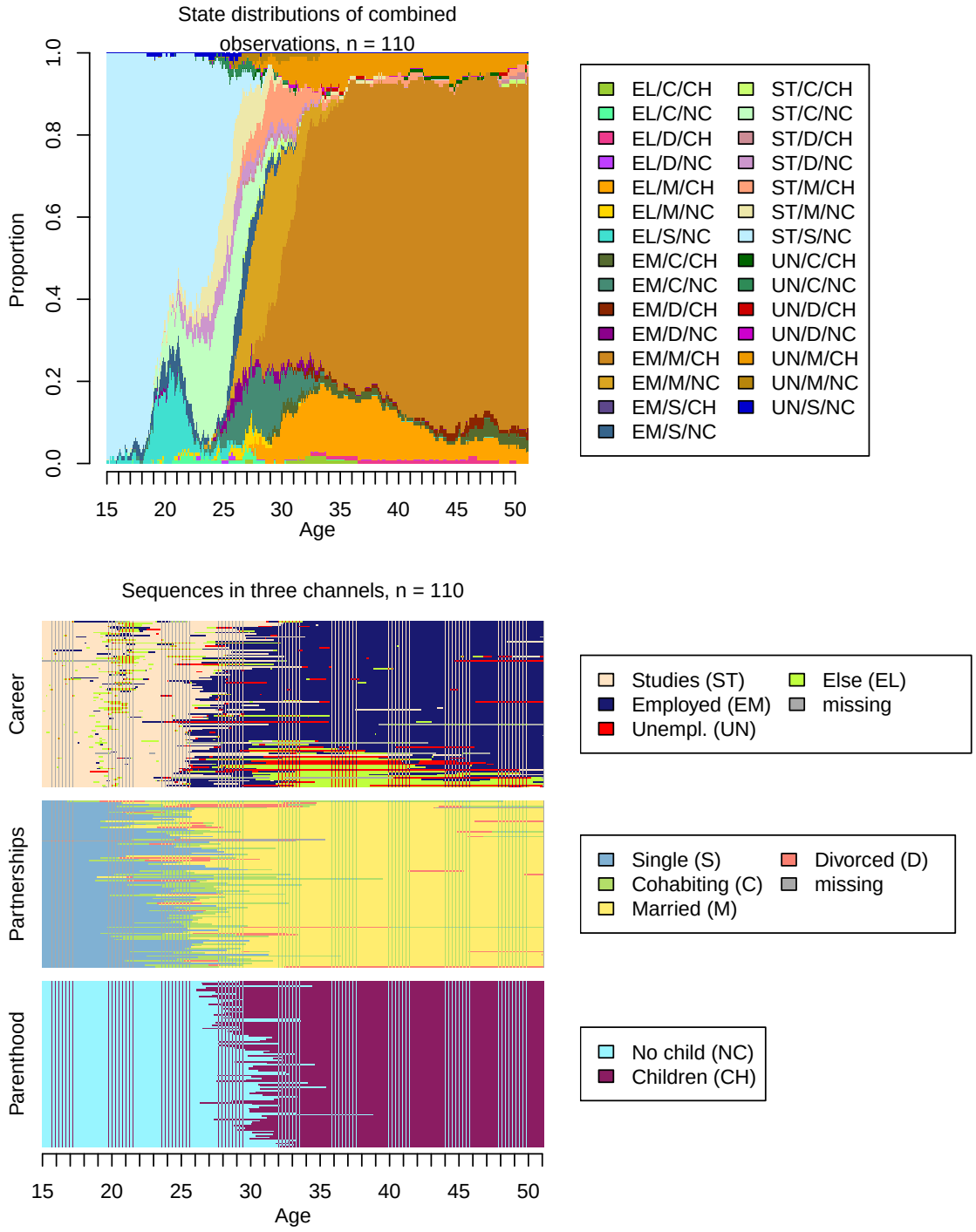


Figure 5.1: Visualizing three-channel life sequenced data. State distributions of combined observations (top) show the prevalence of (combined) life states at each time point. Sequence index plots (bottom) show the sequence(s) for each individual; here the observations in three life domains are plotted separately. Sequences are ordered by multidimensional scaling scores.

State distribution plots (also called tempograms or chronograms; Billari & Piccarreta, 2005; Widmer & Ritschard, 2009) summarize information in the whole data. Such graphs show the change in the prevalence of states in the course of time (see an example in Figure 5.1). Again, the horizontal axis represents time (here age) but vertical axis is now a percentage scale. These plots simplify the overall patterns but do not give information on transitions between different states. Other summary plots include, e.g., *transversal entropy plots* (Billari, 2001), which describe how evenly states are distributed at a given time point, and *mean time plots* (Gabadinho et al., 2011), which show the mean time spent in each state across the time points.

Visualizing multichannel data is not a straightforward task. Section 1.2 discussed the problems of dealing with multichannel sequences. Combining states into a single-channel representation often works well if the alphabet is small and states at each time point are either completely observed or completely missing. In other cases it can be preferable to preserve the multichannel structure. In Article III we propose the so-called *stacked sequence plot* where sequence data are plotted separately for each channel according to some criterion such as the scores from multidimensional scaling. The order of the subjects is kept the same in each plot and the plots are stacked on top of each other. Since the time axes are horizontally aligned, comparing timing in different life domains should be relatively easy. This approach also protects the privacy of the subjects; even though all data are shown, combining information across channels for a single individual is difficult unless the data are very small. State distribution plots can then be used to show information on the prevalence and timing of combined states on a more general level.

Figure 5.1 illustrates state distributions and stacked sequence index plots for three-channel life sequence data with monthly observations between ages 15–50. The data are a subset of the NEPS data used in Article IV, presenting individuals with long education and later family. At the start of the follow up, at age 15, almost all individuals are studying, single, and childless. Around the age of 20, many are out of workforce due to, e.g., military service or voluntary work. Many individuals form residential partnerships while studying. From the sequence index plots we can see that most cohabit with one or more partners before marrying. Some have children during their studies, while others first move to employment. After the age of 30, most individuals are married with children, typically employed. Some, however, stay out of employment for years: between the ages 30–40, 15–30% of the individuals in this subset are out of employment.

5.2 Visualizing hidden Markov models

Markovian models are often visualized as directed graphs where vertices (nodes) present states and edges (arrows, arcs) show transition probabilities between states. In Article III, we extend this basic graph in the hidden Markov model framework by presenting hidden states as pie charts, with emission probabilities as slices, and by adjusting the thickness of edges according to transition probabilities. Such graph allows for presenting a complex model in a very efficient way, guiding the viewer to the most important aspects of the model.

Figure 5.2 illustrates a HMM with five hidden states for the data visualized in Figure 5.1. Following the common convention, hidden states are presented as vertices and transition probabilities are shown as edges. Here, the width of the edge depends on the probability of the transition; the most probable transitions are thus easy to detect. Vertices are drawn

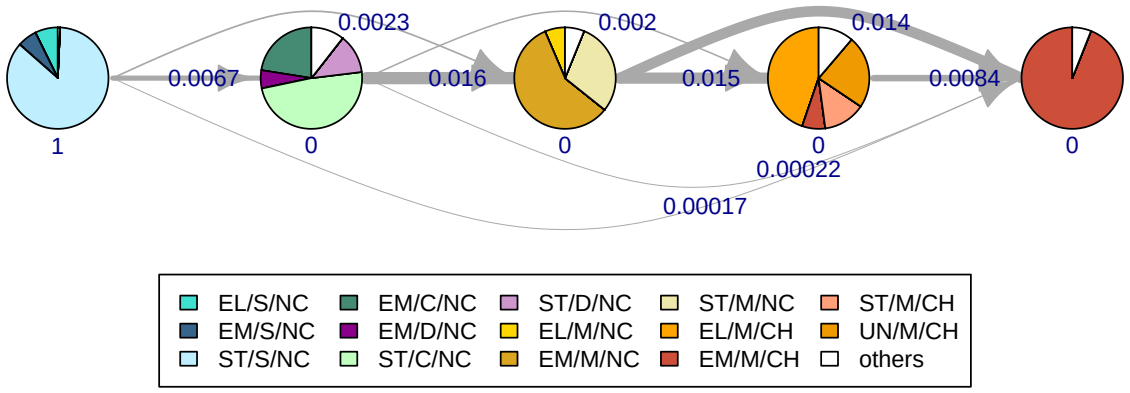


Figure 5.2: Illustrating a left-to-right hidden Markov model as a directed graph. Pies represent the hidden states, with emission probabilities as slices. Arrows illustrate transition probabilities between the hidden states. Probabilities of starting in each state are shown below the pies. The model is estimated for the data visualized in Figure 5.1. As indicated by the arrows, transitions back to preceding hidden states are not allowed here. The combined states show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren). The label "others" refers to combined states with joint emission probabilities less than 0.05.

as pie charts where the slices represent emitted observations or – in a multichannel case – combinations of observed states across channels. Also here, the size of the slice is proportional to the emission probability of the observed state (or in a multichannel model, the product of the emission probabilities across channels). For emphasizing the relevant information, observations with small emission probabilities can be combined into one category. Initial state probabilities are given below the respective vertices.

The graph shows the essence of the hidden states and the dynamics between them. Note that the transition probabilities are small as the data consists of monthly observations and typically individuals remain in one state for years. All individuals start from the first hidden state (initial probability is 1) where they are single and childless and mostly studying. The most likely transition from this hidden state (excluding transitions to the same state) is to the second hidden state, where individuals are cohabiting or separated, typically studying or working. The third hidden state represents childless marriage, mostly studying or working. Transitions out of this hidden state are almost as probable to the fifth state as they are to the fourth state. These both represent married parents; some move out of employment for some time, while others continue working. On very small probabilities (less than 0.00025), some transition to the last hidden state straight from the first or the second hidden state.

Chapter 6

Comparison of methods

This thesis compares three approaches suitable for the analysis of categorical life sequences: the model-free data mining method of SA and two model-based probabilistic approaches, EHA and (M)HMMs. Articles I and II compare SA and EHA in two different life course settings – multiple life domains and recurrent events – while Article IV discusses the usage of SA and MHMMs. Table 6.1 extends Table 1 in Article I into a summary of the basic differences of the three approaches discussed in this thesis.

Table 6.1: Basic differences of event history analysis (EHA), sequence analysis (SA), and mixture hidden Markov models (MHMM)

	EHA	SA	MHMM
Unit of analysis	Event, transition	Trajectory	Trajectory
Basic tool	Transition rate	Dissimilarity matrix	Initial, transition, and emission probabilities
Direction of inference	Prospective	Retrospective	Prospective
Mode of inference	Conditional	Unconditional	Conditional
Type of inference	Comparison of rates	(Dis)similarity of trajectories	Comparison of probabilities
Aim of inference	Individual level	Population level	Population & individual level

SA is a descriptive tool for finding patterns and creating an overall picture of whole trajectories. No assumptions about the data-generating mechanisms are needed. Joint analysis of multichannel data is straightforward and SA helps in developing an intuitive understanding of complex relationships. As shown in Articles I and II, SA is able to reveal typical and atypical patterns in life courses.

EHA is a predictive method which requires structured hypotheses and well-defined systems of hazard models. Analysing individual-level event histories is useful for drawing inferences about the effects of covariates on the occurrence and timing of events of interest. In EHA we can account for censoring and unobserved individual characteristics that affect the timing and duration of states.

HMMs and MHMMs can be used for data of most versatile types: time series and panel data, one or multiple subjects, independent sequences or multichannel data, continuous or

categorical observations. The main difference to EHA is the inclusion of the latent level, i.e., one or more unobservable statuses that may be constant (cluster memberships) or vary in time (hidden states). MHMMs are able to reveal patterns in the population as well as on individual level; clustering identifies groups of similar life trajectories, hidden states and transition probabilities describe dynamics within groups, and most probable hidden state paths give information on individual trajectories. Some of the main issues are related to the Markov assumption; first-order models are (relatively) simple, but assuming that only the previous state has an effect to the present may be problematic. Higher-order models can account for a longer history but are much more complicated. When applying HMMs to life course data, also the homogeneity assumption needs to be contemplated; is it reasonable that the transition probabilities remain the same throughout the follow-up?

EHA is valuable when analysing a few well-specified events. When the number of states and transitions between the states increases, joint analysis may become too complex. Hidden Markov modelling and especially SA are easily applied for multichannel sequence data with multiple states and various patterns of transitions between them.

Analysis of sequence data using model-based methods often suffers from long estimation times – the larger the data and the more complex the model, the longer the time required for estimation. Often, estimating a set of candidate models is required for finding the best model structure, increasing the estimation time even further. Model-free SA is typically relatively fast to apply.

External information can be used to explain differences in life courses in each of the approaches. When applying SA, currently only time-constant covariates can be used. Time-varying variables can only be included as additional parallel state sequences (channels). In SA, the focus is on holistic patterns and as covariates are typically applied on the cluster level and the choice of the clustering result is at least to some extent subjective, one should regard the associations as suggestive and be cautious when drawing inferences on individual level. For this purpose, analyzing individual-level event histories is a more appropriate approach.

At present, time-varying covariates are only possible in model-based analysis. Covariates may have an effect on changes in the (observed or unobserved) status or the timing of an event. In the MHMM framework, covariates can also be used to predict or explain whole patterns of life courses (through cluster membership probabilities; see Article III).

The choice of the method(s) depends on the type of the data and the aims of the study. Applying different approaches provides versatile information on the phenomena of interest, as the methods capture time in different ways. Descriptive SA can also be used as a starting point for modelling. In Article IV, SA was used to sort complex individual life courses into clusters. Hidden Markov models were then used to compress information of multichannel sequences into more general life stages and to describe the dynamics between them. In another recent paper, Rossignon, Studer, Gauthier, and Le Goff (2016) used SA for identifying (time-varying) typologies of childhood co-residence trajectories which were further included as covariates in EHA to explain the probability of leaving parental home. As these examples reveal, model-free and modelling approaches are not substitutes but complete each other in the analysis of life sequence data.

Summary of original publications

Article I compares sequence analysis and event history analysis in the analysis of life history calendar data. EHA is used to estimate cumulative prediction probabilities of multiple life events. Regarding SA, different dissimilarity metrics are explored and compared. As an example, we study transitions to adulthood in three life domains in a Finnish cohort born in 1959 and assess the relationship between life trajectories and excess depressive symptoms in midlife.

We find that the two approaches complement each other. Model-free SA is useful in obtaining an overview of multichannel sequence data with multiple states and transitions, offering means for large-scale comparative analysis across populations or cohorts. EHA requires structured hypotheses; in complex settings analysis may be challenging. Time-varying covariates and conditioning on individual-level history are only possible in EHA.

Article II presents SA and EHA approaches in the analysis of recurrent events and shows how these methods can complement each other in an empirical analysis of co-residential partnership histories. As a substantive question, we study how family background and childhood socio-emotional characteristics are related to later partnership formation and stability in a Finnish cohort born in 1959.

With SA we find eight partnership clusters that differ in the number and timing of partnerships. With discrete-time EHA we are able to capture a notable part of variation due to time-invariant individual characteristics that in previous studies have been left to the unobserved random part. We find that especially high self-control of emotions during childhood is associated with the probability of partnership transitions in adulthood, e.g., a lower risk of dissolution. High social activity in childhood is related to men's tendency to form first and subsequent partnerships faster.

Article III introduces the R package seqHMM for the analysis of hidden Markov models (HMMs) and mixture hidden Markov models (MHMMs) for categorical sequence data. We formulate the HMM for multichannel sequences and extend it to the mixture model with or without external covariates. In seqHMM, we provide functions for estimation and inference of the HMM and the MHMM, as well as some special cases such as latent class models and Markov models. The package provides several alternatives for efficient and flexible model estimation and supports fast parallel computation. We also introduce new approaches and easy-to-use tools for visualizing multichannel sequence data and hidden Markov models.

Article IV illustrates the comparative nature of SA and (M)HMMs in an empirical analysis of large life sequence data from a German cohort born in 1955–1959. We study two differ-

ent approaches for analysing clustered life sequence data with sequence analysis and hidden Markov models. In the first approach we use SA clusters as fixed and estimated HMMs separately for each group. States found with SA are used as suggestions for hidden states in the HMM. In the second approach we treat SA clusters as suggestive and used them as a starting point for the estimation of MHMMs. Analyses are conducted with the seqHMM package.

Model estimation with complex sequence data turns out to be challenging due to computational issues. Here, the second approach with undecided numbers and contents of hidden states and clusters turns our to be unfeasible. With the first approach we end up in eight clusters which describe life trajectories that differ by the timing and occurrence of career and family states. We find that in the HMM framework, information in life sequence data can be compressed into hidden states describing different life stages and clusters representing general patterns in life courses. Hidden states are able to capture general life stages that include not only rather stable episodes but also life stages characterized by relatively rapid change.

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. doi: 10.1007/s10680-007-9134-6

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods*, *16*(4), 129–147. doi: 10.1080/01615440 .1983.10594107

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, *16*(3), 471–494.

Aisenbrey, S., & Fasang, A. (2010). New life for old ideas: The "second wave" of sequence analysis – bringing the "course" back into the life course. *Sociological Methods & Research*, *38*(3), 420–462. doi: 10.1177/0049124109357532

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological methodology*, *13*(1), 61–98. doi: 10.2307/270718

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781412984195

Allison, P. D. (2009). *Fixed effects regression models*. Los Angeled, CA: SAGE publications. doi: 10.4135/9781412993869

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York, NY: Springer-Verlag.

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*(2), 91.

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. doi: 10.1111/j.1467-985X.2006.00440.x

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, Germany: Springer-Verlag. doi: 10.1007/ 10104_2014_20

Batagelj, V. (1988). Generalized ward and related clustering problems. In H. H. Bock (Ed.), *Classification and related methods of data analysis* (pp. 67–74). Amsterdam, Netherlands: North-Holland.

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. doi: 10.1214/aoms/1177699147

Belli, R., Stafford, F., & Alwin, D. (2008). *Calendar and time diary: methods in life course research*. Sage Publications, Inc.

Berchtold, A. (1999). The double chain Markov model. *Communications in Statistics-Theory and Methods*, *28*(11), 2569–2589. doi: 10.1080/03610929908832439

Berchtold, A., & Raftery, A. E. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 328–356. doi: 10.1214/ss/1042727943

Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociological Methodology*, *41*(1), 195–221. doi: 10.1111/j.1467-9531.2011.01235.x

Billari, F. C. (2001). The analysis of early life courses: Complex descriptions of the transition to adulthood. *Journal of Population Research*, *18*(2), 119–142.

Billari, F. C., Fürnkranz, J., & Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population/Revue Européenne de Démographie*, *22*(1), 37–65. doi: 10.1007/s10680-005-5549-0

Billari, F. C., & Piccarreta, R. (2005). Analyzing demographic life courses through sequence analysis. *Mathematical Population Studies*, *12*(2), 81–106. doi: 10.1080/08898480590932287

Blossfeld, H.-P., & Rohwer, G. (2001). *Techniques of event history modeling: New approaches to causal analysis*. New York, NY: Psychology Press.

Bonetti, M., Piccarreta, R., & Salford, G. (2013). Parametric and nonparametric analysis of life courses: An application to family formation patterns. *Demography*, *50*(3), 881–902. doi: 10.1007/s13524-012-0191-z

Breen, R., & Moisio, P. (2004). Poverty dynamics corrected for measurement error. *The Journal of Economic Inequality*, *2*(3), 171–191. doi: 10.1007/s10888-004-3227-9

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks.

Brown, C. C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics*, *31*(4), 863–872. doi: 10.2307/2529811

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. New York, NY: Cambridge University Press.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic Medicine*, *74*(4), 366–376. doi: 10.1097/PSY.0b013e31825474cb

Deville, J.-C., & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. *Journal of Econometrics*, *22*(1), 169–189. doi: 10.1016/0304-4076(83)90098-2

Dijkstra, W., & Taris, T. (1995). Measuring the agreement between sequences. *Sociological methods & research*, *24*(2), 214. doi: 10.1177/0049124195024002004

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

Elzinga, C. H. (2003). Sequence similarity: a nonaligning technique. *Sociological Methods and Research*, *32*(1), 3–29. doi: 10.1177/0049124103253373

Elzinga, C. H., & Studer, M. (2015). Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, *44*(1), 3–47. doi: 10.1177/0049124114540707

Fasang, A. E., & Liao, T. F. (2014). Visualizing sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods & Research*, *43*(4), 643-676. doi: 10

.1177/0049124113506563

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. doi: 10.18637/jss.v040.i04

Gauthier, J.-A., Bühlmann, F., & Blanchard, P. (2014). Introduction: Sequence analysis in 2014. In *Advances in sequence analysis: Theory, method, applications* (pp. 1–17). New York, NY: Springer-Verlag.

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. doi: 10.1111/j.1467-9531.2010.01227.x

George, L. K. (1993). Sociological perspectives on life transitions. *Annual Review of Sociology*, 353–373. doi: 10.1146/annurev.so.19.080193.002033

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, *56*(296), 841–868. doi: 10.1080/01621459.1961.10482130

Grelet, Y. (2002). Des typologies de parcours: méthodes et usages. *Notes de Travail Génération 92*, *20*.

Halpin, B. (2010). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research*, *38*(3), 365–388. doi: 10.1177/0049124110363590

Halpin, B. (2014). Three narratives of sequence analysis. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 75–103). New York, NY: Springer-Verlag.

Halpin, B., & Chan, T. W. (1998). Class careers as sequences: An optimal matching analysis of work-life histories. *European Sociological Review*, *14*(2), 111–130. doi: 10.1093/oxfordjournals.esr.a018230

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, *29*(2), 147–160. doi: 10.1002/j.1538-7305.1950.tb00463.x

Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods & Research*, *38*(2), 235–264. doi: 10.1177/0049124109346164

Hougaard, P. (2012). *Analysis of multivariate survival data.* New York, NY: Springer-Verlag.

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. doi: 10.2105/AJPH.2015.302752

Ip, E. H., Zhang, Q., Rejeski, J., Harris, T., & Kritchevsky, S. (2013). Partially ordered mixed hidden Markov model for the disablement process of older adults. *Journal of the American Statistical Association*, *108*(502), 370–384. doi: 10.1080/01621459.2013.770307

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). Hoboken, NJ: John Wiley & Sons.

Kohonen, T. (2001). *Self-organizing maps* (Vol. 30). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-642-56927-2

Kruskal, J. B., & Liberman, M. (1983). The symmetric time-warping problem: from continuous to discrete. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison* (pp. 125–161). Reading, MA: Addison-Wesley.

Lee, E. T., & Wang, J. W. (2003). *Statistical methods for survival data analysis.* Hoboken,

NJ: John Wiley & Sons.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, *38*(3), 389–419. doi: 10.1177/0049124110362526

Lesnard, L., & Kan, M. Y. (2011). Investigating scheduling of work: a two-stage optimal matching analysis of workdays and workweeks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 349–368. doi: 10.1111/j.1467-985X.2010.00670.x

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Levine, J. (2000). But what have you done for us lately? *Sociological methods & research*, *29*(1), 34–40. doi: 10.1177/0049124100029001002

Levy, R. (2005). Why look at life courses in an interdisciplinary perspective? *Advances in Life Course Research*, *10*, 3–32. doi: 10.1016/S1040-2608(05)10014-8

Lopez, A. (2008). *Markov models for longitudinal course of youth bipolar disorder* (Doctoral dissertation, University of Pittsburgh, Ann Arbor, MI). Retrieved from `http://d-scholarship.pitt.edu/6524/1/LopezAdrianaApril23.pdf`

Lu, J., Pan, J., Zhang, Q., Dubé, L., & Ip, E. H. (2015). Reciprocal Markov modeling of feedback mechanisms between emotion and dietary choice using experience-sampling data. *Multivariate Behavioral Research*, *50*(6), 584–599. doi: 10.1080/00273171.2015.1033510

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

Marteau, P.-F. (2008). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 306–318. doi: 10.1109/TPAMI.2008.76

Massoni, S., Olteanu, M., & Rousset, P. (2009). Career-path analysis using optimal matching and self-organizing maps. In J. C. Príncipe & R. Miikkulainen (Eds.), *Advances in self-organizing maps* (pp. 154–162). Berlin, Germany: Springer-Verlag.

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. doi: 10.1016/j.socscimed.2009.10.001

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(2), 317–334. doi: 10.1111/1467-985X.00641

Mills, M. (2011). *Introducing survival and event history analysis*. London, UK: Sage Publications. doi: 10.4135/9781446268360

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods*, *4*(2), 139. doi: 10.1037/1082-989X.4.2.139

Paliwal, K. (1993). Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer. In *Proceedings of the 1993 ieee international conference on acoustics, speech, and signal processing* (Vol. 2, pp. 215–218). doi: 10.1109/ICASSP.1993.319273

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.

Piccarreta, R. (2012). Graphical and smoothing techniques for sequence analysis. *Sociological Methods & Research*, *41*(2), 362–380. doi: 10.1177/0049124112452394

Piccarreta, R. (2015). Joint sequence analysis association and clustering. *Sociological Methods & Research*, 0049124115591013. doi: 10.1177/0049124115591013

Piccarreta, R., & Lior, O. (2010). Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(1), 165–184. doi: 10.1111/j.1467-985X.2009.00606.x

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. doi: 10.1016/0167-8116(90)90028-L

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi: 10.1109/5.18626

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 528–539. Retrieved from `http://www.jstor.org/stable/2345788`

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. doi: 10.1007/s11336-007-9001-8

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *116*(1), 5–24. doi: 10.1177/0759106312454635

Robette, N., Bry, X., & Lelièvre, É. (2015). A global interdependence approach to multidimensional sequence analysis. *Sociological Methodology*, 0081175015570976. doi: 10.1177/0081175015570976

Rossignon, F., Studer, M., Gauthier, J.-A., & Le Goff, J.-M. (2016). Childhood co-residence structures and homeleaving: A combination of survival and sequence analyses. In R. Gilbert & S. Matthias (Eds.), *Proceedings of the international conference on sequence analysis and related methods, lausanne, june 8-10, 2016* (pp. 383–427).

Rousset, P., Giret, J.-F., & Grelet, Y. (2012). Typologies de parcours et dynamique longitudinale. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *114*(1), 5–34. doi: 10.1177/0759106312437142

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review*, *17*(2), 119–144. doi: 10.1093/esr/17.2.119

Settersten, R. A., Jr., & Mayer, K. U. (1997). The measurement of age, age structuring, and the life course. *Annual Review of Sociology*, *23*, 233–261. doi: 10.1146/annurev.soc.23.1.233

Steele, F. (2011). Multilevel discrete-time event history analysis with applications to the analysis of recurrent employment transitions. *Australian & New Zealand Journal of Statistics*, *53*(1), 1–20. doi: 10.1111/j.1467-842X.2011.00604.x

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers*, *2013*.

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. doi: 10.1111/rssa.12125

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, *40*(3), 471–510. doi: 10.1177/

0049124111415372

Tufte, E. R. (1961). *The visual display of quantitative information.* Chesire, CT: Graphics Press.

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. doi: 10.1177/0049124186015001009

van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, *20*, 213–247. Retrieved from `http://www.jstor.org/stable/271087` doi: 10.2307/271087

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454.

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. doi: 10.3102/10769986024002179

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, MA: Elsevier.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, *13*(2), 260–269. doi: 10.1109/TIT.1967.1054010

Ward, J., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244. doi: 10.1080/01621459.1963.10500845

Widmer, E. D., & Ritschard, G. (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research*, *14*(1), 28–39. doi: 10.1016/j.alcr.2009.04.001

Wiggins, L. M. (1955). *Mathematical models for the interpretation of attitude and behavior change: the analysis of multi-wave panel* (Unpublished doctoral dissertation). Columbia University, New York, NY.

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

Wu, L. L. (2000). Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological Methods Research*, *29*(1), 41-64. doi: 10.1177/0049124100029001003

Zhang, Q., Jones, A. S., Rijmen, F., & Ip, E. H. (2010). Multivariate discrete hidden Markov models for domain-based measurements and assessment of risk factors in child development. *Journal of Computational and Graphical Statistics*, *19*(3), 746–765. doi: 10.1198/jcgs.2010.09015

I

Eerola, M. and Helske, S. (2016) Statistical analysis of life history calendar data. *Statistical Methods in Medical Research.* 25(2), 571–597. doi:10.1177/0962280212461205

Statistical analysis of life history calendar data

Mervi Eerola* and Satu Helske†

Abstract

The life history calendar is a data-collection tool for obtaining reliable retrospective data about life events. To illustrate the analysis of such data, we compare the model-based probabilistic event history analysis and the model-free data mining method, sequence analysis. In event history analysis, we estimate instead of transition hazards the cumulative prediction probabilities of life events in the entire trajectory. In sequence analysis, we compare several dissimilarity metrics and contrast data-driven and user-defined substitution costs. As an example, we study young adults' transition to adulthood as a sequence of events in three life domains. The events define the multistate event history model and the parallel life domains in multidimensional sequence analysis. The relationship between life trajectories and excess depressive symptoms in middle-age is further studied by their joint prediction in the multistate model and by regressing the symptom scores on individual-specific cluster indices. The two approaches complement each other in life course analysis; sequence analysis can effectively find typical and atypical life patterns while event history analysis is needed for causal inquiries.

Keywords: Distance-based data; Life course analysis, Life history calendar; Multidimensional sequence analysis; Multistate model; Prediction probability

1 Introduction

Follow-up studies, which register prospective events in time, are the golden standard of reliable data collection in developmental studies and life course analysis. Yet these can be expensive and sometimes difficult to perform. Retrospective data collection is used mainly when a very large sample is required, the classical example being rare outcomes and case-control designs. Recently, however, retrospective data collection has been used in survey studies to obtain detailed information about multiple life domains and individuals' multiple activities.[1] *The life history calendar* (LHC), also called an event-history calendar, is a data-collection tool for obtaining reliable retrospective data about life events.[2] The advantage of a life history calendar is that the order and proximity of important transitions in multiple life domains can be studied at the same time. The time window of a life history calendar can be years or even an entire life-span. As a data collection tool, it encourages respondents to incorporate temporal changes as cues in the reporting of events. It has shown the ability to provide data of remarkably high quality.[1]

While life course epidemiology studies the relationship between exposure and disease, problems of special interest to psychologists and social scientists point to an understanding of individuals' behaviour and choices in their lives. These choices are often reflected in the amount of time devoted to different activities. Individuals also have several social roles in their lives, and in these roles they share values and resources which may form their decisions and experiences in a similar way. These

*Department of Mathematics and Statistics, Assistentinkatu 7, 20014 University of Turku, Finland; tel: +358-2-3335437, +358-40-5622913; email: mervi.eerola@utu.fi

†Methodology Centre for Human Sciences/Department of Mathematics and Statistics, University of Jyväskylä

links have been of interest especially in life course studies carried out by social scientists. Linking different life domains (e.g. education, family formation, health, working life) of a single individual is an effort to study the life course as an interdependent system of life processes, and makes the analysis multidimensional and dynamic at the same time. This is the focus of our article when evaluating methods for the statistical analysis of life history calendar data. We believe that the approach taken by sociologists and psychologists can be valuable also to health scientists. Variable life patterns can have effects, for example, on chronic diseases or on patients' differential response to clinical treatments.

Traditionally, life course data have been analysed by event history methods. There is a vast literature on the basic principles and on more advanced methods based on the theory of counting processes (e.g. Andersen et al.[3]). These methods are valuable when studying the time course of a few well-specified life events but when the number of states, and accordingly the number of transitions between the states, increases, joint analysis of the model especially for prediction purposes becomes rather elaborate. In this article, we compare two approaches to life course analysis: model-based probabilistic event-history analysis (EHA) and a more recent type of approach of model-free data-mining, sequence analysis (SA). The latter is well known in bioinformatics but has provided novel insight to the diversity of life trajectories and their relationship to life satisfaction and depressiveness. We emphasize the differences, but also the complementary tasks of the methods. As an example, we study young adults' pathways to adulthood and consequent depressive symptoms in middle age in a cohort established in Central Finland in 1968. The cohort members have been followed for 42 years, from age 8 until age 50.

The article is structured as follows. In Section 2, some concepts and principles of prospective and retrospective approaches to life course analysis are contrasted, the first in terms of predictive probabilities and the latter in terms of typologies of life sequences. Section 3 provides comparative analysis of the cohort data and some sensitivity analysis. Finally, Section 4 presents a methodological discussion about the different informational content of the two approaches.

2 Prospective and retrospective analysis of the life course

From a methodological point of view, the timing and order of events is of fundamental relevance in life course analysis. Events represent transitions, marking developmental stages in life, while the role and statuses accompanying such transitions feature the essential characteristics of the life course.[4] *Trajectories* are sequences of previously occupied life states, which provide a long-term view of usually one dimension of an individual's life course. *Transitions* between the states, which are of course embedded in the life trajectories, provide a short-term view of the dynamics of the life course. Historically, transitions have been more important concepts because they relate directly to important changes in life history.

Recently, more attention has been given to micro-settings and diversity of the dynamics involved in the individual's different activities, roles, and relationships. This change in scope has emphasized the analysis of whole trajectories instead of events. The role of transitions and trajectories as the basic unit of analysis is described in the next sections.

2.1 Event history analysis

Prospective analysis is based on short-term predictions of transitions in the life course. These predictions can be modified by some informative covariates $\mathbf{Z}$ which themselves may vary in time. A concise review of event history methods can be found, for example, in Andersen and Keiding.[5] Here we prefer, however, an approach based on a *marked point process* $(T, X) = \{(T_n, X_n), n \geq 1\}$. Rather than a system of states accompanied with a transition matrix, we model the life course as a sequence

of events by specifying a pair of random variables, the occurrence time T and a mark X identifying the event. An extensive overview of such models and theory is given by Arjas.[6]

Let $N_x(t) = \sum_{n\geq 1} 1\{T_n \leq t, X_n = x\}$ be a process counting x-specific events in an individual's life course such that $\sum_x N_x(t) = N(t)$ is the total number of life events by time t. Since life history calendar data is often recorded on a yearly basis, we define the discrete *event-specific hazard* in the age interval $t = 1, 2, ...$ as the conditional probability of a change in the value of N_x

$$p_x(t) = P(\Delta N_x(t) = 1 | \mathcal{F}_{t-1}^N) \tag{1}$$

given the internal history $\mathcal{F}_{t-1}^N$ of the counting process. We will denote the history of the occurrence times and marks by time t as H_t. The crude hazard that some event occurs in the interval t, regardless of which one, is the sum over the event-specific hazards, $p(t) = \sum_x p_x(t)$.

The likelihood contribution of an individual's life history can be interpreted as a product of a sequence of multinomial trials over the intervals. Since $\Delta N_x(t)$ can only have the value of 1 or 0 in a short interval t, the outcome of the multinomial trial within each interval can be read from its value. This determines which one of the x-components of N contributes to the likelihood. For a generic individual, the likelihood contribution by time t is

$$L(t) = \prod_{s\leq t} \prod_x p_x(s)^{\Delta N_x(s)} (1 - p(s))^{1-\Delta N(s)}. \tag{2}$$

While the hazard gives a very short-term prediction of the life course, the *prediction process* associated with a marked point process gives a long-term prediction of some random event related to (T, X) for the whole observed trajectory.[7–10] We can then view the prediction process as the conditional distribution of that random event given the history H_t. The prediction probabilities are again functions of event-specific hazards, so modelling the hazards brings external explanatory information to the prospective analysis of the whole life trajectory.

In Section 3.2, we shall consider in detail the specification and estimation of prediction probabilities in a multistate model. For a tutorial on event history analysis and prediction probabilities, we refer to Putter et al.[10] In the next section, we contrast the model-based predictions of life events, extended to the whole observed trajectory, with the model-free approach of sequence analysis. Since it is still less familiar than event history analysis to health scientists, we give a more extensive overview of its basic principles.

2.2 Life sequence analysis

A completely different approach is taken in sequence analysis (SA), originally used in bioinformatics to organize, classify, and parse protein and DNA sequence data.[11] In the 1980s, data mining methods were developed to analyse molecular sequences as texts (e.g. TGACT = Thymine-Guanine-Adenine-Cytosine-Thymine). Comparing sequences corresponds to comparing amino acids in protein sequences or nucleotides in DNA sequences at each position. The goal is to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. This is accomplished by aligning the sequences pairwise. Gaps are inserted between the elements so that identical or similar characters are aligned in successive columns. Mismatches between the sequences can have biological interpretations as point mutations and gaps as insertion or deletion mutations introduced in one or both lineages since their divergence from a common ancestor.

In the life course setting, sequence analysis was first introduced by the social scientist A. Abbott.[12] He criticized the event-oriented method as being unable to reveal life patterns when focusing only on isolated events. Aligning life sequences provided correspondences with similar life patterns, while mismatches and gaps corresponded to differential timing and/or a lack of certain life events or episodes. Studying trajectories as the basic units allowed them to be interpreted as connected series of experiences or summaries of lives, not isolated events.[13]

Table 1: Basic differences of sequence analysis and event history analysis.

Method	Sequence analysis	Event-history analysis
Unit of analysis	sequence	event
Basic tool	distance matrix	transition rate
Direction of inference	retrospective	prospective
Mode of inference	static, unconditional	dynamic, conditional
Type of inference	alignment of sequences	comparison of rates
Aim of inference	population-level	individual-level

While event-history analysis models the risk of life events with explanatory covariates, sequence analysis aims at forming typologies of life trajectories based on their similarity and characterizing them by means of covariates. To assess similarity, pairwise distances of the sequences are first calculated. The distance matrix is then used as data for clustering to find similar life patterns. Table 1 summarizes the basic differences of the two methods. We notice that, from a statistical point of view, they have in many respects completely different approaches. One can expect that they also provide different types of information about the life course.

2.2.1 Probabilistic sequence analysis

We start with reviewing a probabilistic approach to SA to more clearly contrast the prospective and retrospective probabilistic life course analyses, and then focus on the non-probabilistic sequence analysis that has been used exclusively in life sequence analysis to date. We follow closely Durbin et al.[11] in the probabilistic SA presentation.

Sequence alignment depends on a scoring model, on the algorithm for optimizing the scoring, and on statistical methods to evaluate the goodness of the results. In a probabilistic scoring model, the *substitution score* measures the relatedness of sequences in the observed data with the expected case, where matching occurs only randomly at each position. The log odds ratio of the scoring models for the whole sequences compares the log of observed and "expected by chance" models.

Consider two sequences, x and y with lengths m_x and m_y. Let x_i be the symbol of ith site of x and y_j be the symbol of the jth site of y. In the case of DNA sequences, the symbols are elements of $\{A, T, C, G\}$ so there are $K = 4$ symbols. We want to assign a score to the alignment that measures the relative likelihood that the sequences are related as opposed to being unrelated. The unrelated scoring model assumes that a symbol, say a, occurs independently with frequency q_a. For sequences of equal length, the unrelated or random model is then of the form

$$P(x, y|R) = \prod_i q_{x_i} \prod_j q_{y_j} \tag{3}$$

whereas the related or match model M is the product of joint probabilities for the whole alignment

$$P(x, y|M) = \prod_i p_{x_i y_i}. \tag{4}$$

Here p_{ab} is the joint probability of elements a and b occurring as an aligned pair. The ratio of the models is the odds ratio

$$\frac{P(x, y|M)}{P(x, y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}. \tag{5}$$

To have an additive score model, we take a logarithm of the odds ratio which can further be written as

$$log\frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}} = \sum_i log\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} = \sum_i s(x_i, y_i). \tag{6}$$

The substitution costs $s(a, b)$ for each aligned pair of elements can be arranged in a $K \times K$ matrix which gives a statement about the probability of ab occurring jointly. The probabilities p and q are based on biological theory.

An optimal *alignment algorithm* to minimize the total cost of dissimilarity (or maximize the score of similarity) is based on dynamic programming. Optimal matching (OM) computes generalized Levenshtein distances[14] by minimizing the cost of elementary operations: substitution, insertion, or deletion of an element. Insertions or deletions are jointly called *indels*. The cost of a gap (one or more conjoined indels) is often set as the length of the gap, but opening and extending gaps can also be given different weights. OM quantifies the effort needed to transform one sequence to another. Example 1 illustrates a possible alignment of two sequences and the OM operations needed to compute the cost.

Example 1

sequence 1: AAAABBBB
sequence 2: AAA−BBCC

The alignment above contains five matching elements, two mismatches, and a gap of length 1. For defining the cost of this alignment, sequence 2 is transformed to sequence 1 using an insertion of an element A and two substitutions of an element C with B (shown bold).

AAAABBBB → AAAABBBB → AAAABBBB
AAABBCC → AAA**A**BBCC → AAAABB**BB**

The cost of the alignment is the sum of the costs of the operations. Transforming sequence 1 would lead to exactly the same result. The best possible alignment with the lowest cost is found using dynamic programming.

A global alignment algorithm is, for example, the Needleman–Wunsch algorithm.[15] The idea is to build up an optimal alignment, using previous solutions for optimal alignments of smaller subsequences. To find the alignment with the lowest score, a matrix D is allocated. The value $D(i, j)$ is the score of the best alignment between the initial segments $x_{1...i}$ and $y_{1...j}$ and can be built recursively. First $D(0, 0) = 0$ is initialized. The matrix is then filled from top left to bottom right with

$$D(i, j) = \min \begin{cases} D(i-1, j-1) + s(x_i, y_j) \\ D(i-1, j) - o \\ D(i, j-1) - o, \end{cases} \tag{7}$$

where $s(x_i, y_j)$ is the cost of a substitution and o of an indel. In the first row x_i is aligned with y_j; in the second row x_i is aligned with a gap in y; and in the third row y_j is aligned with a gap in x. The best score up to (i, j) will be the smallest of these. The equation is applied repeatedly until the matrix is filled. The value in the final cell $D(m_x, m_y)$ is the best score for an alignment of x with y.

The significance of a particular alignment M can be assessed, for example, by Bayesian model comparison. The posterior of alignment M is

$$P(M|x, y) = \frac{P(x, y|M)P(M)}{P(x, y)}, \tag{8}$$

where $P(M)$ is the prior probability of M and $P(x, y|M)$ the likelihood of data given alignment M. The Bayes factor of the odds ratio is then

$$log\left(\frac{P(x,y|M)}{P(x,y|R)}\right) + log\left(\frac{P(M)}{P(R)}\right) \tag{9}$$

where R is the random model.

2.2.2 Non-probabilistic sequence analysis

In life sequence applications, only non-probabilistic sequence analysis has been used to date. These methods are either based on sequence editing and pairwise alignment of sequences as in the probabilistic case, or on counting common sequence attributes (non-alignment methods).

Substitution costs. The most important difference is that in pairwise alignment the substitution cost $s(x_i, y_j)$ is not a log odds ratio as in probabilistic SA but rather a given constant, defined by the analyst. At least three alternatives have been used. The first derives costs from *substantive theory* that often suggests some order between the states. Different substantive questions have of course different interpretations for the similarity of states. In social science applications, a theory-based cost matrix is often preferred because the timing of events and the similarity of states are considered conceptually separate issues (e.g. Halpin[16]).

Subjectivity in cost definition can be reduced by *data-driven costs* which are inversely proportional to transition frequencies from state A to B and B to A.[17,18] The time-independent cost of substituting A to B is then

$$2 - p(A, B) - p(B, A)$$

where $p(A, B)$ is the estimated proportion of transitions from A to B. The substitution cost is therefore symmetric.

A third alternative is to calculate pairwise distances from some theory-driven *prototypes*.[19]

Sequence alignment. When the cost matrix is defined, pairwise distances between the sequences are calculated as in probabilistic SA. Optimal matching (OM) algorithm, described in the previous section, has been used most often in life sequence analysis. However, changing the order of states with insertions and deletions (indels) have been criticized for warping the time in an unnatural way.[20,21] A generalization of the Hamming distance[22] is a special case of optimal matching where indels are not used, and thus only states at the same position (time) are aligned.

The assumption of independent positions within a sequence may be a reasonable approximation to reality in bioinformatics but unrealistic in life course analysis. Most of the criticism of life sequence analysis has been directed at the ignorance of time order, which is so fundamental in prospective analysis (e.g. Wu[23]). In the probabilistic setting, it would be natural to model a life sequence as a Markov chain and generalize the independent elements (iid) assumption by assuming homogeneity or non-homogeneity of the chain.

In non-probabilistic SA, several ad-hoc alternatives for *duration-dependence* have been proposed to account for the length of time spells in sequence comparison. Stovel et al.[24] used a decay-function, which depends on a specific index period. Halpin[16] suggested a variant of OM that makes substitutions and indels cheaper for long spells than short spells. Marteau[25] used time-warping, which locally compresses or expands the time-scale to minimize the distance to the other sequence. Lesnard[21] proposed a time-dependent cost matrix for each time unit, depending on the neighbouring states (dynamic Hamming distance).

Non-alignment metrics. Elzinga[26–28] has taken a completely different approach, based on combinatorial methods, which does not require any cost matrix. The distance between two sequences is

generally defined as

$$d(x,y) = A(x,x) + A(y,y) - 2A(x,y), \tag{10}$$

where A is some sequence attribute. Some natural attributes are shown in table 2. As an illustration, the distance between the two sequences in example 1, now based on the LCS metric (the length of the longest common subsequence), is shown in example 2.

Example 2

sequence 1: **AAA**A**BB**BB
sequence 2: **AAABB**CC

The longest common subsequence of sequences 1 and 2 is AAABB of length 5, so the distance based on the LCS metric is $8 + 7 - 2 \times 5 = 5$.

While being intuitively meaningful and more objective than user-defined cost matrices, these distance criteria usually produce quite different results compared to alignment methods and have not been used often in real applications. Table 2 summarizes differences of some distance metrics used in life sequence analysis.

Censoring. A common problem in life history data are censored observations which in sequence analysis amounts to sequences of uneven length. The assumption of uninformative censoring in EHA is closely related to prediction; the predictions of observable participants are assumed to also be valid for the censored cases. In SA, the problem is how an incomplete observation window for some individuals affects the distance values. The solution is either to simply use shorter sequences or to extend the state space with a new "missing" state. Table 2 summarizes how censoring is handled with different metrics.

Multiple life domains. In the case of one life domain only, the alignment procedure is straightforward and most problems are related to the choice of the distance metric and the definition of the substitution costs. Multiple interdependent life domains complicate analysis in that not only does the state space grow rapidly, but the meaningfulness of the substitution costs also becomes more of an issue. For non-alignment metrics no methods for multi-domain sequences have been proposed to date. For alignment methods at least two approaches have been suggested.

In the *extended alphabet* approach, the letter corresponding to a particular state is replaced by a combination of letters (e.g. being simultaneously in states A, C, G, and J is denoted by ACGJ).[13,17,29] This can extend the state space rapidly. A conceptual problem is that the same cost matrix is applied to all states although it is not straightforward what a substitution of one state with another means in this approach. Gauthier et al.[30] define instead a separate cost matrix for each life domain $c = 1, \ldots, C$ and take an *average* of the costs at each position. If $s_c(x_i, y_j)$ is the cost for aligning x_i with y_j for the life-domain c, the average substitution (or indel) cost is calculated as

$$s(x_i, y_j) = \frac{\sum_{c=1}^{C} s_c(x_i, y_j)}{C}. \tag{11}$$

Typology of sequences. Once the distance matrix has been obtained with some of the alternative metrics in table 2, the goal is to find a typology of sequences by means of clustering methods. In life course studies, the differences between sequences should somehow be related to the timing of events, lengths of episodes determined by onset events, and the complete lack of some events or episodes.

Several alternatives are again available. In life sequence applications, Ward's agglomerative algorithm[31] is most commonly used because it tends to produce more equal-sized clusters than other

clustering algorithms, and this has been preferable for interpretation purposes. At each step, the algorithm combines the two clusters that minimize the within-cluster variability. No unique typology may exist if several pairs of sequences have the same distance value (i.e. there are ties) because a random start of the clustering algorithm can lead to different clustering results.

To determine the optimal number of clusters, generalizations of the usual goodness-of-fit statistics, coefficient of determination R^2 and F-test for non-Euclidian metrics have been used .[32] The sums of squares

$$SS = \frac{1}{n} \sum_{x=1}^{n} \sum_{y=x+1}^{n} d(x, y) \tag{12}$$

are now based on the chosen dissimilarity criterion $d(x, y)$ between sequences x and y. The pseudo R^2 and pseudo F-test, although defined as usual as the ratio of the between and within sum of squares, and that multiplied with the ratio of the degrees of freedom, can now have a different interpretation than in the Euclidean metric.

3 Application to life history calendar data

3.1 The JYLS Study

We illustrate the differences of the prospective and retrospective approaches with the Jyväskylä Longitudinal Study of Personality and Social Development (JYLS), ongoing in Finland. The participants, born in 1959, have been followed from age 8 to 50.[33] In 1968, twelve randomly selected second-grade classes in Jyväskylä, Central Finland, were chosen for the study. All of the pupils participated, so the initial attrition was zero. The original sample consisted of 173 girls and 196 boys. During the follow-up, no systematic attrition has been found.[34,35]

A LHC was used to retrospectively collect information about partnership status, children, studies, and work, as well as other important life events. The occurrence, timing, and duration of the transitions were recorded annually from age 15 to age 42 (in 2001[36]) and from age 42 to age 50 (in 2009) during interviews in which 275 participants gave reports based on memory and visual aids provided by the LHC-sheet (Figure 1). The information collected with the LHC was complemented using other sources of information, such as life situation questionnaires and interviews at ages 27, 36, and 42.

Figure 1: A section of the first life history calendar of the JYLS study.

Year													
Marriage/cohab. Age	15	16	17	18	19	20	21	22	23	24	25	...	42
Partner(s)													
Children	15	16	17	18	19	20	21	22	23	24	25	...	42
First child													
Second child													
⋮													
Other parenthood													
Education	15	16	17	18	19	20	21	22	23	24	25	...	42
Type of education													
Work	15	16	17	18	19	20	21	22	23	24	25	...	42
Fulltime work													
⋮													

Table 2: Comparison of non-probabilistic sequence analysis metrics.

	Alignment			Either[a]	Non-alignment		
Method	**Optimal matching**	**Hamming**	**Dynamic Hamming**	**Longest common subsequence**	**Longest common prefix/postfix**	**Number of common subsequences**	**Number of matching subsequences**
Operations/ attributes	Substitution, indels	Substitution	Time-varying substitution	Indels/subsequence	Substring	Subsequences	
Cost definition	substitutions: user-defined, transition probabilities, prototypes; higher indel costs favour substitutions, lower indels (in OM)			Constant	Not relevant		
Computing	Dynamic programming	Sum of substitutions		Dynamic programming	Direct comparison	Dynamic programming	
Principle of similarity	Most common states				Exact prefix/postfix	Same order of states (ignoring repetition)	Same order of states (counting repetition)
Sequences of uneven length	Insert/delete elements	Add missing states		No action			
Multidimensional sequences	Possible				Not yet possible		

[a]The longest common subsequence metric can be seen as either an alignment metric (using only indels) or a non-alignment metric (with the length of the longest common subsequence as the attribute).

We compared event history methods and sequence analysis in a setting where the dynamics of three inter-dependent life domains – partnership formation, parenthood, and employment – are studied in parallel. In EHA, we specified a multistate model for the event-specific transitions and in SA we specified domain-specific cost matrices. As a more substantive question, we studied the relationship between different life paths and excess depressive symptoms in middle age. These were assessed at age 42 using a shortened version of General Behavior Inventory (GBI).[37,38]

3.2 A multistate model

We note first that all events (partnership formation, child births, and career events) can be repeated several times in an individual's life course, making some simplification necessary. We limited the state space to the first transitions in each domain. In particular, we defined "employment" as the year when the person definitively had entered working life. The timings of initial partnership (either marriage or cohabitation) and parenthood are usually easily defined, but the onset of steady employment requires some thought. We defined it as the year which was followed by two subsequent years of employment. Studying and working in the same year was coded either in accordance with the subject's individual situation. The hazards for these events are shown in figure 2, while a histogram of GBI depression scores at age 42 is presented in figure 3.

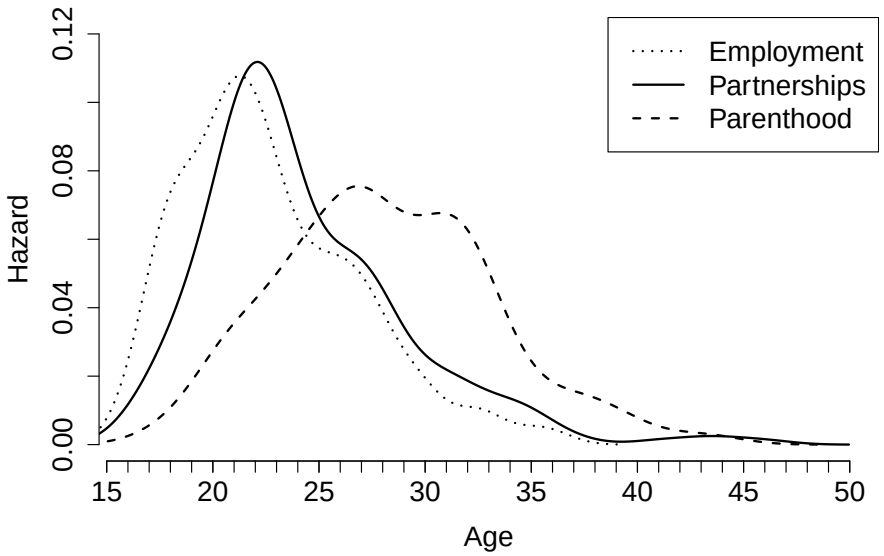


Figure 2: JYLS data: smoothed hazards of initial partnership, parenthood, and employment by age.

We were interested in how the timing of initial partnership and steady employment affect the joint prediction of remaining childless and having excess depressive symptoms at age 42. Excess depressive symptoms was defined as a higher than median GBI score value ($GBI_{med} = 1.44$).

For the sake of simplicity, we excluded cases who had become a parent before the prediction time (age 20) and also one case who had incomplete information on the transitions. This led to a sample size of 260 cases. Figure 4 shows the possible transitions between the states.

The events of interest were denoted by W=entering working life, P=forming an initial partnership, and C=becoming a parent, and their occurrence times by T_W, T_P, and T_C, respectively. The time interval of the LHC recordings was one year and we denote this interval by t, where $t = 20, \ldots\ldots, 42$. Because of the coarse data, it was possible for two or all three events of interest to occur within the same year. Since in that case we do not know the order of events, we simply multiply the discrete hazards in that year in the prediction formulae.

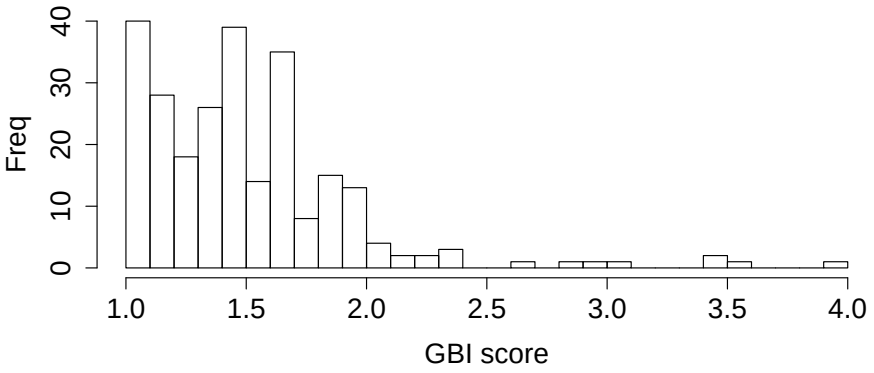


Figure 3: JYLS data: histogram of GBI depression scores at age 42.

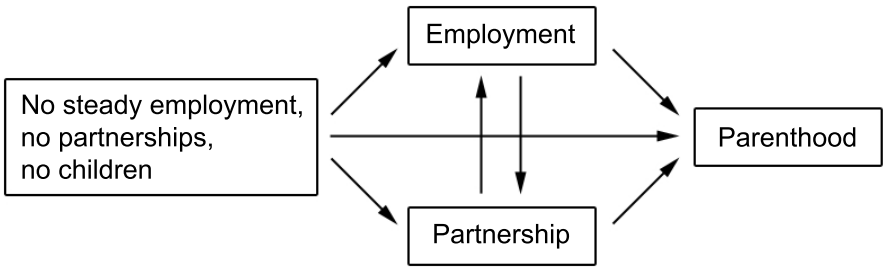


Figure 4: JYLS data: the multistate event history model.

Event-specific hazards. The discrete hazard of entering working life (W) at age t when neither an initial partnership (P) nor parenthood (C) has yet occurred, can be written in the general form as

$$p_W(t) \quad = \quad P(T_W = t | T_W \geq t, T_P \geq t, T_C \geq t). \tag{13}$$

Since any of the events P, W, or C can occur first, a similar hazard model can be defined for initial partnership and parenthood. If both P and W have already occurred at times $w \lessgtr v < t$, the conditional hazard of having a first child at age t is then

$$p_{C|WP}(t|v,w) \quad = \quad P(T_C = t | T_W = v, T_P = w, T_C \geq t). \tag{14}$$

Other conditional hazards are defined in an obvious way.

We used piecewise constant logistic hazard models where

$$p_x(t) = (1 + exp(-\beta' Z_t))^{-1} \tag{15}$$

is the discrete hazard of event x. The effect of the preceding events was modelled with time-dependent covariates which were simple indicators because the sample size did not allow for more complicated modelling. For example, in the hazard $p_{P|W}(t|v)$, the covariate $Z_t(W) \equiv 1$, $t \geq v$, when W occurred at v, whereas in $p_P(t)$ the covariate $Z_t(W)$ was not defined. Although possible, no other covariates were used in the models. Men and women were both included in the final model because no apparent differences in the effects of timing of partnership and work on the response event were found in separate analyses.

Prediction probabilities. In the multistate model the possible paths of not having children within the prediction interval depend on the occurrence times of initial partnership and steady employment.

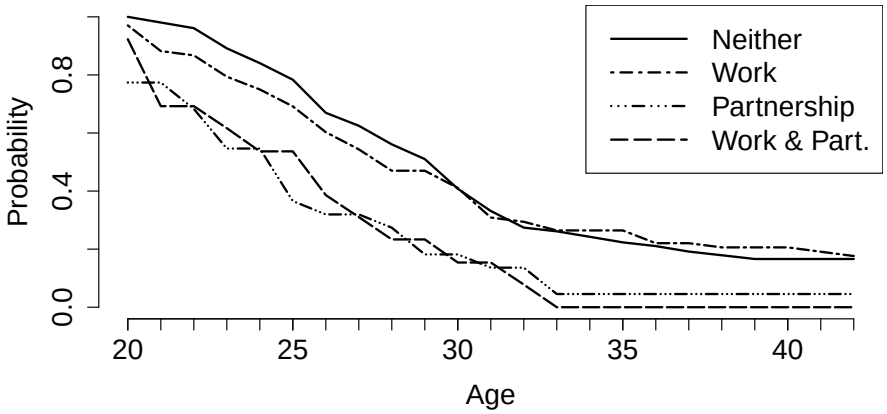


Figure 5: Survival probabilities of not having children for individuals who have or have not entered working life or initial partnership by the prediction time, age 20. "Neither" corresponds to no initial partnership nor employment by age 20.

The most complicated situation is when nothing has yet happened by the prediction time t. In this case, we must account for all possible timings of partnership and employment. We then have the prediction

$$
\begin{aligned}
&P(T_C > u | T_W > t, T_P > t, T_C > t) = \\
&\quad \prod_{s=t+1}^{u} (1 - p_W(s) - p_P(s) - p_C(s)) \\
&\quad + \sum_{s=t+1}^{u} \prod_{r=t+1}^{s-1} (1 - p_W(r) - p_P(r) - p_C(r)) p_W(s) \\
&\quad \times P(T_C > u | T_W = s, T_P > s, T_C > s) \\
&\quad + \sum_{s=t+1}^{u} \prod_{r=t+1}^{s-1} (1 - p_W(r) - p_P(r) - p_C(r)) p_P(s) \\
&\quad \times P(T_C > u | T_W > s, T_P = s, T_C > s) \\
&\quad - \sum_{s=t+1}^{u} \prod_{r=t+1}^{s-1} (1 - p_W(r) - p_P(r) - p_C(r)) p_W(s) p_P(s) \\
&\quad \times P(T_C > u | T_W = s, T_P = s, T_C > s). \qquad (16)
\end{aligned}
$$

The last sum accounts for the paths in which P and W occur within the same year and their order is unknown.

The other paths are special cases of (16). In particular, when initial partnership (P) and entering working life (W) have occurred by the prediction time t, the prediction is simply, for $0 < v \lesseqgtr w < t < u$,

$$
P(T_C > u | T_W = v, T_P = w, T_C \geq t) = \prod_{s=t+1}^{u} (1 - p_{C|WP}(s|v, w)). \qquad (17)
$$

The prediction probability is a function of the prediction time t and the prediction interval $I = (t, u]$ and its realizations depend on the history H. By letting one of them be variable and fixing the values of the other two, we can obtain different views of the life course dynamics. In figure 5, we obtain the usual survival probability $S(u)$ of not having children by age u when fixing the prediction

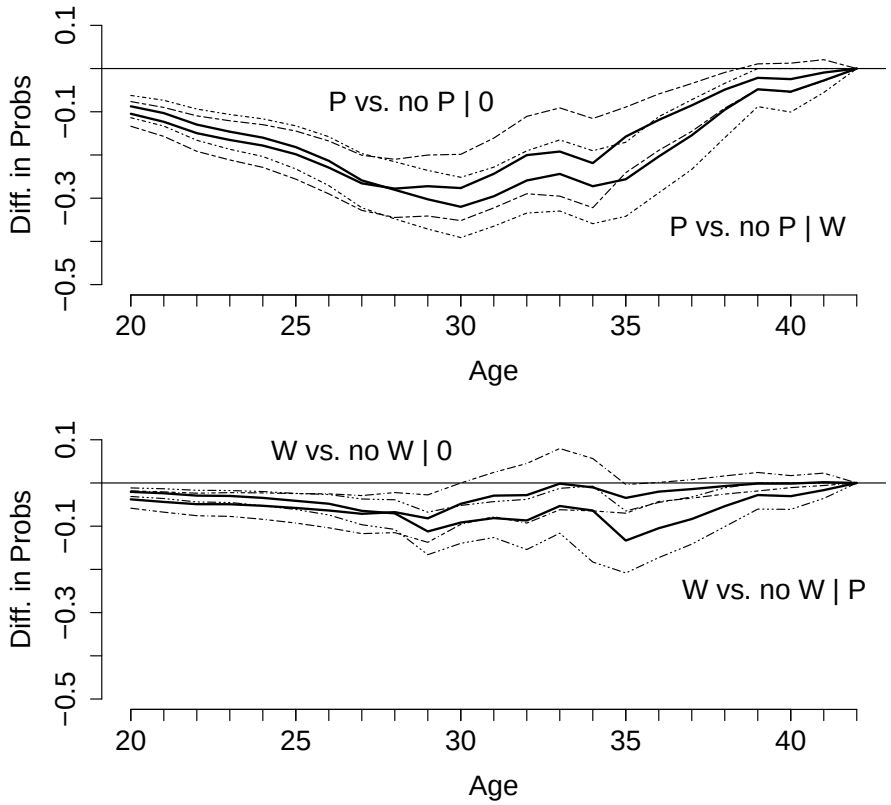


Figure 6: Innovation gains in predicting no children by age $u = 42$ from observing employment (W) and initial partnership (P) at the prediction time $t = 20, ..., 42$, given that nothing/the other one has occurred previously (0 = nothing has yet happened). The confidence intervals are based on 5000 bootstrap samples of the data.

time at $t = 20$ and history at H_{20} and letting the prediction interval vary with ages $u = 21, ..., 42$. We notice that half of those who had formed initial partnership already by age 20, had children by age 25, whereas the effect of employment by age 20 had a much smaller effect on early parenthood compared to those who had neither formed initial partnership nor entered working life at that age.

Factual and counterfactual predictions. Instead of fixing the prediction time t we now identify it with the variable occurrence time $t = 20, ..., 42$ of either initial partnership or employment. When comparing these predictions at age $u = 42$, we obtain a visual representation of the effect of timing of initial partnership P and employment W on the prediction of no parenthood by age 42. In figure 6, we consider the difference of the two predictions:

$$P(T_C > u | T_P = t, T_W > t, T_C > t) - P(T_C > u | T_P > t, T_W > t, T_C > t). \tag{18}$$

This is the *innovation gain* from observing initial partnership at age $t = 20, ..., 42$ related to the prediction of not having children by age 42, given no steady employment by age t. If a person actually forms an initial partnership at age t, the first probability is a *factual* prediction of not having children by age u, given the history, and the second probability is a *counterfactual* prediction of the same event.

At all ages, both initial partnership and employment decreased the probability of remaining childless compared to the situation where neither has occurred yet. The 95% confidence limits show, however, that the timing of steady employment had a significant effect only if it occurred before age 30 if no partnership had been formed yet. Initial partnership around ages 28 to 31 decreased the

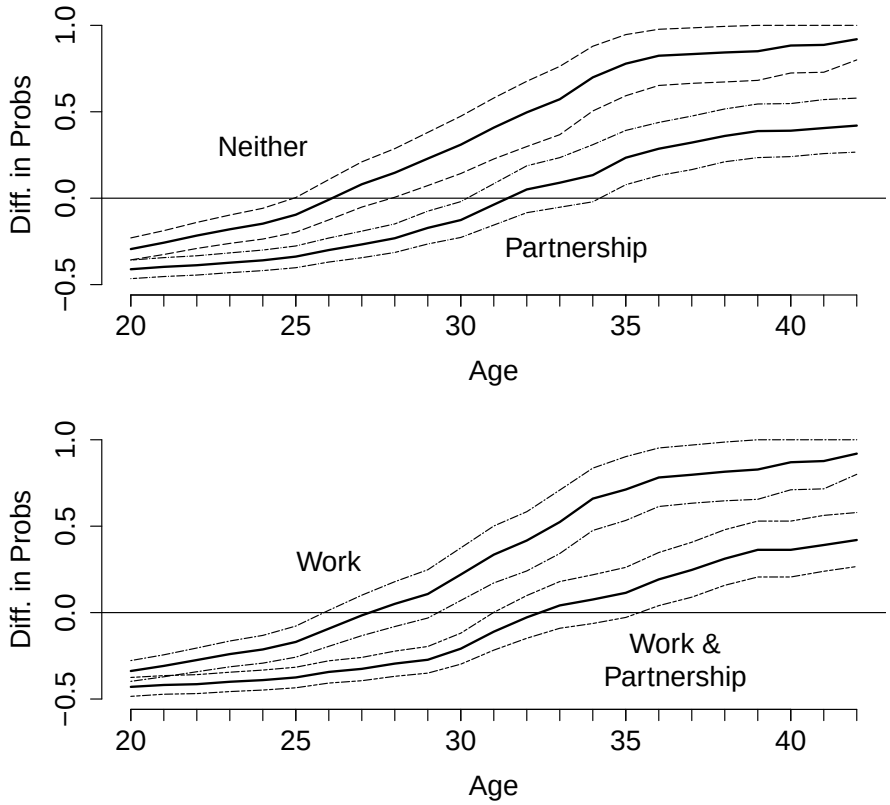


Figure 7: Difference in the joint prediction probabilities of excess depressive symptoms and having children versus not having children by age 42, given that employment or/and initial partnership have occurred at the prediction time $t = 20, ..., 42$. "Neither" corresponds to no partnership or employment (yet) at the time of prediction. The confidence intervals are based on 5000 bootstrap samples of the data.

prediction of remaining childless the most, but had a significant effect at any age. It should be noted that, while controlling for the history effect, the size of the innovation gain from observing initial partnership depends on the length of the remaining prediction interval.

Joint prediction probability. Finally, to evaluate the relationship between possible histories of family formation and employment with depressive symptoms (D) in middle age, we compared the joint prediction of parenthood/no parenthood and excess depressive symptoms at age 42, given the history of partnership and employment. For the case of having children, we then have

$$P(T_C \leq 42, D_{42} > d^* | H_t) = P(D_{42} > d^* | T_C \leq 42, H_t)(1 - P(T_C > 42 | H_t)) \tag{19}$$

with obvious changes for the case of no children.

The first probability on the right is evaluated only at age $u = 42$, so it only affects the last terms at time $u = 42$ in the prediction formulae. It is the cross-sectional logistic probability for a higher than median GBI score d^* at age $u = 42$, depending on family formation and employment

$$p_D(42) = logit(P(D_{42} > d^* | \mathbf{Z}_{42})) = \alpha + \beta_1 Z_{42}(W) + \beta_2 Z_{42}(P) + \beta_3 Z_{42}(C) \tag{20}$$

where $Z_{42}(C) = 1$ for the case when $T_C \leq 42$ and $Z_{42}(C) = 0$ for the case when $T_C > 42$. Since all these covariates were indicators, the occurrence times did not make a difference.

Figure 7 shows the differences in the joint prediction probabilities of having children versus not having children by age 42 and excess depressive symptoms at that age, given initial partnership or employment at the time of prediction. This analysis provides "limiting" ages for increasingly higher prediction of excess depressive symptoms in middle age and remaining childless, compared to having children. We find that if initial partnership is formed later than at age 31, the difference of these joint probabilities becomes positive and increasing. For steady employment but no initial partnership, this age limit is about 27 years. For those who have no initial partnership nor steady employment at the prediction time, this limit is reached already at age 26. By age 34, the prediction of excess depressive symptoms and no children is already about 80% higher than the prediction of excess predictive symptoms and children.

This analysis shows that, having estimated the event-specific hazards, we can evaluate joint predictions of events related to both dynamic and non-dynamic parts of a multistate model. Including explanatory covariates in the hazard models (which we did not do), would allow to compare predictions of hypothetical individuals with different histories and characteristics.

3.3 Multidimensional sequence analysis

In sequence analysis, instead of transitions we studied the distribution of individuals in the states year by year. This difference corresponds to annually evaluating the prevalence of the states instead of incidence. We define the state space of partnership, parenthood, and career histories from age 15 to 50 as shown in table 3.

Table 3: Life domains and respective states for three-domain sequence analysis.

Life domain	States
Partnership	single, in partnership, divorced/separated/widowed
Parenthood	no children, has children (biological, adopted, foster)
Career	studying, working, other (unemployed, out of labour force)

Unlike in the EHA example, we did not restrict the analysis to the first events but used all events for the three life domains. This state space results in 18 possible state combinations for each year. The transition matrix was sparse, but in non-probabilistic sequence analysis and with domain-specific costs this is not a serious issue.

In our data, sequence lengths vary because of the two data collection phases and small differences in ages: 215 participants have sequences of length 36, 14 participants of length 35, and 46 participants of length 28.

Dissimilarity criteria. We compared six dissimilarity metrics suitable for multidimensional sequence analysis. They were based on different definitions of the substitution costs. In optimal matching (OM) and Hamming distance, we used either user-defined or data-driven substitution costs. In dynamic Hamming distance, they were based on estimated transition probabilities taking into account the neighbouring states of the previous and the following year.[21] The LCS criterion (the length of the longest common subsequence) corresponds to OM with the specific choice of substitution cost 2 and indel cost 1.

Since Hamming distance does not allow indels, censored positions were replaced by a "missing" state. The effect of different costs for missing states was investigated by defining costs 0, 0.5, or 1 times the largest substitution cost. With larger costs, the sequences with missing states tend to form their own uninformative cluster. Thus, using no cost at all resulted in the best results.

Table 4: Partnership, parenthood and career-related substitution costs based on theory (user-defined) or transition probabilities.

	User-defined				Transition probabilities			
	→ S	→ P	→ D	→ *	→ S	→ P	→ D	→ *
Single (S) →	0	2	3	0	0	1.89	2.00	0
Partnership (P) →	2	0	1	0	1.89	0	1.80	0
Divorced/sep. (D) →	3	1	0	0	2.00	1.80	0	0
Missing (*) →	0	0	0	0	0	0	0	0

	User-defined			Transition probabilities		
	→ N	→ C	→ *	→ N	→ C	→ *
No children (N) →	0	3	0	0	1.94	0
Has children (C) →	3	0	0	1.94	0	0
Missing (*) →	0	0	0	0	0	0

	User-defined				Transition probabilities			
	→ S	→ W	→ O	→ *	→ S	→ W	→ O	→ *
Studying (S) →	0	3	1.5	0	0	1.77	1.87	0
Working (W) →	3	0	1.5	0	1.77	0	1.67	0
Other (O) →	1.5	1.5	0	0	1.87	1.67	0	0
Missing (*) →	0	0	0	0	0	0	0	0

Combined substitution costs. The substitution cost matrix was defined separately for each three domain and then averaged. The domain-specific costs for OM and Hamming distance are shown in table 4. For dynamic Hamming, time-specific costs resulted in 36 distinct substitution cost matrices (not shown here). It should be noted that the absolute numbers in user-defined substitution costs have no meaning since the information is only relative. In our application, the states "single" and "divorced" were the most distant because forming a partnership was regarded as one step in the developmental process to adulthood. In another study, these could be interpreted as similar, as both indicate a state of "living without a partner". Compared to the transition-based costs, this is the main difference in the partnership domain. For career domain, transitions from states "studying" to "other" (or vice versa) were the least common (highest cost in the cost matrix based on transition probabilities), but in the user-defined matrix the corresponding cost was set relatively low, due to the versatile nature of the state "other". The indel costs were set to half of the largest substitution cost, making them equally costly. For averaging, the costs in each matrix were scaled to have the same range in order to give equal weight to each life domain.

Typology of sequences. Ward's agglomerative clustering was used to find a typology of life sequences, applying the six dissimilarity criteria for solutions starting from 2 to 15 clusters. Based on dendrograms, the goodness-of-fit statistics, and interpretability of the clusters, an eight-cluster solution was chosen.

The goodness-of-fit statistics in table 5 for the chosen eight cluster solutions suggested that clustering based on the Hamming distance with theory-based substitution costs fits the data best. It covered around 45% of sequence variation ($F = 31.56$) and resulted in interpretable clusters where all three life domains were well represented. In comparison, the second best criterion, dynamic

Table 5: Goodness-of-fit statistics for eight cluster solutions obtained with six distance measures based on transition probabilities or user-defined costs.

Dissimilarity measure	Pseudo R^2	Pseudo F
Hamming distance (user-defined)	0.453	31.56
Dynamic Hamming distance (trans. prob.)	0.433	29.18
Optimal matching (user-defined)	0.406	26.09
Hamming distance (trans. prob)	0.395	24.93
Optimal matching (trans. prob)	0.369	22.33
Length of longest common subsequence	0.358	21.23

Figure 8: The dendrogram of the clustering based on Hamming distance with user-defined substitution costs.

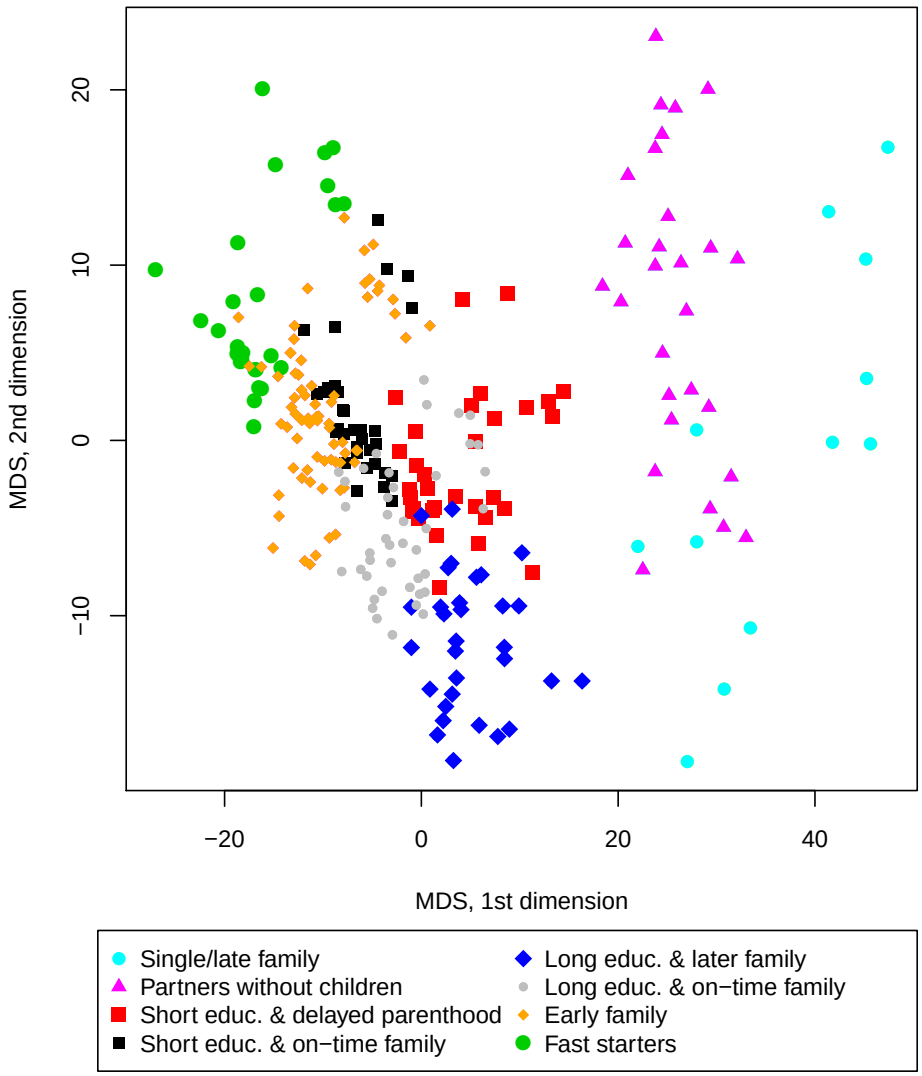


Figure 9: Scatter plot of the cluster-specific MDS scores based on the first two dimensions of multidimensional scaling. Dissimilarities were computed by using theory-based Hamming distance for the three-dimensional sequences.

Hamming, resulted in clusters where the family-related life domains dominated and the career domain was hardly represented. The dendrogram based on the Hamming distance in figure 8 supported the eight cluster solution.

Time-preserving Hamming distance instead of OM seems more reasonable for sequences of uneven length. In OM, using indels in our data would mean aligning, for example, a state at age 15 with a state at age 23 in another sequence. Within the same metric, user-defined substitution costs gave better results than costs based on transition probabilities in this three-domain setting. However, preliminary studies with only one life domain suggested the opposite so no general guidelines can be given.

We present the results with the Hamming distance, at the same time illustrating different ways of investigating the clustering results by sequence plots, by comparing sequence variation in clusters, by reducing dimensionality in the multivariate categorical analysis with multidimensional scaling and finally, by using the cluster indicators in the regression of depressive symptom scores on cluster membership.

Table 6: Logistic regression of depression score on cluster membership.

	β	s.e.	p	OR
Short education & delayed parenthood	−0.21	0.37	0.58	0.81
Short education & on-time family	−0.51	0.37	0.16	0.60
Long education & late family	−0.59	0.39	0.14	0.56
Partners without children	−0.24	0.40	0.55	0.79
Early family	0.18	0.25	0.46	1.20
Single/late family	1.61	0.77	0.04	5.00
Long education & early partnership	−0.05	0.33	0.87	0.95
Fast starters	0.47	0.40	0.24	1.60

Multidimensional scaling (MDS). MDS provides a concise visual representation of cluster results, first by showing how well the clusters actually separate but also by providing a visual aid when the original sequences are ordered according to MDS scores. The first few scaling dimensions capture the most prominent variation in the sequences. The rotation of the solution is arbitrary, but principal component axes can be used for achieving a meaningful rotation. The resulting dimensions often sort the sequences according to an attribute, such as the timing of some transition.

In figure 9, the sequences were plotted as points on the plane spanned by the first two MDS dimensions with cluster identification. The eigenvalues of the MDS solutions with different dimensions supported two MDS dimensions. Correlation between the original Hamming distances and the distances computed from two-dimensional MDS scores was 0.93. The timing of initial partnership and parenthood seemed to separate the clusters best (1st principal component dimension); length of education follows (2nd dimension). Clusters of individuals with no children were clearly separated from the others, which were more or less connected but not completely overlapping.

Sequence plots. *Index plots* show the individual life courses, merely re-organizing the original data according to the similarity defined by clustering (figure 10). Ordering according to some MDS dimension assists in interpretation. *State distribution plots* show the prevalence of states at each time point. We combined the different life domains to give an overview of the dynamics of the state distribution (figure 11).

Sequence variability. Shannon's entropy[39,40] is often used as a measure of disorder of a system. In life sequence analysis, entropy is used to characterize variation in the states within one sequence, or more interestingly, within and between clusters of sequences. When entropy is 0, all cases (of a cluster) are in the same state. When entropy is 1, there are equally large amount of cases in each state. Important transition times are easily seen as peaks in the cluster-specific plots (figure 12).

Regression analysis. External explanatory variables can be taken into account either in the clustering phase (covariance analysis instead of ANOVA), or as independent variables in a multinomial analysis of cluster membership indicators. We used the membership indicators as explanatory factors in a logistic regression predicting higher than median depression scores as in EHA. The "single/late family" cluster was the only one that shows statistically significant differences (with higher odds of having excess depressive symptoms). This result supports the finding of Salmela-Aro et al.[41] that postponing or lack of some stages in the transitory process to adulthood anticipate lower life satisfaction in adulthood. Note that although we used individual-specific cluster membership indicators in the regression models, the cluster characteristics may not be representative to all members of the cluster. Clustering was based on the matrix of pairwise distances, not on the individual sequences any more. It was therefore expected that only the most different clusters (here singles) would have

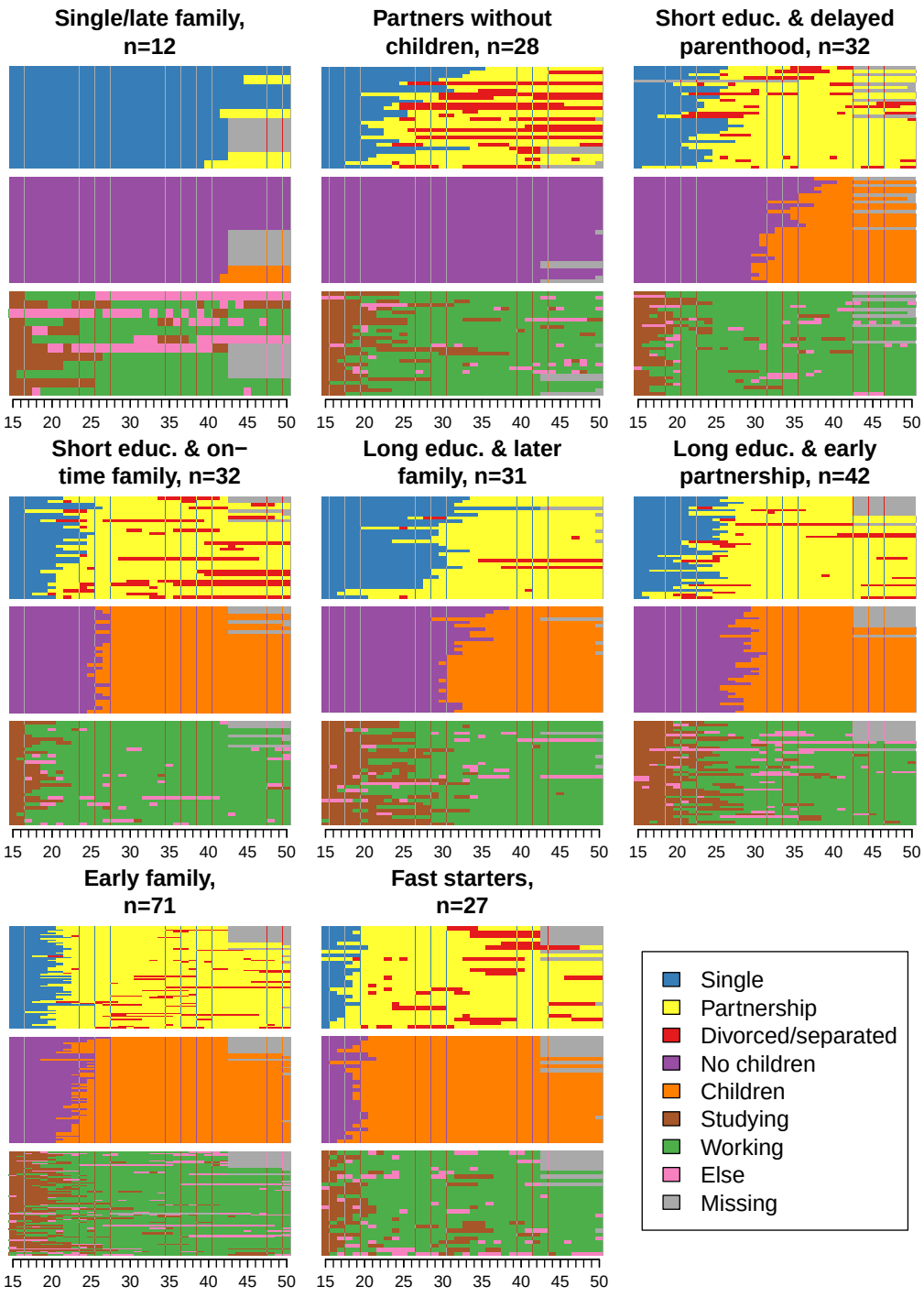


Figure 10: Index plots of partnership (top), parenthood (middle), and career (bottom) in the eight clusters based on Hamming distance. The sequences are ordered according to the first dimension of multidimensional scaling that represents the timing of partnership and children.

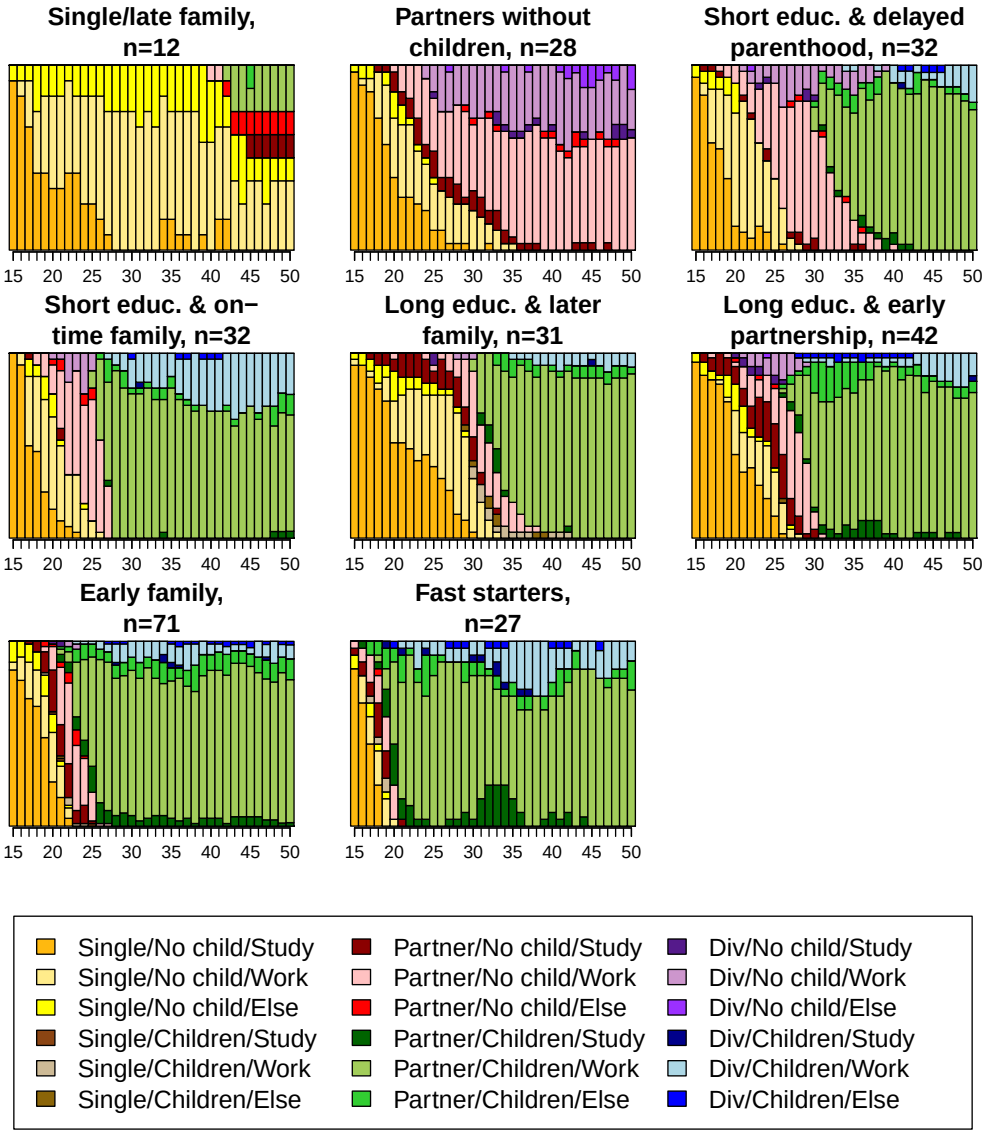


Figure 11: State distribution plots of combined partnership, parenthood, and career states in the eight clusters based on Hamming distance. Note, that "divorced" can mean either a broken marriage or cohabitation. Positions with missing states in any life domain are excluded from the plots.

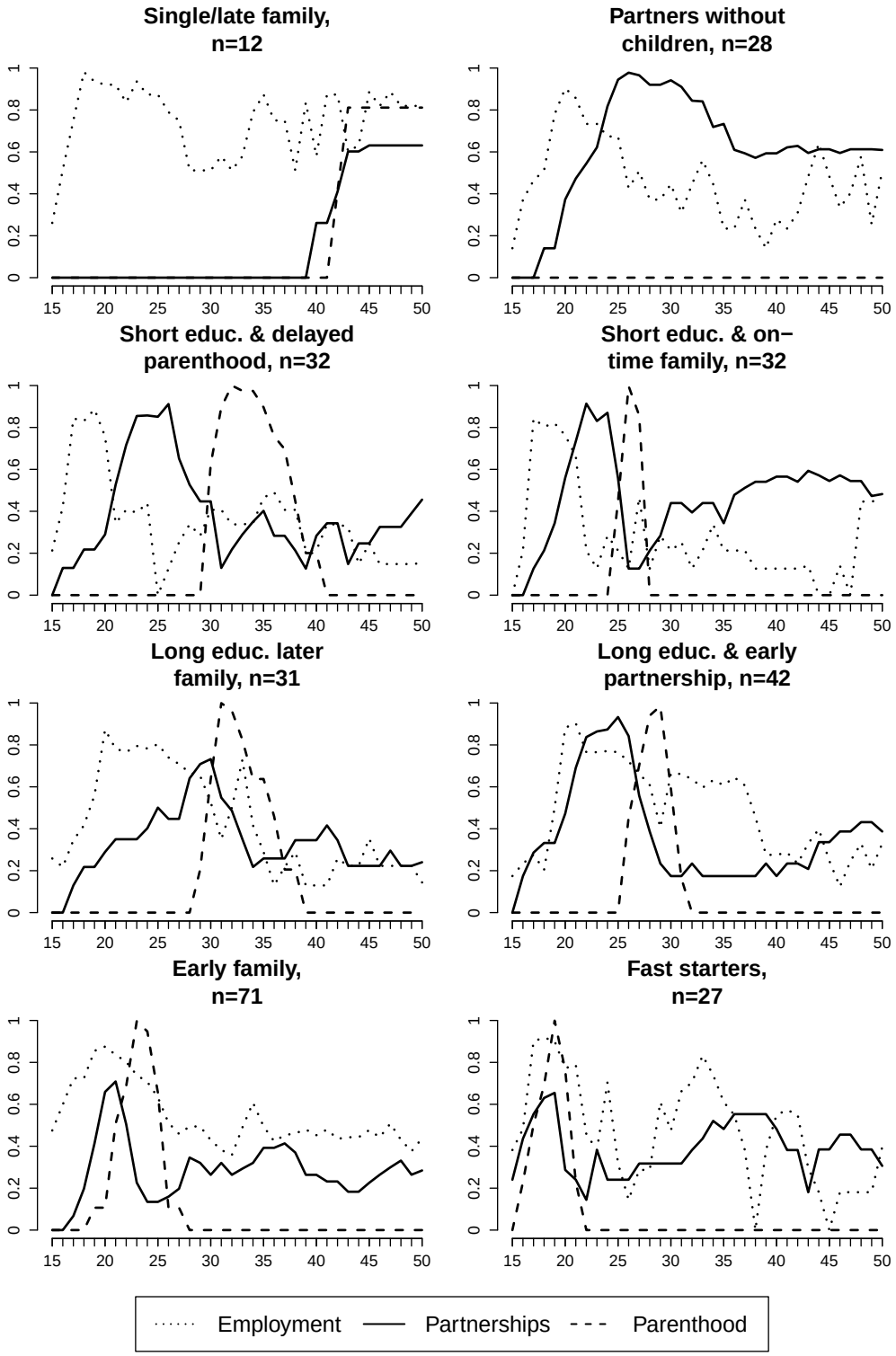


Figure 12: Transversal entropies of partnerships, parenthood, and career sequences in clusters based on Hamming distance.

a significant role in the regression analysis. We conclude that, unlike making individual-level predictions of parenthood given the history of partnership and employment, the aim of sequence analysis was to find subpopulations or clusters of individuals whose life courses were similar in terms of the timing of initial partnership, parenthood and employment.

Computations. Sequence analyses were carried out with the *TraMineR* library in *R.*[42] Logistic hazard models and programs calculating the prediction probabilities and their bootstrap intervals in section 3.2 were implemented with R.

4 Discussion

We compared two approaches of analysing data collected with a life history calendar: the model-based probabilistic method of event history analysis and the model-free data mining method of sequence analysis. Traditionally, EHA models the risk of a transition from one state to another, but here we instead estimated the cumulative prediction probabilities of life events in a multistate model to have a more comparable setup with sequence analysis. Instead of transitions, the analysis was extended to the entire observed trajectory, which was the unit of analysis in SA as well. In sequence analysis, we compared several dissimilarity metrics and contrasted data-driven and user-defined substitution costs. To illustrate the two methods, we studied young adults' transition to adulthood as a sequence of landmark events in several life domains. These landmark events defined our multistate event-history model and the parallel life domains in multidimensional SA. Finally, we analysed the relationship between life trajectories and excess depressive symptoms at age 42 by first estimating their joint predictions in the multistate model and then by using the individual-specific cluster indices of multidimensional SA in a further explanatory analysis for depressive symptoms.

When the same life course problem was analysed with both methods, we found that the two approaches complement each other. SA is a descriptive tool synthesizing large amount of information to obtain a broad picture of multidimensional data. As other dimension-reducing methods, SA helps developing an intuitive understanding of complex relationships but the resulting clusters should not be given a confirmatory status. Finding descriptions for the clusters mirrors the rather subjective way of naming factors in factor analysis. In our case, sequence analysis could reveal typical and atypical patterns of young adults transition process to adulthood which supported the earlier findings that no normative pathway to adulthood exists any longer. Individuals' enhanced opportunities to make choices in their own lives increases diversity in the life course. These individual choices are affected by various governmental and other external decisions, the results of which are difficult to conceive at the population level. In particular, sequence analysis has offered new means for large scale comparative analysis of life patterns across nations and between age cohorts.

Multistate event history analysis, on the other hand, is a predictive method which requires structured hypotheses and a well-defined system of hazard models. This is opposite to the data mining approach of SA in which no assumptions about the data generating mechanisms are made, or needed, for that matter. We believe, however, that the analysis of increasingly complex life course data, combining perhaps both biological and behavioural data, will require methods that at the initial stage can reveal underlying structures and help generating causal hypotheses for further analysis. Causal inquiries can only be addressed with proper "book keeping" of risk sets for transitions. Thus, correct individual-level conditioning of the history is possible only in EHA. Multiple time scales, inherent in many life course problems, and their separate effects can only be quantified by modelling. Furthermore, time-varying covariates indicating individual status changes or contextual changes in time are only possible in model-based analysis.

Sequence analysis has been criticized for violating the basic principles of prospective analysis because the "past" and the "future" are treated symmetrically in vertical alignment. In this sense,

it is not suitable for any causal analysis. Subjectivity of substitution cost specification and non-uniqueness of clustering results have also raised scepticism about its usefulness. In recent years, several improvements have been suggested to the specification of substitution costs, to handle censoring, and to preserve timing and the order of states in sequence analysis.[43] They all modify the substitution cost matrix in some way because this is the only way of tuning the values of the distance matrix. According to our examinations, also Elzinga's non-alignment methods[26–28] seem promising, but no multidimensional method exists yet. As a data mining method, SA is best suited for large register-based data sets. With small data sets and large state space, all trajectories tend to be unique. If the substitution cost matrix is based on estimated transition probabilities, small data sets run out of observations. This was shown by Helske et al.,[44] who used Hidden Markov models to cluster life sequences probabilistically.

Statistical analysis of life sequences still has many unresolved questions, compared to the well developed theory of event history analysis. Sequence analysis is less conventional, but its use is expected to increase in the future, especially now that there is an easy-to-use software available in R. Event history analysis will certainly remain the main tool for analytical life course studies. We believe that although the prediction probabilities are not a standard tool in EHA, they are valuable for synthesizing information in a multistate model. Although the probabilistic statements and programming require careful specification, the probabilities can be estimated in a straightforward manner from state-specific hazards. Confidence intervals can be calculated, for example, by bootstrapping (as we did) or, in a fully parametric case, analytically (cf. Eerola[8]).

As in epidemiology, prevalence indicates what is typical or atypical at a particular time, whereas incidence is related to change, the underlying concept in all causal inquiry. Life course analysis is obviously dynamic, but the complex pattern of interacting factors also requires "zooming" into details. Therefore, one could summarize the complementary advantages of the methods: while sequence analysis provides detailed information about "how things are", event history analysis answers the "why".

Acknowledgements. We thank The Jyväskylä Longitudinal Study of Personality and Social Development, led by Lea Pulkkinen, for letting us use the data in our study. We thank especially Katja Kokko and Eija Räikkönen for their comments.

References

[1] Belli RF, Stafford FP, Alwin DF. Calendar and time diary: methods in life course research. Sage Publications, Inc; 2008.

[2] Caspi A, Moffitt TE, Thornton A, Freedman D, et al. The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. International Journal of Methods in Psychiatric Research. 1996;6(2):101–114.

[3] Andersen PK, Borgan Ø, Gill RD, Keiding N. Statistical models based on counting processes. Springer Verlag; 1993.

[4] Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. Journal of Epidemiology and Community Health. 2003;57(10):778–783.

[5] Andersen PK, Keiding N. Multi-state models for event history analysis. Statistical Methods in Medical Research. 2002;11(2):91.

[6] Arjas E. Survival Models and Martingale Dynamics. Scandinavian Journal of Statistics. 1989;p. 177–225.

[7] Arjas E, Eerola M. On predictive causality in longitudinal studies. Journal of statistical planning and inference. 1993;34(3):361–386.

[8] Eerola M. Probabilistic causality in longitudinal studies. vol. 92 of Lecture Notes in Statistics. Springer-Verlag; 1994.

[9] Klein JP, Keiding N, Copelan EA. Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients. Statistics in Medicine. 1993;12(24):2315–2332.

[10] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine. 2007;26(11):2389–2430.

[11] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.

[12] Abbott A. Sequence Analysis: New Methods for Old Ideas. Annual Review of Sociology. 1995;21(1):93–113.

[13] Pollock G. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2007;170(1):167–183.

[14] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10; 1966. p. 707–710.

[15] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970;48(3):443–453.

[16] Halpin B. Optimal Matching Analysis and Life-Course Data: The Importance of Duration. Sociological Methods & Research. 2010;38(3):365–388.

[17] Stovel K, Savage M, Bearman P. Ascription into achievement: Models of career systems at Lloyds Bank, 1890–1970. The American Journal of Sociology. 1996;102(2):358–399.

[18] Rohwer G, Pötter U. TDA User's Manual; 2004.

[19] Wiggins R, Erzberger C, Hyde M, Higgs P, Blane D. Optimal matching analysis using ideal types to describe the lifecourse: an illustration of how histories of work, partnerships and housing relate to quality of life in early old age. International Journal of Social Research Methodology. 2007;10(4):259–278.

[20] Hollister M. Is Optimal Matching Suboptimal? Sociological Methods & Research. 2009;38(2):235–264.

[21] Lesnard L. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. Sociological Methods & Research. 2010;38(3):389–419.

[22] Hamming RW. Error detecting and error correcting codes. Bell System Technical Journal. 1950;29(2):147–160.

[23] Wu LL. Some Comments on "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect". Sociological Methods Research. 2000;29(1):41–64.

[24] Stovel K. Local sequential patterns: The structure of lynching in the Deep South, 1882–1930. Social Forces. 2001;79(3):843–880.

[25] Marteau PF. Time warp edit distance with stiffness adjustment for time series matching. IEEE transactions on pattern analysis and machine intelligence. 2008;31(2):306–318.

[26] Elzinga CH; Citeseer. Sequence similarity: a nonaligning technique. Sociological Methods and Research. 2003;32(1):3–29.

[27] Elzinga CH. Sequence analysis: Metric representations of categorical time series. Manuscript. 2006;.

[28] Elzinga CH, Liefbroer AC. De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. European Journal of Population/Revue européenne de Démographie. 2007;23(3):225–250.

[29] Han SK, Moen P. Clocking out: Temporal patterning of retirement. American Journal of Sociology. 1999;105(1):191–236.

[30] Gauthier JA, Widmer ED, Bucher P, Notredame C. Multichannel sequence analysis applied to social science data. Sociological Methodology. 2010;40(1):1–38.

[31] Ward JH Jr. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963;p. 236–244.

[32] Studer M, Ritschard G, Gabadinho A, Müller N. Discrepancy analysis of complex objects using dissimilarities. Advances in Knowledge Discovery and Management. 2010;292(4):3–19.

[33] Pulkkinen L, Lyyra AL, Kokko K. Life success of males on nonoffender, adolescence-limited, persistent, and adult-onset antisocial pathways: follow-up from age 8 to 42. Aggressive Behavior. 2009;35(2):117–135.

[34] Pulkkinen L. The Jyväskylä Longitudinal Study of Personality and Social Development. In: Pulkkinen L, Kaprio J, Rose RJ, editors. Socioemotional development and health from adolescence to adulthood. Cambridge University Press, New York; 2006. p. 29–55.

[35] Pulkkinen L, Kokko K. Tiivistelmä [Summary]. In: Pulkkinen L, Kokko K, editors. Keski-ikä elämänvaiheena [Middle-age as a stage of life]. Jyväskylän yliopisto, Jyväskylä; 2010. p. 5–13.

[36] Kokko K, Pulkkinen L, Mesiäinen P. Timing of parenthood in relation to other life transitions and adult social functioning. International Journal of Behavioral Development. 2009;33(4):356–365.

[37] Depue R. General Behavior Inventory; 1987. Ithaca, NY: Department of Psychology, Cornell University.

[38] Kokko K, Pulkkinen L. Unemployment and psychological distress: Mediator effects. Journal of Adult Development. 1998;5(4):205–217.

[39] Shannon C. A mathematical theory of communication. The Bell System Technical Journal. 1948;27(7):379–423.

[40] Billari FC. The Analysis of Early Life Courses: Complex Descriptions of the Transition to adulthood. Journal of Population Research. 2001;18(2):119–142.

[41] Salmela-Aro K, Kiuru N, Nurmi JE, Eerola M. Mapping pathways to adulthood among Finnish university students: Sequences, patterns, variations in family-and work-related roles. Advances in Life Course Research. 2011;16(1):25–41.

[42] Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software. 2011;40(4):1–37.

[43] Aisenbrey S, Fasang AE. New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. Sociological Methods & Research. 2010;38(3):420–462.

[44] Helske J, Eerola M, Tabus I. Minimum description length based hidden Markov model clustering for life sequence analysis. In: Proceedings of the Third Workshop on Information Theoretic Methods in Science and Engineering, August 16–18, 2010, Tampere, Finland; 2010. .

II

Helske, S., Steele, F., Kokko, K., Räikkönen E., and 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. doi:http://dx.doi.org/10.14301/llcs.v6i1.290

Partnership formation and dissolution over the life course: applying sequence analysis and event history analysis in the study of recurrent events

Satu Helske, University of Jyvaskyla
satu.helske@jyu.fi
Fiona Steele, London School of Economics
Katja Kokko, University of Jyvaskyla
Eija Räikkönen, University of Jyvaskyla
Mervi Eerola, University of Turku

(Received February 2014 Revised September 2014) http://dx.doi.org/10.14301/llcs.v6i1.290

Abstract

We present two types of approach to the analysis of recurrent events for discretely measured data, and show how these methods can complement each other when analysing co-residential partnership histories. Sequence analysis is a descriptive tool that gives an overall picture of the data and helps to find typical and atypical patterns in histories. Event history analysis is used to make conclusions about the effects of covariates on the timing and duration of the partnerships. As a substantive question, we studied how family background and childhood socio-emotional characteristics were related to later partnership formation and stability in a Finnish cohort born in 1959. We found that high self-control of emotions at age 8 was related to a lower risk of partnership dissolution and for women a lower probability of repartnering. Child-centred parenting practices during childhood were related to a lower risk of dissolution for women. Socially active boys were faster at forming partnerships as men.

Keywords: partnership formation, partnership dissolution, sequence analysis, event history analysis, recurrent events

1 Introduction

During the life course many events (such as marriages, child births, unemployment etc.) can occur several times to an individual. In this paper we present two approaches to the analysis of recurrent events for discretely measured data and show how these methods can complement each other when analysing co-residential partnership histories of a representative sample of Finnish men and women now in their fifties. The first method, *sequence analysis*, is a descriptive technique which we used to summarize all partner transitions made by individuals over the whole observation period. We grouped similar histories of forming and dissolving partnerships and searched for typical and atypical patterns. In contrast, *event history analysis* is a model-based method which we used to model the probability of making a transition to or from partnership in a given time interval as a function of possibly time-varying individual characteristics. Specifically, we examined how home background and socio-emotional characteristics in childhood were related to later partnership formation and stability, whether these effects differed between women and men, and if they played a part in a tendency to repartner.

1.1 Partnerships in a life course perspective

Establishment of an intimate relationship has been recognized as one of the milestones during the transition to adulthood (e.g. Shanahan, 2000). In the past, this typically meant the start of the first and only marriage. However, the choice of union type is now no longer confined to traditional life-long marriage, as cohabitation has become an integral part of family life in Western countries (Kennedy & Bumpass, 2008; Kiernan, 2001). Furthermore, it is increasingly common for people to enter a union more than once during their lives. As a result, partnership trajectories have become diverse according to the type and number of unions formed during the life course. Regarding the first union, cohabiting unions have been consistently found to be less stable than marriages (Poortman & Lyngstad, 2007). In the case of the second and higher-order unions, the picture is more complex. In general, second unions have been shown to be as stable as the first unions, when selection based on individual characteristics is controlled for (Aassve, A., Burgess, S., Propper, C., & Dickson, M., 2006; Lillard, Brien, & Waite, 1995; Poortman & Lyngstad, 2007; Steele, Kallis, Goldstein, & Joshi, 2005; Steele, F., Kallis, C., & Joshi, H., 2006).

It is likely that second and higher-order unions differ from the first union in that they often involve individuals with more complex life histories, including multiple spells of partnerships, children from previous relationships, and the continuing influence of previous partners and their family members (Poortman & Lyngstad, 2007; Teachman, 2008). Higher-order unions also involve individuals who have learned about the process of break up. Going through this often painful process may have caused people to be more cautious the next time (Furstenberg & Spanier, 1984), which may lead to less commitment to and fewer investments in the second union compared to the first. Furthermore, marriage market conditions have also changed because people are older when they search for a partner for the second time, and therefore the pool of potential partners is more restricted (Teachman, 2008). Thus, it is likely that the factors linked to the dissolution of second and higher-order unions are not the same as those linked to the disruption of the first union.

The life course perspective (Elder, 1998) suggests that partnership transitions are inter-related with other areas of life, such as parenthood. However, empirical evidence regarding the association between partnership dissolution and having children is somewhat mixed. Earlier research has found different, even opposite, effects of having children on partnership dissolution across countries and in different family situations with regard to, for example, the number, age, and residence of children (Coppola & Di Cesare, 2008; Lillard & Waite, 1993; Lyngstad & Jalovaara, 2010; Steele et al., 2005; Svarer & Verner, 2008).

1.2 Partnership transitions in context

A life course perspective suggests that decisions regarding life transitions are constrained by various contextual factors (e.g. Elder, 1998; Shanahan, 2000), as well as by the individual's development prior to the transitions (Räikkönen, Kokko, Chen, & Pulkkinen, 2012). Our study focused on the associations between partnership transitions and individual (i.e. gender and socio-emotional behaviour) and family characteristics.

Empirical studies have demonstrated that, in general, women undergo family-related transitions for the first time at a younger age than men (e.g. Elder, 1998; Kokko, Pulkkinen, & Mesiäinen, 2009; Räikkönen et al., 2012; Ross, Schoon, Martin, & Sacker, 2009). Furthermore, the timing of family transitions may also be more closely interlinked among women than among men (Kokko et al., 2009). It has been shown that early motherhood may weaken women's subsequent attachment to the labour market (e.g. Rönkä & Pulkkinen, 1998). No such association has been found among men (Rönkä, Kinnunen, & Pulkkinen, 2000).

To the best of our knowledge, the effects of childhood socio-emotional behaviour have not been studied in previous analyses of partnership formation and dissolution. However, indirect support for the links between childhood socio-emotional behaviour and adult partnership transitions can be found in previous research. First, there is evidence that child behavioural problems predisposes individuals to earlier parenthood (e.g. Kokko et al., 2009; Rönkä et al., 2000), especially among women (Kokko et al., 2009). In contrast, adaptive behaviour in childhood, such as shyness, has been shown to be related to later parenthood in men (Caspi, Elder, & Bem, 1988). Second, low self-control of emotions in childhood has been found to be a risk factor for later marital problems (Kinnunen & Pulkkinen, 2003). Third, there is evidence that high self-control of emotions in both genders, and social activity in women, contribute to favourable

adult development (Pulkkinen, 2009). On the basis of these earlier studies, we anticipated that high self-control of emotions would be connected to fewer and longer-lasting partnerships. Also, we expected that women with lower self-control of emotions and socially active men would form their first partnerships sooner.

An individual's family of origin may also influence union formation behaviours throughout adulthood. Accordingly, it has been shown that individuals who come from a less-advantaged family in terms of low socio-economic status (SES) tend to undergo their first partnership transition at an earlier age than individuals from a high SES background, for whom the later timing of transitions is more typical (e.g. Berrington & Diamond, 2000; Rönkä et al., 2000; Ross et al., 2009; Steele et al., 2006). Higher SES of the family of origin has also been linked to an increased risk of partnership dissolution (Bumpass, Martin, & Sweet, 1991; Lyngstad, 2006). In British cohorts, Steele et al. (2006) found that after a break-up, women from a higher SES background took longer to repartner, whereas Goldstein, Pan, and Bynner (2004) found no such effect among men. Family breakdown in childhood has been linked to earlier establishment of one's own partnership (Aassve, Burgess, Propper, & Dickson, 2006; Berrington & Diamond, 2000; Steele et al., 2006), as well as to a higher risk of partnership dissolution (Amato, 1996; Gähler, Hong, & Bernhardt, 2009; Steele et al., 2006), suggesting that union behaviours transfer at least to some extent from parents to their children.

Besides individual and family factors, the socio-historical context promotes variability in transition behaviours (e.g. Elder, 1998; Shanahan, 2000). The present study was based on longitudinal data collected for a representative sample of individuals born in Finland in 1959 (Pulkkinen, Lyyra, & Kokko, 2009; Pulkkinen & Kokko, 2010; Pulkkinen, 2009). Regarding partnership transitions in Finland, the mean age at first marriage was 25.9 years for women and 28.1 years for men in 1986–1990 (Statistics Finland, 2010). Cohabitation before marriage or as an alternative to marriage was very popular then, just as it is now (Statistics Finland, 1994). Among women born in 1938–42, 13% had cohabited, but among women born in 1958–62, 51% had cohabited before marriage and 33% as an alternative to marriage. Since the mid-1980s, the mean age at first marriage has risen: in 2009, the mean age was 30.2 years for women and 32.5 years for men (Statistics Finland, 2010). Most men and women marry only once; in 2009 11% of married women and 12% of married men had remarried. In 2009, the total divorce rate in Finland was 50% and the mean age at the time of divorce was 41.3 years for women and 43.8 for men. Of marriages entered into in 1985, 39% had ended in divorce by 2009. Due to the popularity of cohabitation in Finland, in this article our definition of a partnership includes both marital and non-marital cohabitational unions, which are treated as substitutes for each other.

2 Methods

2.1 Sample

We analysed data from the Finnish Jyväskylä Longitudinal Study of Personality and Social Development (JYLS). The study, established in 1968 by Lea Pulkkinen, includes all students from 12 randomly sampled second-grade school classes in Jyväskylä, Central Finland (Pulkkinen, 2009). All the pupils participated. The original sample consisted of 173 girls and 196 boys, of whom the majority (94%) were born in 1959. All participants were native Finns and they have been followed from age 8 to 50. During the follow-up, no systematic attrition has been found in the JYLS sample and the participants have continued to be representative of their Finnish birth cohort (Pulkkinen, 2009; Pulkkinen & Kokko, 2010).

During two data collection phases in 2001 at age 42 and in 2009 at age 50, life history calendars (LHC; adapted from Caspi, Moffitt, Thornton, Freedman, & others, 1996; Kokko, Pulkkinen, & Mesiäinen, 2009) were used to retrospectively collect information about partnership status, children, education and work, as well as other important life events. The occurrence, timing and duration of the transitions were recorded annually first from age 15 to age 42, and later from age 42 to age 50, during interviews in which altogether 275 participants (77% of the original sample still alive at age 50) gave reports based on their memory and visual aids provided by the LHC-sheet.

The information collected with the LHCs was confirmed and complemented using other sources, such as life situation questionnaires and interviews at ages 27, 36, 42, and 50. We were able to derive almost complete partnership data between ages 15–42, but missing information due to non-response during the last phase of data collection at age 50 led to incomplete histories for 22% of the

participants. The length of the follow-up varies between individuals because of the two data collection phases and small differences in their ages. Altogether 215 participants were followed for 36 years, 14 participants for 35 years, and 46 participants for only 28 years.

2.2 Variables

In addition to subjects' annual partnership histories, we used information from their parenthood histories to derive a time-varying binary indicator of whether or not the individual was a *parent* to biological or adopted children in a given year.

Socio-economic status (SES) based on father's occupation (or mother's if she was the sole provider or had a higher status), was coded 0 if blue-collar and 1 if a white-collar worker (Pitkänen, Lyyra, & Pulkkinen, 2005).

Family structure at age 14 was coded 0 if the participant lived with both parents and 1 if the parents had divorced or a parent had died (Kokko & Pulkkinen, 2000).

Child-centred parenting was an average score of five dichotomous variables based on age 27 recollections of parenting practices and home environment (parental relationship, physical punishment, maternal supervision, relationship with the father, and *family structure*; Kokko & Pulkkinen, 2000). Missing data were imputed (Pitkänen, Kokko, Lyyra, & Pulkkinen, 2008).

Child socio-emotional behaviour at age 8 was assessed using two subscales: *social activity* and *high self-control of emotions* (including emotional stability, constructiveness, and compliance; see Kokko, Pulkkinen, Mesiäinen, & Lyyra, 2008; Pulkkinen, Kokko, & Rantanen, 2012). Each item was rated by teachers on a scale from 0 (never) to 3 (often).

2.3 Statistical methods

Sequence analysis (SA) is a model-free data-mining type of approach that provides an overview of individual sequences over the whole observation period, including the most common transitions and time spent in each partnership state. The aim of SA is to measure pairwise (dis)similarity of the sequences, which is often followed by some kind of clustering method to find typologies of whole trajectories. *Event history analysis* (EHA; also known as survival, duration, or failure-time analysis) is used for the study of factors that influence the timing of transitions. The response variable in EHA is the duration between becoming at risk of experiencing the event of interest and the time that the event occurs.

2.3.1 Sequence analysis

SA was originally developed in bio-informatics to organize, classify, and parse protein and DNA sequence data (Durbin, Eddy, Krogh, & Mitchison, 1998). In the social sciences, Abbott introduced the use of SA in life course analysis in the mid-1980s (Abbott, 1983; Abbott, 1995; Abbott & Tsay, 2000). The basic idea in SA is to measure the distance or dissimilarity of two sequences consisting of the succession of categorical states describing the trajectories. Two major issues are essential for SA. The first concerns the composition of sequences: how many and what type of states? The second issue is related to determining the dissimilarities between the sequences: which dissimilarity measure to use and, for some measures, how to assign the "cost" of converting one state to another? Typical steps in SA include the following: 1) creating sequences using a finite set of states; 2) choosing and implementing a method for computing pairwise dissimilarities between sequences; 3) analysing the dissimilarities (e.g. cluster analysis and/or multi-dimensional scaling); 4) graphical illustration and examination of sequence data.

Definition of states

Technically, the number of states does not have to be restricted (though finite), but for practical and interpretational reasons the state space is often relatively limited. Definition of the states requires careful consideration. In the present application, for example, defining divorced as single, or distinguishing partnership states by the type of union instead of order, would give a different viewpoint. In previous research it has been common to group all co-residential partnerships together as one state (e.g. Aassve, Billari, & Piccarreta, 2007; Gauthier, Widmer, Bucher, & Notredame, 2010; Salmela-Aro, Kiuru, Nurmi, & Eerola, 2011) or to separate marriages from cohabitations (e.g. Barban & Billari, 2012; Elzinga & Liefbroer, 2007; Piccarreta & Lior, 2010). Usually these have been combined with information on children.

We coded annual partnership states for each individual based on the *order* of the partner: 1) living single (never had a co-residential partner), 2)

living with the first partner, 3) with the second partner, 4) with at least the third partner, or 5) living divorced/separated/widowed. Widowhood was very rare and thus it was merged with the other states of living without a previous partner. Transitions between the states were more restricted than in most studies of partnership sequences: only the last two could be revisited, except for the rare event of going back to a previous partner. Without separating partnerships by order it would have been difficult or even impossible to distinguish sequential partnerships.

Dissimilarities of sequences

There are several methods for measuring sequence dissimilarity, optimal matching (OM) being the most well-known (e.g. McVicar and Anyadike-Danes, 2002). In OM the goal is to find the best alignment of two sequences. Their dissimilarity is computed from the operations needed to transform one sequence into the other using insertions, deletions, and substitutions of states. Roughly, the more operations needed, the more distant the sequences are. The operations can be given different costs to reflect the amount of dissimilarity between the states. Another completely different type of approach by Elzinga is based on counting or measuring common sequence attributes such as sub-sequences (Elzinga, 2006; Elzinga & Liefbroer, 2007). These methods do not require definition of any costs.

In the present study, we use generalized Hamming distance (Hamming, 1950; Lesnard, 2010) which compares states at the same time positions in each sequence. This performs well in our data where the observed sequence lengths vary across individuals, and where the timing of the partnership transitions is regarded as very important. To assess the closeness of two partnership histories, sequences are aligned year by year (see Example 1). Shorter sequences are complemented with missing states to achieve equal sequence lengths required to compute Hamming distances. Partnership states at each age are compared and each comparison is given a cost (see Table 1). Only the ratio of the costs is important and usually the absolute numbers have no substantive meaning; multiplying the costs by a constant does not change the results. The dissimilarity of the histories is simply the sum of the costs.

Example 1

Computing generalized Hamming distances between artificial partnership histories. The costs are given for a comparison of partnership states at each age. See Table 1 for definition of states and costs.

Age	20	21	22	23	24	25	26	27	28
Sequence 1	S	S	S	P1	P1	P1	P1	P1	P1
Sequence 2	S	S	S	S	S	S	P1	P1	*
Cost	0	0	0	2	2	2	0	0	0

Dissimilarity = 6

Age	20	21	22	23	24	25	26	27	28
Sequence 1	S	S	S	P1	P1	P1	P1	P1	P1
Sequence 3	P1	P1	P1	P1	P1	D	P2	P3	P3
Cost	2	2	2	0	0	2	2	3	3

Dissimilarity = 16

Definition of the costs depends not only on the states themselves but also on the research question of interest: which states are regarded as close and which as distant? The most common strategies have been to assign the costs based on theory or transition probabilities between the states. The latter way is automatic and has been said to reduce subjectivity (Aisenbrey & Fasang, 2010; Gauthier, Widmer, Bucher, & Notredame, 2009). However, it is not suitable for many cases such as the present study, where most of the partnership transitions are impossible and the probabilities of the transitions provide little information on the dissimilarities between the states. Setting the costs is an ongoing debate and many modifications to the basic options have been suggested (e.g. Aisenbrey & Fasang, 2010; Gauthier et al., 2009; Halpin, 2010; Hollister, 2009; Lesnard, 2010).

Table 1. Costs for Hamming distance computations

		Sequence 2					
		S	P1	P2	P3	D	*
Sequence 1	Single (S)	0	2	3	5	5	0
	1st partnership (P1)	2	0	1	3	2	0
	2nd partnership (P2)	3	1	0	2	2	0
	3rd+ partnership (P3)	5	3	2	0	2	0
	Divorced/separated (D)	5	2	2	2	0	0
	Missing (*)	0	0	0	0	0	0

Note. Costs were defined to measure how distant different partnership states are regarded

We set costs that would lead to clusters that separate histories of stable and unstable partnerships from those with long periods of living single or divorced/separated. The last two were seen as distant states (cost = 5) because forming a partnership was regarded as one step in the developmental process to adulthood. Second partnerships were very common, so the cost of alignment with the first partnership state was set low (cost = 1). Aligning any state to a missing state was defined to have zero cost to ensure that sequences were grouped together according to the known parts of the histories, not with other sequences with missing information.

For the JYLS data, other dissimilarity measures including optimal matching, dynamic Hamming distance (Lesnard, 2010), the length of the longest common subsequence, and the number of common subsequences were considered together with different cost definitions. Generalized Hamming with the costs presented in Table 1 gave the most meaningful clusters and the best goodness-of-fit, as measured by the proportion of the variation explained by the clusters (pseudo coefficient of determination).

Clustering sequences

The dissimilarities between all partnership sequences are collected in a matrix that can be used to cluster similar histories together. We used Ward's agglomerative algorithm (Ward Jr., 1963). At each step, the algorithm combines the two clusters (at the first step, sequences) that minimize within-cluster variability and maximize inter-cluster variability. It is commonly used to cluster sequences since it usually produces more equal-sized clusters than other algorithms (Aisenbrey & Fasang, 2010). We also tested other clustering options but, as also found by Aassve et al. (2007), most of them (single, average, and complete linkage) resulted in one large cluster and many residual clusters with only a handful of sequences, even several clusters with only one sequence. This is not desirable for the purpose of interpretation and possible further analyses. With our dissimilarities, the "partition around medoids" method (PAM) (Kaufman & Rousseeuw, 2009) was the best competitor, but not as good as Ward in terms of pseudo-R^2 (for pseudo-R^2 see Studer, Ritschard, Gabadinho, & Müller, 2011). Choosing the best

number of clusters is not straightforward. Our decision was based on the dendrogram, interpretability of the clusters, and change in measures including pseudo-R^2, pseudo F (Studer et al., 2011), Hubert's C, and Hubert's Gamma (Hubert & Arabie, 1985). See Studer (2013) for a review of measuring the quality of clustering of sequence data.

External information can be taken into account after clustering or at the clustering phase. We used regression trees (Breiman, Friedman, Olshen, & Stone, 1984) to group similar partnership histories using information on subjects' home background and socio-emotional behaviour in childhood as predictors. The idea of regression trees is to recursively partition data into clusters using values of a predictor, creating binary splits for the values of a variable for which the highest pseudo-R^2 is achieved. The tree is grown until no further significant splits (assessed through a permutation F-test) are found (Studer et al., 2011).

We studied whether sex and socio-emotional characteristics and home background during childhood predicted future partnership histories using regression tree methods with the same Hamming distances as previously.

Graphical illustrations

There are many options for graphical description of sequence data. The most common choices include cross-sectional state distribution plots and sequence index plots. State distributions plotted for each time point show the change in the prevalence of states in the course of time. Sequence index plots show the whole partnership histories for the individuals. Plotting all sequences at once in a random order is usually not very informative. Clustering eases interpretation by grouping similar histories together, and multi-dimensional scaling or some other criterion is often used to order sequences more meaningfully.

Software

The TraMineR package in R (Gabadinho, Ritschard, Müller, & Studer, 2011) was used for the SA presented in this paper. Alternatives include TDA (Rohwer & Pötter, 2004) and the Stata packages SQ (Brzinsky-Fay, Kohler, & Luniak, 2006) and SADI (Halpin, 2014). To our knowledge, TraMineR has been the most versatile and widely used software for SA in recent years. However, the new SADI package in Stata appears to have the potential to become a strong competitor.

2.3.2 Discrete-time event history model

SA is a useful tool for obtaining an overview of histories. However, as the focus is the whole trajectory, SA cannot be used to study how the factors of interest – especially those which vary over time – are related to the timing and duration of each co-residential partnership. EHA is a highly flexible approach for the study of how individual time-invariant and time-varying characteristics influence the timing of partnership transitions.

Moving in with the first partner is a milestone for an individual, but it may not be the only partnership (marriage or cohabitation) that is established during their life time. Instead of focusing only on the timing of the first partnership we can analyse the duration of all episodes of living without a partner. These are periods during which an individual is continuously "at risk" of establishing a new partnership. Individuals not living with a partner in a given time interval constitute what is referred to as the "risk set" for partnership formation. An individual's first episode starts at the beginning of the follow-up and it ends when the individual moves in with a partner for the first time or is censored because of loss to follow-up. Individuals stay out of the risk set as long as they are living with the same partner. A new episode begins at dissolution when the individual is again "at risk" of forming a new partnership.

The durations of episodes from the same individual are likely to be correlated, which invalidates the independence assumption of standard statistical methods. This correlation is due to unmeasured time-invariant individual characteristics that affect the risk of forming any (new) partnership. The variation in the risks between individuals is generally called unobserved heterogeneity or individual frailty (e.g. Vaupel, Manton, & Stallard, 1979). Recurrent events data can be viewed as having a two-level hierarchical structure where the events are nested within individuals. These types of hierarchical data can be analysed with multilevel or random effects models (e.g. Goldstein, 2011; Raudenbush & Bryk, 2002).

Many life transitions, such as partnerships, are formed in continuous time, but it is not always possible or practical to collect data as such. Often, event times are recorded in time intervals such as months or years because finer measurement (e.g. daily accuracy in a study spanning several years) would not be informative. At other times it is not possible to observe the occurrence times as frequently as would be preferred. In both cases the discrete-time model can be used as an approxi-

mation to a continuous-time model (e.g. Allison, 1982).

The two LHCs from the JYLS study contain yearly information on individuals' partnership statuses. We were interested in both the formation and dissolution of partnerships. However, annual accuracy was not always frequent enough to distinguish between consecutive partnerships. To properly define who was in the risk set of moving in with a new partner (i.e. living without a partner) at the start of a given time interval, artificial six-month intervals were created and the partnership status of the latter part of the year changed to "single" for those who had dissolved and formed a partnership during the same year (29 cases from 24 individuals).

Random effects model for repeated partnership formation

In our annual data, a partnership beginning "at age t" occurs during the one-year interval $[t, t+1)$. Suppose that t_{ij} is the number of years for which individual j is observed in episode i, where an episode is a continuous period of time unpartnered. We form a data set with one record per year for each individual (a person-episode-period file) and define a binary indicator y_{tij} for each year $t = 1, \ldots, t_{ij}$ such that

$$y_{tij} = \begin{cases} 1 & \text{if episode } i \text{ of an individual } j \\ & \text{ends in partnership formation at } t \\ 0 & \text{otherwise} \end{cases}$$

The discrete-time hazard function is defined as

$$p_{tij} = P\left(y_{tij} = 1 \middle| y_{t'ij} = 0 \text{ for } t' < t\right),$$

which is the conditional probability that a partnership is formed during interval t of episode i of individual j given that they have not moved in with a partner before interval t.

A logistic regression model is commonly used to model the dependence of p_{tij} on the duration unpartnered by interval t and a vector of (possibly time-varying) explanatory variables $\boldsymbol{x}_{tij}$:

$$\log\left(\frac{p_{tij}}{1-p_{tij}}\right) = \alpha' \boldsymbol{z}_{tij} + \beta' \boldsymbol{x}_{tij} + u_j,$$

where $\boldsymbol{z}_{tij}$ is a vector of functions of t and $\alpha' \boldsymbol{z}_{tij}$ defines the baseline hazard function. Polynomials and step functions are common choices for modelling the time-dependency. Unobserved variation between individuals (frailty) is represented by u_j, which is usually assumed to follow a normal distribution $N(0, \sigma_u^2)$. The random effect shifts the log-odds of partnering up or down for the individual j while the effects of duration and covariates are assumed to be constant across individuals. Conditional on u_j, the durations of episodes for the same individual are assumed to be independent. A similar model is specified for the risk of partnership dissolution.

A two-state model

We can extend the above model to study transitions between two (or more) states. That model considers transitions from a single state to living with a partner and the individual is dropped from observation after forming a partnership (unless they separate and re-enter the risk set). In a two-state model the durations of all episodes living with and without a partner are examined. Exit from one state implies entry to the other. Examples of the use of multistate models to study partnership transitions include Aassve et al. (2006), Goldstein et al. (2004), and Steele et al. (2006).

We denote by S_{tij} the state of individual j's ith episode at the start of interval t. Now y_{tij} is the binary indicator of a transition of either type, forming (F) or dissolving (D) a partnership. The conditional probability of a transition from state s $(s = F, D)$, during interval t, given that a transition has not yet occurred in that episode, is now

$$p_{stij} = P\left(y_{tij} = 1 \middle| y_{t'ij} = 0 \text{ for } t' < t, S_{tij} = s\right),$$

and the multilevel event history model for transitions between the two states can be written as

$$\text{logit}\left(p_{stij}\right) = \alpha_s' \boldsymbol{z}_{stij} + \beta_s' \boldsymbol{x}_{stij} + u_{sj}, \quad s = F, D$$

Note that the baseline logit-hazard, covariates, coefficients, and random effects can all vary across states, as indicated by the s subscripts.

Software

Random effects models for recurrent events and multiple states can be fitted in most mainstream statistical software packages such as R, SAS and Stata, and also with more specialist software including MLwiN and Sabre. The packages may vary in the estimation procedures used, leading to differences in parameter estimates and computational times (see Steele (2011) for a detailed summary). In our study, event history models were fitted using the xtlogit procedure in Stata which implements maximum likelihood via Gauss–Hermite quadrature.

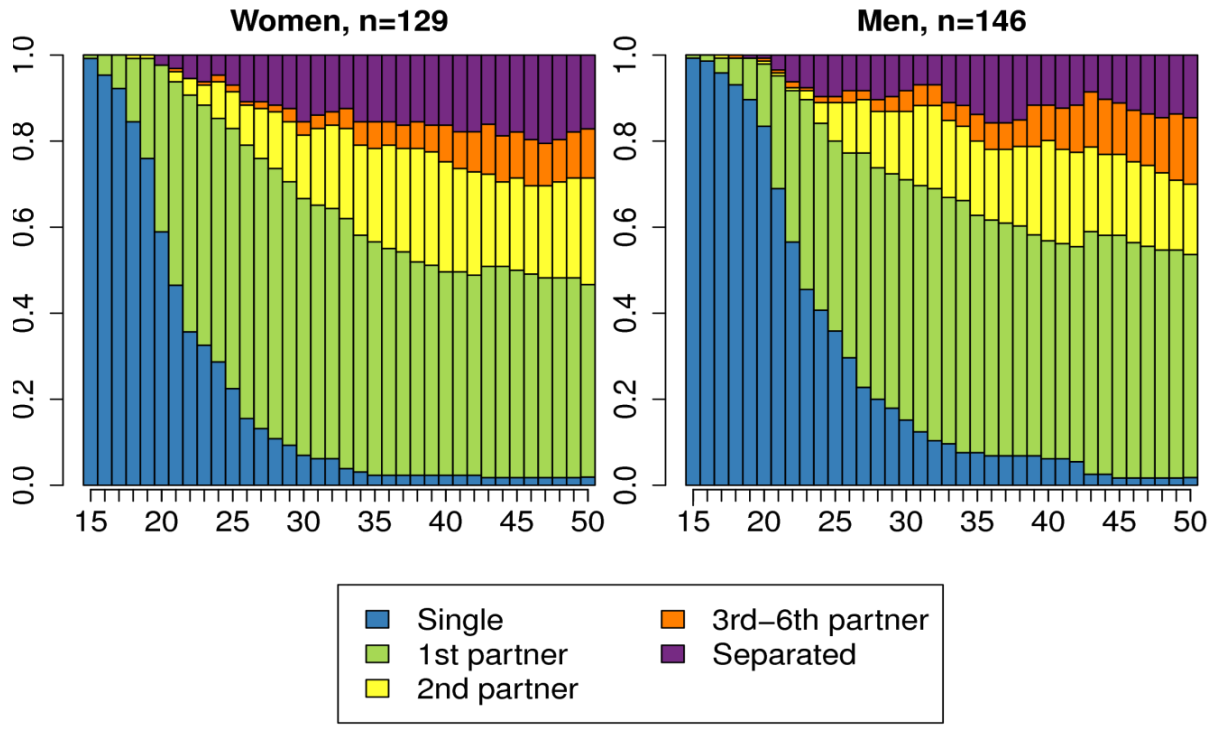
3 Results

3.1 Sequence analysis: trajectories of partnerships

Sequence analysis was used to provide an overall view of partnership histories, to obtain descriptive information on typical and atypical trajectories, and to explore how much childhood socio-emotional characteristics and family background predict future histories.

Figure 1 presents the prevalence of partnership states at each age for women and men. On average, men formed their first partnership later than women. Women spent more time living as divorced or separated than men, but from this figure we cannot see the duration of these periods.

Figure 1: State distribution plots of partnership histories for women and men between ages 15–50 in JYLS data



Notes. Missing states are not included in the yearly proportions. The change in proportions at 43 is due to individuals who were lost to follow-up.

Table 2 shows the average number of years that women and men spent in each partnership state. Women had longer first and second partnerships than men, but there was a lot of variation.

Table 2. Mean and standard deviation of years spent in each partnership state since age 15 for women and men in the JYLS data

	Women		Men	
State	Mean	S.D.	Mean	S.D.
Single	7.8	5.7	10.1	6.7
1st partnership	16.3	11.1	14.9	10.7
2nd partnership	5.2	1.6	4.3	7.4
3rd–6th partnership	1.6	5.0	1.9	5.1
Divorced/separated	4.0	5.8	3.0	5.2
Missing	1.1	2.7	1.8	3.6

Table 3 shows the most frequent types of history ignoring the time spent in each state. Two out of three individuals had settled in to their first or at most second partnership. Since the transitions between states are rather limited due to several being absorbing, there are few possible histories. Except for the differences in the number of partners and dissolutions, the histories only differ by whether or not the individuals had lived alone between their partnerships. Taking account of the durations of episodes adds little additional information: the number of the JYLS participants is limited compared to the length of the follow-up so most of the sequences are unique.

Table 3. The most common partnership histories in JYLS data, when durations are omitted

State	Freq.	%
S-P1	122	44.4
S-P1-D-P2	59	21.5
S-P1-D	25	9.1
S-P1-D-P2-D-P3	14	5.1
S-P1-D-P2-D	10	3.6
S	9	3.3
Total	239	86.9

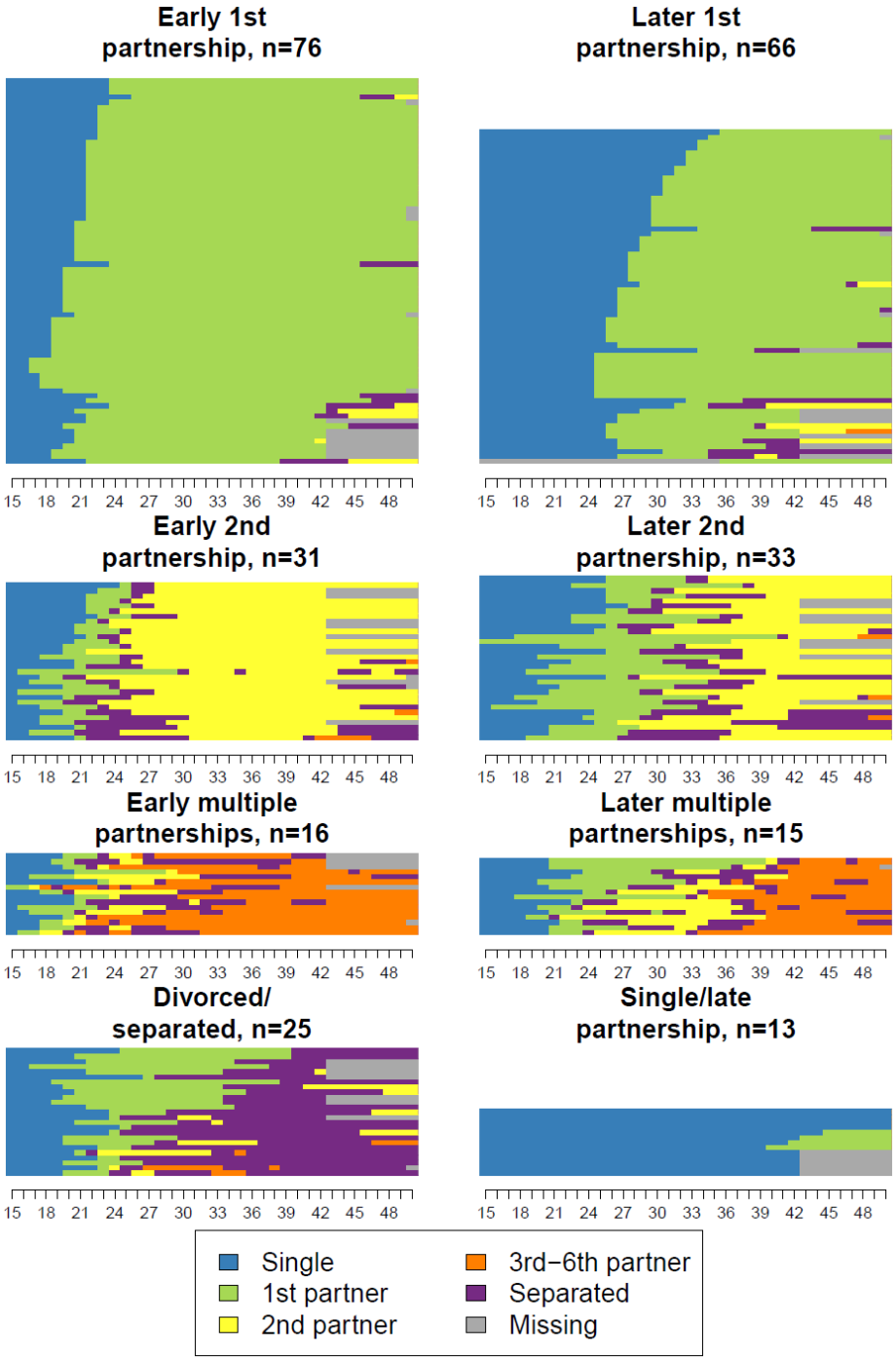
Notes. S=single, P1=1st partnership, P2=2nd partnership, P3=3rd–6th partnership, D=Divorced/separated/widowed.

3.1.1 Clustering sequences

Solutions with between 2 and 15 clusters from Ward's algorithm were studied, and the eight-cluster solution was chosen based on the criteria described in Section 2.3.1. These clusters explained 61% of the variation between the histories. Sequence index plots of the clusters are shown in Figure 2.

There were four larger clusters of relatively stable partnership histories with one or two partners that only differ in timing. Men were in the majority among those who have established a (typically long-lasting) late initial partnership, but in the "later second partnership" group the majority were women (Table 4). There emerged also two male-dominated clusters which included individuals with multiple partnerships, either earlier or later in life. Some of these individuals had experienced multiple partnerships but settled down after early adulthood, and others had not formed long-lasting partnerships at all. The last two clusters showed histories of living without a partner; some (typically women) had a partnership that ended in separation or divorce, while others (typically men) had never lived with a partner or had entered their first partnership very late.

Figure 2: Eight clusters of partnership histories using generalized Hamming distances as a measure of dissimilarity and Ward's method for clustering



Notes. Multidimensional scaling was used to order sequences.

Table 4: Proportion of partnership clusters and the percentage of women

Cluster	Size (*n*)	Size (%)	Women (%)
Early 1st partnership	76	27.6	55.3
Later 1st partnership	66	24.0	34.8
Early 2nd partnership	31	11.3	45.2
Later 2nd partnership	33	12.0	60.6
Early multiple partnerships	16	5.8	43.8
Later multiple partnerships	15	5.5	33.3
Divorced/separated	25	9.1	60.0
Single/late partnership	13	4.7	23.1
Total	275	100	46.5

3.1.2 Clustering with external information

Using the regression tree method described in Section 2.3.1, only two of the covariates were statistically significant predictors of cluster membership; these formed altogether three clusters of the data (Figure 3).

The first and the most effective split of the data was achieved with child-centred parenting (CCP). More child-centred parenting practices in the family of origin (CCP $>$ 0.4) was related to more stable partnership histories with usually one or two partners. The second split was for the lower values of CCP and self-control of emotions (SCE). On average, individuals with lower values of CCP and SCE had more partners compared to those who also had lower values of CCP but higher SCE. Altogether grouping on CCP and SCE explained only 3.5% of the variability between the partnership histories, so most important sources of sequence variation was the timing and the number of partnerships.

3.2 Event history analysis: transitions to and from partnerships

Event history analysis was used to examine the timing of partnership formation and dissolution and how the rate of partnership transitions depends on individual history and characteristics.

As can be seen from the partnership clusters in the previous section and again in Table 5, recurrent partnerships were common: almost a half of both women and men had established at least two partnerships (marriages or cohabitations) during the follow-up period. Third and subsequent partnerships were less common, especially among women.

Figure 3. Regression tree of partnership histories with two significant splitting variables: child-centred parenting (CCP, scores 0–1) and high self-control of emotions (SCE, scores 0–3)

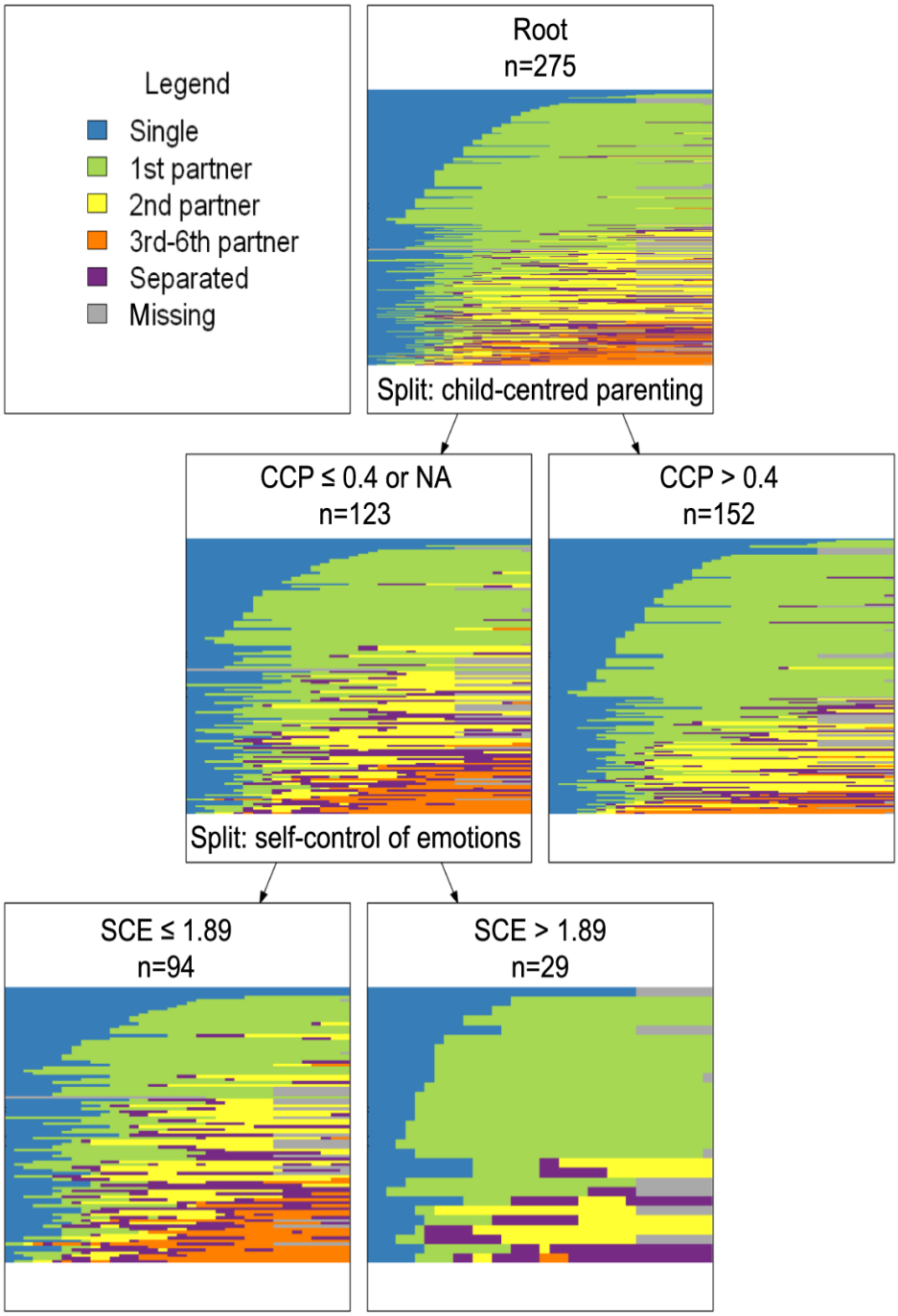


Table 5. Participants in the JYLS study by sex and the number of cohabitating partnerships

				3+ partners	
	No partners	1 partner	2 partners	Individuals	Partnerships
Women	3	66	43	17	25
Men	6	79	33	27	42

Notes. Higher-order partnerships (3th–6th) are combined into one category due to their small number.

Table 6 shows the means of the age at forming partnerships, duration of partnerships and time before forming new partnerships (not accounting for right-censoring). On average, first partnerships were formed around age 22 among women and age 24 among men. The youngest formed their first partnership (cohabitation) at 15 and the oldest at 35 (women) and 45 (men). On average, a new partnership was formed 2–3 years after dissolution of the previous partnership but there was considerable variation, with a maximum duration of over 20 years.

The average duration of first partnerships that ended in dissolution during the follow-up was about 8 years. Second partnerships were of a similar length to first partnerships for women and two years shorter among men. Higher-order partnerships lasted 4–5 years on average.

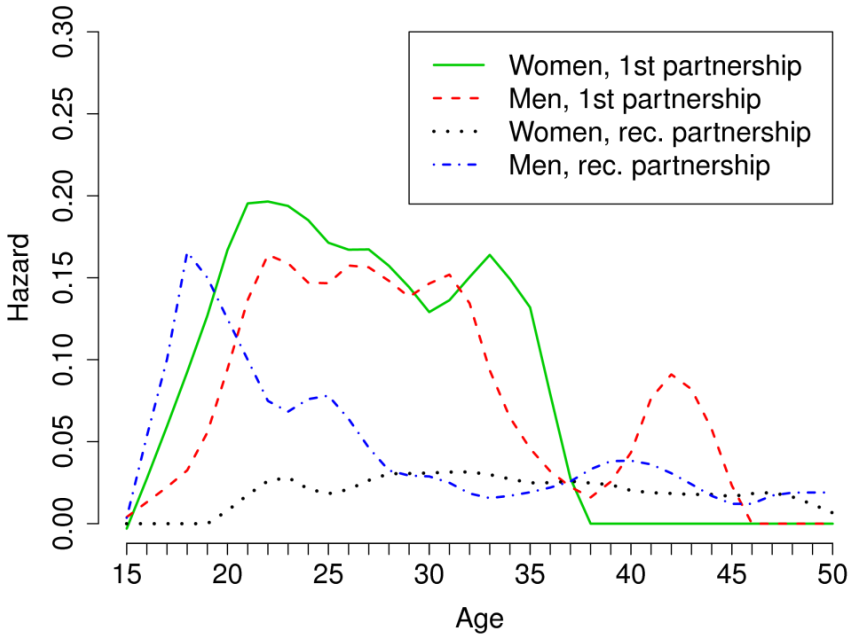
Table 6. Timing of partnership events: mean ages at forming partnerships, years since dissolution before forming a new partnership, and duration of partnerships that had ended in separation in the JYLS data

		Formation					Dissolution		
		Age		Time since diss.			Duration		
Sex	Partner	Mean	S.D.	Mean	S.D.	N	Mean	S.D.	N
Female	1st	22.17	4.16			126	8.54	6.51	68
	2nd	32.07	7.91	3.45	3.36	60	8.20	6.13	25
	3rd+	36.17	7.95	2.72	2.42	25	4.38	3.13	13
Male	1st	24.30	5.20			139	8.14	7.41	74
	2nd	31.22	7.53	2.68	3.19	59	5.97	6.30	31
	3rd+	36.56	9.04	2.39	2.74	42	4.92	4.30	18

Note. Right censoring was not accounted for.

Hazards of forming first and recurrent partnerships were computed from the data. The hazard at a given age is the proportion who were newly partnered from all individuals in the risk set (those who were not living with a partner yet/anymore). The hazard function is plotted in Figure 4 using locally weighted scatterplot smoothing (lowess) to show the change in the rate of partnership formation by age. We also see that on average women formed their first partnerships earlier than men. On the other hand, those men who *had* established and dissolved their first partnerships young (before age 25) seemed to form subsequent partnerships quicker than young women in the same situation. There was an especially high peak for teenagers, but the risk set at that age was very small. In this study, the oldest age at first partnership was 35 for women, but is some suggestion that for men the hazard of first partnership increased in their early 40s (although, again, the risk set is small).

Figure 4. Hazard functions of the formation of first and recurrent partnerships for women and men



Note. Hazards were smoothed with lowess (locally weighted scatterplot smoothing) using 20–25% of the closest points.

3.2.1 Partnership formation

Since preliminary analyses (not all shown here) revealed large differences between women and men in the timing of partnership formation and dissolution and in the factors related to these transitions, separate event history models were fitted for women and men. Based on the hazard functions shown in Figure 4, a piecewise constant function was chosen as the best representation of the baseline hazard for partnership formation. The timing of first partnership was categorized into three periods: early (15–22 years), on-time (23–32), and late (33–50). The last category is wider than would be preferred, since it is unlikely that, for example, a 33-year-old and a 50-year-old have a same risk for establishing especially the first partnerships. However, as no women in our sample established their first partnership after age 35 it was not possible to use narrower age categories. Time since, and the duration of the last partnership were also considered (using linear, quadratic, logarithmic, categorical functions of time) as well as the type of the previous partnership (marriage/cohabitation), but these variables did not show significant effects for either sex and were excluded from the models. Covariates measured in childhood were treated as time-invariant, while parenthood status and

existence of previous partners were time-dependent.

We first studied the main effects of the covariates and their interactions with age and a previous partnership indicator. Interactions with age were considered to test the proportional hazards assumption, while interactions with previous partnership were tested to determine whether covariate effects differ for first and recurrent partnerships. Variables with effects that were significant at the 5% level were then tested together in one model, with non-significant effects dropped one by one. None of the interactions between age and any covariate were significant.

Tables 7 and 8 show the final random effects models for partnership formation for women and men respectively. There was little evidence of unobserved heterogeneity among women (σ_u was estimated close to 0), but among men the additional of random effects led to a significant improvement in fit ($\hat{\sigma_u} = 0.607$, significance assessed through likelihood ratio test). The "risk" of forming an initial partnership was estimated to be the highest among 23–32 year-olds for both sexes, but the differences between the age categories were small and not statistically significant at the 5% level. Among men and women who had already dissolved at least one partnership, the risk of repartnering was significantly higher among 15–22 year-olds than for the other age groups.

Table 7. Logistic model of partnership formation for women

	Est.	s.e.	p	OR	OR 95% CI
Constant	−3.579	0.541	0.000		
Had previous partner(s)	1.841	0.670	0.006	6.302	(1.697,23.410)
Age 15–22	0.550	0.500	0.272	1.733	(0.650,4.620)
Age 23–32	0.975	0.507	0.054	2.651	(0.982,7.157)
Prev. partners * Age 15–22	1.883	0.716	0.009	6.571	(1.615,26.738)
Prev. partners * Age 23–32	−0.116	0.566	0.838	0.891	(0.294,2.702)
Has child(ren)	1.232	0.312	0.000	3.429	(1.861,6.318)
Prev. partners * Has child(ren)	−0.935	0.411	0.023	0.393	(0.175,0.879)
High SCE	0.025	0.138	0.856	1.025	(0.782,1.344)
Prev. partners * High SCE	−0.737	0.232	0.001	0.479	(0.304,0.754)
Higher SES	−0.058	0.208	0.782	0.944	(0.628,1.419)
Prev. partners * Higher SES	−0.889	0.394	0.024	0.411	(0.190,0.889)
Random effect SD σ_u	0.001	0.012			

Notes. Estimated coefficients and odds ratios (OR) are shown together with standard errors, p-values and 95% confidence intervals (CI) for the odds ratios. The last age category (33–50) was chosen as the reference category. SCE = self-control of emotions (scores 0–3), SES = socio-economic status based on the parents' (mainly fathers') occupational status during the subject's childhood (higher/lower).

Table 8. Logistic model of partnership formation for men

	Est.	s.e.	p	OR	OR 95% CI
Constant	−3.410	0.480	0.000		
Had previous partner(s)	0.127	0.499	0.799	1.136	(0.427,3.023)
Age 15–22	−0.795	0.451	0.078	0.451	(0.186,1.093)
Age 23–32	0.334	0.409	0.414	1.396	(0.627,3.109)
Prev. partners * Age 15–22	2.763	0.731	0.000	15.855	(3.787,66.373)
Prev. partners * Age 23–32	0.580	0.490	0.237	1.785	(0.683,4.668)
Has child(ren)	2.849	0.370	0.000	17.275	(8.372,35.643)
Prev. partners * Has child(ren)	−2.302	0.469	0.000	0.100	(0.040,0.251)
Social activity	0.251	0.137	0.067	1.285	(0.982,1.682)
Random effect SD σ_u	0.607	0.127			

Notes. Estimated coefficients and odds ratios (OR) are shown together with standard errors, p-values and 95% confidence intervals (CI) for odds ratios. The last age category (33–50) was chosen as the reference category.

Altogether three childhood factors were associated with partnership formation: socio-economic status (SES, Table 7), self-control of emotions (SCE, Table 7), and social activity (Table 8). Being from a higher SES family background was associated with a longer time to repartner for women. High self-control of emotions that was found to predict cluster membership in the regression tree analysis of SA, was also a predictor in the event history analysis of partnership formation: women who had higher self-control of emotions at age 8 had a lower risk of forming a new partnership following a dissolution. The effect of social activity was significant at the 10% level for men: being more socially active at age 8 was associated with forming partnerships sooner. The effect was the same for first and recurrent partnerships.

Parents were faster at forming first partnerships, although only ten participants had a child before forming any co-residential partnerships. There was some evidence that fathers also formed recurrent partnerships faster compared to childless men ($\hat{\beta} = 2.849 - 2.302 = 0.548$, s.e. $= 0.300$, p-value $= 0.068$).

Child-centred parenting, which was found to be the most important covariate in the regression tree, was not a significant predictor of partnership formation for either sex after controlling for the effects of other covariates. Childhood family structure was not significant in either model after controlling for the other childhood variables.

3.2.2 Partnership dissolution

Partnership dissolutions were explored in a similar way to formations. Time was captured in the models by two different variables: the age at the start of the current partnership and the duration of the partnership. Different functional forms (linear, quadratic, logarithmic, and categorical) were studied for both variables. Covariates measured during childhood were treated as time-invariant; type of partnership (marriage/cohabitation), parenthood status, and existence of previous partners as time-dependent. Child-centred parenting and family structure (included in CCP) were correlated, which induced multicollinearity in the model for women. Both variables were considered important and included irrespective of the large standard error of CCP in the common model.

Tables 9 and 10 show the results from the event history models of partnership dissolutions for women and men respectively. The random effect

standard deviations were large but non-significant. The age effect was linear and decreasing for women. For men, the estimated effects of age and age squared formed a quadratic curve: the risk decreased until 42 years of age and then slightly increased (the age at which the hazard reached its minimum was found by taking the square root of the first derivative of the quadratic function). For men, the effect of the duration of the current partnership was linear and decreasing. For women, the risk of partnership dissolution was quadratic, increasing until 12 years into the partnership and then decreasing.

Table 9. Logistic model of partnership dissolution for women

	Est.	s.e.	p	OR	OR 95% CI
Constant	−1.589	0.594	0.007		
Age at partnership formation	−0.055	0.018	0.003	0.946	(0.913,0.982)
Partnership duration	0.095	0.055	0.086	1.100	(0.987,1.225)
(Partnership duration)^2	−0.004	0.002	0.046	0.996	(0.991,1.000)
Married	−1.109	0.249	0.000	0.330	(0.204,0.534)
High self-control of emotions	−0.397	0.173	0.022	0.672	(0.479,0.944)
Broken family at 14	0.532	0.248	0.032	1.702	(1.048,2.766)
Child-centred parenting	−0.636	0.476	0.182	0.529	(0.208,1.347)
Random effect SD σ_u	0.518	0.254			

Notes. Estimated coefficients and odds ratios (OR) are shown together with standard errors, p-values and 95% confidence intervals (CI) for the odds ratios.

Table 10. Logistic model of partnership dissolution for men

	Est.	s.e.	p	OR	OR 95% CI
Constant	1.211	1.495	0.418		
Age at partnership formation	−0.254	0.105	0.019	0.782	(0.637,0.961)
(Age at partnership formation)^2	0.003	0.002	0.055	1.003	(1.000,1.007)
Partnership duration	−0.040	0.019	0.032	0.961	(0.926,0.997)
Broken partnership(s)	0.757	0.272	0.005	2.132	(1.252,3.630)
Has child(ren)	−0.701	0.228	0.002	0.496	(0.317,0.776)
High self-control of emotions	−0.443	0.158	0.005	0.642	(0.471,0.875)
Random effect SD σ_u	0.385	0.247			

Notes. Estimated coefficients and odds ratios (OR) are shown together with standard errors, p-values and 95% confidence intervals (CI) for odds ratios.

Previous experience of dissolution increased the risk of subsequent separation or divorce among men but not among women. Married women were less likely to dissolve their partnerships compared to cohabiting women, but cohabiting and married men did not differ in their risk of dissolution. Motherhood did not change the risk of dissolution but fathers had a lower risk than men without children.

Three childhood characteristics were connected to the risk of dissolution: self-control of emotions, family disruption, and child-centred parenting. High self-control of emotions at age 8 decreased the risk of dissolution for both sexes and all partnerships, while child-centred parenting was associated with a lower risk of dissolution for women. The experience of a broken family during childhood was associated with a higher risk of partnership dissolution among women, but not men.

4 Summary and discussion

This paper had two aims: (i) to describe the use of complementary statistical methods, sequence analysis and event history analysis, in a study of recurrent events; and (ii) to apply both techniques in a study of partnership formation and dissolution over the life course.

4.1 Statistical analysis

Sequence analysis was used to build an overall picture of partnership histories from age 15 to 50. Using Ward's clustering method, eight clusters were found, which together explained over 60% of sequence variation. These differed from each other according to the number, timing, and duration of partnerships. Another clustering method, that uses external information for the division of the data, was also studied. Regression tree analysis was used to divide data into clusters based on childhood covariates. Two significant predictors of partnership histories – high self-control of emotions and child-centred parenting – were found, which altogether explained only 3.5% of the variability of partnership histories. In contrast, the three-cluster solution using Ward's method without external information resulted in $R^2 = 35\%$, which increased to 61% for the chosen eight-cluster solution. Hence, the predictive power of those covariates alone was very low, although this was to be expected as we did not account for many factors that previous studies have found to be related to partnership formation and dissolution (e.g. the presence and age of children, educational attainment, employment, income, religiosity, and health-related factors; see e.g. Aassve et al., 2006; Berrington & Diamond, 2000; Jalovaara, 2012; Lyngstad & Jalovaara, 2010; South, 2001; Steele et al., 2006). Many of these other factors are time-varying which is problematic with regression trees, and were therefore beyond the scope of the analysis. However, other life domains could be added as parallel sequences that can then be analysed with multi-dimensional sequence analysis methods (Gauthier et al., 2010; Müller, Sapin, Gauthier, Orita, & Widmer, 2012; Salmela-Aro et al., 2011). In a previous study, Eerola and Helske (2012) compared SA and EHA in a case of multiple parallel life domains using the same JYLS data.

Event history analysis was used to model the probability of partnership transitions between ages 15 to 50 as a function of individual (i.e., social activity and high self-control emotions) and family characteristics (i.e. child-centred home environment, SES, and structure of the family of origin). To account for dependency between the durations of repeated episodes, random effects models for partnership formations and dissolutions were fitted. For all but one model, there was no statistically significant unobserved variation between individuals once the childhood variables were included in the analyses, indicating that these factors captured a substantial part of the variation in partnership formation and dissolution that is due to time-invariant characteristics. A joint model of partnership formations and dissolution (as described in Section 2.3.2) was also fitted for women and men. The idea was to study whether there was correlation between the durations of episodes of living with and without a partner, for example because individuals who separate more rapidly tend to form new partnerships sooner than individuals whose partnerships last longer (as shown by Aassve et al., 2006; Steele et al., 2006 using British data). However, our sample was too small to estimate a joint model, leading to confidence intervals of correlation estimates ranging from −1 to 1.

Sequence analysis and event history analysis provide complementary information on partnership formation. Sequence analysis is a descriptive tool that gives an overall picture of the histories and compresses them in a form that is relatively easy to interpret. Sequences are often shown as colourful lines in an index plot, from which it is – especially

after clustering – easy to see the timing of important partnership transitions and the approximate duration of different episodes. Clustering helps to describe the data and to identify similar patterns in partnership formation by providing typologies of partnership trajectories. However, choosing the number of clusters is to some extent subjective. It is therefore important to consider a range of solutions and to regard the division of life sequences into clusters as suggestive. One should also be cautious about attaching too much meaning to a cluster or a label assigned to it, as the labels given to the clusters are only approximate since borderline cases could also be assigned to other clusters. For example, in the present study most of the members of the "later 1st partnership" cluster had stayed with their first partner but there were also several members who had lived separated or with a new partner for a long time.

Analysis of individual-level event histories is better for drawing inferences about the effects of covariates on the timing of recurring partnership transitions. It can account for censoring and unobserved individual characteristics that affect the timing and duration of partnerships. However, with discretely measured recurrent events, forming the data set can be time-consuming and the size of the person-episode-period type-of-data may be large even when the number of individuals is small, leading to long estimation times when random effects models are used.

Although SA and EHA are both methods for studying longitudinal life course data, their approaches in capturing time are different in many respects and they provide versatile information on the phenomenon of interest. In SA, the focus is on the holistic pattern of the histories and analysis is retrospective in nature. In contrast, in EHA the interest lies in the transitions and the direction of inference is prospective: how much time passes before an event happens. Each episode is as important as the others, no matter how short. In SA, however, especially with the most popular alignment methods for computing sequence dissimilarities such as OM and Hamming, small deviations from a general pattern might not be very influential. For example, in terms of our (rather restricted) state-space (Table 1), hypothetical sequences P1-P2-D-P3-D-P3-P3 and P1-P2-D-P3-P3-P3-P3 would have been regarded as very similar even though the former person had four partners and five transitions and the latter one only three partners and three transitions. The definition of the state-space also matters: had we not separated partnerships by order, distinguishing successive partnerships would have been even more difficult or indeed impossible (as with P1 and P2 in the example sequences above). In such cases, if it is important to treat each episode as distinct, other dissimilarity criteria such as those based on counting common subsequences might be better suited.

SA and EHA are, of course, not the only options suitable for studying discrete longitudinal life course data. For example, trajectory analysis (Nagin, 1999) and latent class analysis (LCA; e.g. Vermunt, Tran, & Magidson, 2008) come in the middle ground of the approaches presented in this paper by using statistical models to create homogenous clusters of similar trajectories. Semi-parametric trajectory analysis can be used for studying binary trajectories such as the histories of living single/in partnership. However, the method is not suited for categorical trajectories with more than two unordered categories. For categorical data, LCA has been used to group trajectories. The standard version of LCA does not take into account the correlation between observations measured in different time periods, but several modifications have been proposed to adjust for the temporal correlation. See Barban and Billari (2012) for a comparison of LCA to SA.

4.2 Partnership formation and dissolution

Different factors related to childhood and current life situation were found to be connected to partnership formation and dissolution for women and men. Contrary to previous research (e.g. Berrington & Diamond, 2000; Rönkä et al., 2000; Ross et al., 2009; Steele et al., 2006), we did not find a significant effect of SES of subjects' fathers on the timing of their first partnerships. In common with previous research by Goldstein et al. (2004) and Steele et al. (2006), the SES of the childhood family was also not connected to men's risk of repartnering, but women with higher SES background had a lower risk.

Many previous studies have shown an increased dissolution risk for higher-order unions, but this has been assumed to be at least partly due to selection on unobserved individual characteristics. Studies that have considered such characteristics have not found an excessive risk of dissolution for recurrent partnerships (Aassve et al., 2006; Lillard et al., 1995; Poortman & Lyngstad, 2007; Steele et al., 2005;

Steele et al., 2006), although few have studied men. Our finding that repartnered men had a higher risk of dissolution was in contrast to studies of British (Aassve et al. 2006) and Norwegian (Poortman and Lyngstad 2007) men, which did not find differences in the dissolution risk by partnership order.

In common with previous studies (e.g. Andersson, 2002; Liefbroer & Dourleijn, 2006; Manning, Smock, & Majumdar, 2004), married women were less likely to dissolve their partnerships (first as well as recurrent) compared to cohabiting women. In contrast, cohabiting and married men did not differ in their risks. Motherhood did not change the risk of dissolution but fathers had a lower risk compared to childless men. However, the models only accounted for having (biological or adopted) children in general. By choosing this conceptualisation of parenthood, some information about the effects of children on the risk of dissolution of partnership is inevitably lost. Earlier research has found different, even opposite, effects of the presence, number, and age of children on partnership dissolution across countries (Coppola & Di Cesare, 2008; Lillard & Waite, 1993; Lyngstad & Jalovaara, 2010; Steele et al., 2005; Svarer & Verner, 2008).

Of the socio-emotional characteristics considered, high self-control of emotions at age 8 was the strongest explanatory variable of partnership transitions. As expected, individuals with high self-control of emotions, indicated by emotional stability and constructive and compliant behaviour (Kokko et al., 2008), had a lower risk of partnership dissolution. For women the probability of repartnering was also lower but, contrary to our expectations, there was no association with the timing of the first partnership. Furthermore, high self-control of emotions was also related to fewer and more stable partnerships for participants who had experienced less child-centred parenting practices during childhood. These results suggest that high self-control of emotions was associated with a more stable family life, even for those individuals with a less supportive family environment in childhood. It is possible that a stable partnership was a part of a cycle of good social functioning linked to child's high self-control of emotions (Pulkkinen, 2009).

In accordance with our expectations, high social activity in childhood was related to men's tendency to form first and also subsequent partnerships faster. Among women, social activity was not related to the timing or pace of partnership events. This difference could be partly due to diverse forms of social activity in boys and girls. In a previous study, Pulkkinen (1995) found that high social activity in boys was more often linked with unfavourable behaviour.

4.3 Limitations and strengths

When interpreting our results, there are some limitations that should be noted. First, our analyses consider only one age cohort of one nationality. Therefore our findings may not generalise to older and younger age cohorts and other nationalities, although many of our results were consistent with previous studies. Second, information on partnerships was gathered using the Life History Calendar (LHC), presented to the JYLS participants during the age 42 and age 50 personal interviews (in 2001 and 2009, respectively). The LHCs covered a time span from age 15 to 50. The long recall period may raise questions about the accuracy of the participants' memory and the validity of the LHC data. However, we do not consider this to be a serious flaw because prospective data on these transitions were also gathered in the JYLS study and these data have been informally used to check the validity of the LHC data (Kokko et al., 2009). Furthermore, previous studies have shown that information gathered with the LHC is reliable (Caspi et al., 1996; Freedman, Thornton, Camburn, Alwin, & Young-DeMarco, 1988). A third limitation of our study is that, in common with most other birth cohort studies where life histories are collected retrospectively, we do not have data on the childhood characteristics and partnership histories of the partners of cohort members.

The two data collection phases led to a high proportion of partnership histories that were right-censored at age 42 (the time of the first phase). We were therefore forced to use missing states for these shorter sequences, which in turn led to problems in the definition of costs in SA. Clustering results made most sense when the cost for aligning any state to a missing state was set to zero. However, this cost setting resulted in Hamming dissimilarities that are not metric distances, as assumed by most clustering methods. Since the chosen clusters were reasonable, and in any case considered suggestive, the use of non-metric dissimilarities is most likely not very serious.

Even though the JYLS study is long and extensive, the moderate sample size imposed many restric-

tions in model building. For example, we were unable to model partnership formations and dissolutions jointly. Moreover, when specifying a piecewise constant baseline hazard function we were forced to use broad age intervals. We considered only a simple indicator of being a parent which did not account for the different aspects of family structure that other studies have found to be related to the risk of partnership formation and dissolution (such as the number, age, and residence of the child(ren) or blended families). We also faced challenges due to the coarse annual measurements and had to be careful when defining the risk sets: for some individuals there seemed to be no unpartnered episodes between two partnerships.

Although the use of the JYLS data imposed methodological restrictions, strengths of the data are the rich covariate information and exceptionally long period of follow-up (from age 8 to 50). This enabled the examination of childhood individual and family characteristics as precursors of partnership transitions measured up to middle-age. In particular, childhood socio-emotional characteristics have not been studied before in this context. As can be seen from the non-significant random effect variances in some of the models, we could capture a notable part of the variation due to time-invariant individual characteristics that in previous studies have simply been left to the unobserved random part. The research question concerned the effects of childhood characteristics on the timing and stability of partnerships. These childhood measures were not used as proxies for the socio-emotional qualities of an adult. Nevertheless, a significant relationship between childhood socio-emotional characteristics and adult personality has been found in the JYLS data (Pulkkinen et al., 2012).

Another contribution of this paper was to demonstrate and compare use of SA and EHA, which to our knowledge is the first attempt to apply both methods in a study of recurrent life events.

Acknowledgements

We appreciate Professor Lea Pulkkinen's contribution to the JYLS over the years. We also thank the referees for their helpful comments and suggestions.

Satu Helske has been funded by the Jyväskylä Graduate School in Computing and Mathematical Sciences (COMAS) and the Finnish Cultural Foundation. Fiona Steele was supported by an ESRC grant for a node of the National Centre for Research Methods (RES-576-25-0032). The major funder of the Jyväskylä Longitudinal Study of Personality and Social Development (JYLS) has been the Academy of Finland, most recently through grant nos. 127125 (Pulkkinen) and 118316 and 135347 (Kokko).

References

Aassve, A., Burgess, S., Propper, C., & Dickson, M. (2006). Employment, family union and childbearing decisions in Great Britain. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 781-804. http://dx.doi.org/10.1111/j.1467-985X.2006.00432.x

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. http://dx.doi.org/10.1007/s10680-007-9134-6

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods, 16*(4), 129-147. http://dx.doi.org/10.1080/01615440.1983.10594107

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*(1), 93-113.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research, 29*(1), 3-33. http://dx.doi.org/10.1177/0049124100029001001

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis – Bringing the "course" Back Into the life course. *Sociological Methods & Research, 38*(3), 420-462. http://dx.doi.org/10.1177/0049124100029001001

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology, 13*(1), 61-98. http://dx.doi.org/10.2307/270718

Amato, P. R. (1996). Explaining the intergenerational transmission of divorce. *Journal of Marriage and the Family, 58*(3), 628-640. http://dx.doi.org/10.2307/353723

Andersson, G. (2002). Children's experience of family disruption and family formation: Evidence from 16 FFS countries. *Demographic Research, 7*(7), 343-364. http://dx.doi.org/10.4054/DemRes.2002.7.7

Barban, N., & 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. http://dx.doi.org/10.1111/j.1467-9876.2012.01047.x

Berrington, A., & Diamond, I. (2000). Marriage or cohabitation: A competing risks analysis of first-partnership formation among the 1958 British birth cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 163*(2), 127-151. http://dx.doi.org/10.1111/1467-985X.00162

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* [Inc.] Belmont, CA: Wadsworth and Brooks.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *Stata Journal, 6*(4), 435.

Bumpass, L. L., Martin, T. C., & Sweet, J. A. (1991). The impact of family background and early marital factors on marital disruption. *Journal of Family Issues, 12*(1), 22-42. http://dx.doi.org/10.1177/019251391012001003

Caspi, A., Elder, G. H., & Bem, D. J. (1988). Moving away from the world: Life-course patterns of shy children. *Developmental Psychology, 24*(6), 824-831. http://dx.doi.org/10.1037/0012-1649.24.6.824

Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., & others. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research, 6*(2), 101-114. http://dx.doi.org/10.1002/(SICI)1234-988X(199607)6:2<101::AID-MPR156>3.3.CO;2-E

Coppola, L., & Di Cesare, M. (2008). How fertility and union stability interact in shaping new family patterns in Italy and Spain. *Demographic Research, 18*(4), 117-144. http://dx.doi.org/10.4054/DemRes.2008.18.4

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511790492

Eerola, M., & Helske, S. (2012). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research.* Advance online publication. http://dx.doi.org/10.1177/0962280212461205

Elder, G. H. (1998). The life course and human development. In W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Volume 1: Theoretical models of human development* (5th ed., pp. 939-991) Hoboken, US: Wiley.

Elzinga, C. H. (2006). Sequence analysis: Metric representations of categorical time series. *Manuscript.*

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue Européenne De Démographie, 23*(3), 225-250. http://dx.doi.org/10.1007/s10680-007-9133-7

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology, 18*, 37. http://dx.doi.org/10.2307/271044

Furstenberg, F. F., Jr., & Spanier, G. B. (1984). The risk of dissolution in remarriage: An examination of Cherlin's hypothesis of incomplete institutionalization. *Family Relations, 33*(3) 33-441. http://dx.doi.org/10.2307/584714

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.

Gähler, M., Hong, Y., & Bernhardt, E. (2009). Parental divorce and union disruption among young adults in Sweden. *Journal of Family Issues, 30*(5), 688-713. http://dx.doi.org/10.1177/0192513X08331028

Gauthier, J., Widmer, E. D., Bucher, P., & Notredame, C. (2009). How much does it cost? Optimization of costs in sequence analysis of social science data. *Sociological Methods & Research, 38*(1), 197-231. http://dx.doi.org/10.1177/0049124109342065

Gauthier, J., Widmer, E. D., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology, 40*(1), 1-38. http://dx.doi.org/10.1111/j.1467-9531.2010.01227.x

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester, UK: John Wiley & Sons.

Goldstein, H., Pan, H., & Bynner, J. (2004). A flexible procedure for analyzing longitudinal event histories using a multilevel model. *Understanding Statistics, 3*(2), 85-99. http://dx.doi.org/10.1207/s15328031us0302_2

Halpin, B. (2014). *SADI: Sequence analysis tools for Stata.* Unpublished manuscript.

Halpin, B. (2010). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research, 38*(3), 365-388. http://dx.doi.org/10.1177/0049124110363590

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 29*(2), 147-160. http://dx.doi.org/10.1002/j.1538-7305.1950.tb00463.x

Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods & Research, 38*(2), 235-264. http://dx.doi.org/10.1177/0049124109346164

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218. http://dx.doi.org/10.1007/BF01908075

Jalovaara, M. (2012) Socio-economic resources and first-union formation in Finland, cohorts born 1969–81. *Population studies*, 66(1), 69-85.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Hoboken, New Jersey: John Wiley & Sons.

Kennedy, S., & Bumpass, L. (2008). Cohabitation and children's living arrangements: New estimates from the United States. *Demographic Research, 19*, 1663-1692. http://dx.doi.org/10.4054/DemRes.2008.19.47

Kiernan, K. (2001). The rise of cohabitation and childbearing outside marriage in Western Europe. *International Journal of Law, Policy and the Family, 15*(1), 1-21. http://dx.doi.org/10.1093/lawfam/15.1.1

Kinnunen, U., & Pulkkinen, L. (2003). Childhood socio-emotional characteristics as antecedents of marital stability and quality. *European Psychologist, 8*(4), 223-237. http://dx.doi.org/10.1027//1016-9040.8.4.223

Kokko, K., & Pulkkinen, L. (2000). Aggression in childhood and long-term unemployment in adulthood: A cycle of maladaptation and some protective factors. *Developmental Psychology, 36*(4), 463-472. http://dx.doi.org/10.1037/0012-1649.36.4.463

Kokko, K., Pulkkinen, L., & Mesiäinen, P. (2009). Timing of parenthood in relation to other life transitions and adult social functioning. *International Journal of Behavioral Development, 33*(4), 356-365. http://dx.doi.org/10.1177/0165025409103873

Kokko, K., Pulkkinen, L., Mesiäinen, P., & Lyyra, A. (2008). Trajectories based on post-comprehensive and higher education and their correlates and antecedents. *Journal of Social Issues, 64*(1), 59-76. http://dx.doi.org/10.1111/j.1540-4560.2008.00548.x

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research, 38*(3), 389-419. http://dx.doi.org/10.1177/0049124110362526

Liefbroer, A. C., & Dourleijn, E. (2006). Unmarried cohabitation and union stability: Testing the role of diffusion using data from 16 European countries. *Demography, 43*(2), 203-221. http://dx.doi.org/10.1353/dem.2006.0018

Lillard, L. A., Brien, M. J., & Waite, L. J. (1995). Premarital cohabitation and subsequent marital dissolution: A matter of self-selection? *Demography, 32*(3), 437-457. http://dx.doi.org/10.2307/2061690

Lillard, L. A., & Waite, L. J. (1993). A joint model of marital childbearing and marital disruption. *Demography, 30*(4), 653-681. http://dx.doi.org/10.2307/2061812

Lyngstad, T. H. (2006). Why do couples with highly educated parents have higher divorce rates? *European Sociological Review, 22*(1), 49-60. http://dx.doi.org/10.1093/esr/jci041

Lyngstad, T. H., & Jalovaara, M. (2010). A review of the antecedents of union dissolution. *Demographic Research, 23*(10), 257-292. http://dx.doi.org/10.4054/DemRes.2010.23.10

Manning, W. D., Smock, P. J., & Majumdar, D. (2004). The relative stability of cohabiting and marital unions for children. *Population Research and Policy Review, 23*(2), 135-159. http://dx.doi.org/10.1023/B:POPU.0000019916.29156.a7

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 165*(2), 317-334. http://dx.doi.org/10.1111/1467-985X.00641

Müller, N. S., Sapin, M., Gauthier, J., 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. http://dx.doi.org/10.1177/0020764010393630

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4*(2), 139. http://dx.doi.org/10.1037/1082-989X.4.2.139

Piccarreta, R., & Lior, O. (2010). Exploring sequences: A graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(1), 165-184. http://dx.doi.org/10.1111/j.1467-985X.2009.00606.x

Pitkänen, T., Kokko, K., Lyyra, A., & Pulkkinen, L. (2008). A developmental approach to alcohol drinking behaviour in adulthood: A follow-up study from age 8 to age 42. *Addiction, 103*(s1), 48-68. http://dx.doi.org/10.1111/j.1360-0443.2008.02176.x

Pitkänen, T., Lyyra, A., & Pulkkinen, L. (2005). Age of onset of drinking and the use of alcohol in adulthood: A follow-up study from age 8–42 for females and males. *Addiction, 100*(5), 652-661. http://dx.doi.org/10.1111/j.1360-0443.2005.01053.x

Poortman, A., & Lyngstad, T. H. (2007). Dissolution risks in first and higher order marital and cohabiting unions. *Social Science Research, 36*(4), 1431-1446. http://dx.doi.org/10.1016/j.ssresearch.2007.02.005

Pulkkinen, L. (1995). Behavioral precursors to accidents and resulting physical impairment. *Child Development, 66*(6), 1660-1679. http://dx.doi.org/10.2307/1131902

Pulkkinen, L. (2009). Personality – a resource or risk for successful development. *Scandinavian Journal of Psychology, 50*(6), 602-610. http://dx.doi.org/10.1111/j.1467-9450.2009.00774.x

Pulkkinen, L., & Kokko, K. (2010). Keski-ikä elämänvaiheena [middle-age as a stage of life]. (pp. 5-13) Jyväskylä, Finland: University of Jyväskylä.

Pulkkinen, L., Kokko, K., & Rantanen, J. (2012). Paths from socioemotional behavior in middle childhood to personality in middle adulthood. *Developmental Psychology, 48*(5), 1283-1291. http://dx.doi.org/10.1037/a0027463

Pulkkinen, L., Lyyra, A., & Kokko, K. (2009). Life success of males on nonoffender, adolescence-limited, persistent, and adult-onset antisocial pathways: Follow-up from age 8 to 42. *Aggressive Behavior, 35*(2), 117-135. http://dx.doi.org/10.1002/ab.20297

Räikkönen, E., Kokko, K., Chen, M., & Pulkkinen, L. (2012). Patterns of adult roles, their antecedents and psychosocial wellbeing correlates among Finns born in 1959. *Longitudinal and Life Course Studies, 3*(2), 211-227.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Rohwer, G., & Pötter, U. (2004). *TDA user's manual.* Bochum: Ruhr-Universität Bochum.

Rönkä, A., Kinnunen, U., & Pulkkinen, L. (2000). The accumulation of problems of social functioning as a long-term process: Women and men compared. *International Journal of Behavioral Development, 24*(4), 442-450. http://dx.doi.org/10.1080/016502500750037991

Rönkä, A., & Pulkkinen, L. (1998). Work involvement and timing of motherhood in the accumulation of problems in social functioning in young women. *Journal of Research on Adolescence, 8*(2), 221-239. http://dx.doi.org/10.1207/s15327795jra0802_3

Ross, A., Schoon, I., Martin, P., & Sacker, A. (2009). Family and nonfamily role configurations in two British cohorts. *Journal of Marriage and Family, 71*(1), 1-14. http://dx.doi.org/10.1111/j.1741-3737.2008.00576.x

Salmela-Aro, K., Kiuru, N., Nurmi, J., & Eerola, M. (2011). Mapping pathways to adulthood among Finnish university students: Sequences, patterns, variations in family- and work-related roles. *Advances in Life Course Research, 16*(1), 25-41. http://dx.doi.org/10.1016/j.alcr.2011.01.003

Shanahan, M. J. (2000). Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. *Annual Review of Sociology,* 667-692. http://dx.doi.org/10.1146/annurev.soc.26.1.667

South, S. J. (2001) The variable effects of family background on the timing of first marriage: United States, 1969–1993. *Social Science Research*, *30*(4), 606-626.

Statistics Finland. (1994). *Suomalainen lapsiperhe [The Finnish family with children].* Helsinki, Finland: Hakapaino Oy.

Statistics Finland. (2010). *Statistical yearbook of Finland 2010.* Helsinki, Finland: Tilastokeskus.

Steele, F. (2011). Multilevel discrete-time event history analysis with applications to the analysis of recurrent employment transitions. *Australian & New Zealand Journal of Statistics, 53*(1), 1-20.

Steele, F., Kallis, C., Goldstein, H., & Joshi, H. (2005). The relationship between childbearing and transitions from marriage and cohabitation in Britain. *Demography, 42*(4), 647-673. http://dx.doi.org/10.1353/dem.2005.0038

Steele, F., Kallis, C., & Joshi, H. (2006). The formation and outcomes of cohabiting and marital partnerships in early adulthood: The role of previous partnership experience. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 757-779. http://dx.doi.org/10.1111/j.1467-985X.2006.00420.x

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research, 40*(3), 471-510. http://dx.doi.org/10.1177/0049124111415372

Svarer, M., & Verner, M. (2008). Do children stabilize relationships in Denmark? *Journal of Population Economics, 21*(2), 395-417. http://dx.doi.org/10.1007/s00148-006-0084-9

Teachman, J. (2008). Complex life course patterns and the risk of divorce in second marriages. *Journal of Marriage and Family, 70*(2), 294-305. http://dx.doi.org/10.1111/j.1741-3737.2008.00482.x

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16*(3), 439-454. http://dx.doi.org/10.2307/2061224

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*. Burlington, MA: Elsevier. 373-385.

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244. http://dx.doi.org/10.1080/01621459.1963.10500845

III

Helske, S. ja Helske, J. (2015) Mixture hidden Markov models for sequence data: the seqHMM package in R.

Mixture Hidden Markov Models for Sequence Data: the seqHMM Package in R

Satu Helske
University of Jyväskylä, Finland

Jouni Helske
University of Jyväskylä, Finland

Abstract

Sequence analysis is being more and more widely used for the analysis of social sequences and other multivariate categorical time series data. However, it is often complex to describe, visualize, and compare large sequence data, especially when there are multiple parallel sequences per subject. Hidden (latent) Markov models (HMMs) are able to detect underlying latent structures and they can be used in various longitudinal settings: to account for measurement error, to detect unobservable states, or to compress information across several types of observations. Extending to mixture hidden Markov models (MHMMs) allows clustering data into homogeneous subsets, with or without external covariates.

The **seqHMM** package in R is designed for the efficient modeling of sequences and other categorical time series data containing one or multiple subjects with one or multiple interdependent sequences using HMMs and MHMMs. Also other restricted variants of the MHMM can be fitted, e.g. latent class models, Markov models, mixture Markov models, or even ordinary multinomial regression models with suitable parameterization of the HMM.

Good graphical presentations of data and models are useful during the whole analysis process from the first glimpse at the data to model fitting and presentation of results. The package provides easy options for plotting parallel sequence data, and proposes visualizing HMMs as directed graphs.

Keywords: multichannel sequences, categorical time series, visualizing sequence data, visualizing models, latent Markov models, latent class models, R.

A CRAN-compliant modification of a manuscript submitted to *Journal of Statistical Software*.

1. Introduction

Sequence analysis is being more and more widely used for the analysis of categorical time series. These data consist of multiple independent subjects with one or multiple interdependent sequences (channels). Sequence analysis is used for computing the (dis)similarities of sequences, and often the goal is to find patterns in data using cluster analysis. However, describing, visualizing, and comparing large sequence data is often complex, especially in the case of multiple channels. Hidden (latent) Markov models (HMMs) can be used to compress and visualize information in such data. These models are able to detect underlying latent structures. Extending to mixture hidden Markov models (MHMMs) allows clustering via latent classes, possibly with additional covariate information. One of the major benefits of using hidden Markov modeling is that all stages of analysis are performed, evaluated, and

compared in a probabilistic framework; e.g. well-known model selection criteria are available for choosing the best clustering solution.

The **seqHMM** package for R (R Core Team 2015) is designed for modeling sequence data and other categorical time series with one or multiple subjects and one or multiple channels using HMMs and MHMMs. The package provides functions for the estimation and inference of models, as well as functions for the easy visualization of multichannel sequences and HMMs. Even though the package was originally developed for researchers familiar with social sequence analysis, knowledge on sequence analysis or social sciences is not necessary for the usage of **seqHMM**. The package is available on Comprehensive R Archive Repository (CRAN) and easily installed via `install.packages("seqHMM")`. Development versions can be obtained from GitHub[1].

There are also other R packages in CRAN for HMM analysis of categorical data. The **HMM** package (Himmelmann 2010) is a compact package designed for fitting an HMM for a single observation sequence. The **hmm.discnp** package (Turner and Liu 2014) can handle multiple observation sequences with possibly varying lengths. For modeling continuous-time processes as hidden Markov models, the **msm** package (Jackson 2011) is available. Both **hmm.discnp** and **msm** support only single-channel observations. The **depmixS4** package (Visser and Speekenbrink 2010) is able to fit HMMs for multiple interdependent time series (with continuous or categorical values), but for one subject only. In the **msm** and **depmixS4** packages, covariates can be added for initial and transition probabilities. The **mhsmm** package (O'Connell and Højsgaard 2011) allows modeling of multiple sequences using hidden Markov and semi-Markov models. There are no ready-made options for modeling categorical data, but users can write their own extensions for arbitrary distributions. The **LMest** package (Bartolucci and Pandolfi 2015) is aimed to panel data with a large number of subjects and a small number of time points. It can be used for hidden Markov modeling of multivariate and multichannel categorical data, using covariates in emission and transition processes. **LMest** also supports mixed latent Markov models, where the latent process is allowed to vary in different latent subpopulations. This differs from mixture hidden Markov models used in **seqHMM**, where also the emission probabilities vary between groups. The **seqHMM** package also supports covariates in explaining group memberships. A drawback in the **LMest** package is that the user cannot define initial values or zero constraints for model parameters, and thus important special cases such as left-to-right models cannot be used.

We start with describing data and methods: a short introduction to sequence data and sequence analysis, then the theory of hidden Markov models for such data, an expansion to mixture hidden Markov models and a glance at some special cases, and then some propositions on visualizing multichannel sequence data and hidden Markov models. After the theoretic part we take a look at features of the **seqHMM** package and at the end show an example on using the package for the analysis of life course data. Appendices include a list of notations and more thorough descriptions of some important algorithms.

[1] https://github.com/helske/seqHMM

2. Methods

2.1. Sequences and sequence analysis

By the term *sequence* we refer to an ordered set of categorical states. It can be a time series, such as a career trajectory or residental history, or any other series with ordered categorical observations, e.g. a DNA sequence or a structure of a story.

As an example we study the `biofam` data available in the **TraMineR** package (Gabadinho, Ritschard, Müller, and Studer 2011). It is a sample of 2000 individuals born in 1909–1972, constructed from the Swiss Household Panel survey in 2002 (Müller, Studer, and Ritschard 2007). The data set contains sequences of annual family life statuses from age 15 to 30. Eight observed states are defined from the combination of five basic states: living with parents, left home, married, having children, and divorced. To show a more complex example, we split the original data into three separate *channels* representing different life domains: marriage, parenthood, and residence. The data for each individual now includes three parallel sequences constituting of two or three *states* each: single/married/divorced, childless/parent, and living with parents / having left home.

Sequence analysis (SA) is statistical analysis of successions of states. It has roots in bioinformatics and computer science (see e.g. Durbin, Eddy, Krogh, and Mitchison 1998), but during the past few decades SA has also become more common in other disciplines for the analysis of longitudinal data. In social sciences SA has been used increasingly often and is now "central to the life-course perspective" (Blanchard, Bühlmann, and Gauthier 2014). SA is model-free data-driven approach, which is used for computing (dis)similarities of sequences. The most well-known method is optimal matching (McVicar and Anyadike-Danes 2002), but several alternatives exist (see e.g. Aisenbrey and Fasang 2010; Elzinga and Studer 2014; Gauthier, Widmer, Bucher, and Notredame 2009; Halpin 2010; Hollister 2009; Lesnard 2010). Also a method for analysing multichannel data has been developed (Gauthier, Widmer, Bucher, and Notredame 2010). Often the goal in SA is to find typical and atypical patterns in trajectories using cluster analysis, but any approach suitable for compressing information on the dissimilarities can be used. The data are usually presented also graphically in some way. So far the **TraMineR** package has been the most extensive and frequently used software for social sequence analysis.

2.2. Hidden Markov models

In the context of hidden Markov models, sequence data consists of *observed states*, which are regarded as probabilistic functions of *hidden states*. Hidden states cannot be observed directly, but only through the sequence(s) of observations, since they emit the observations on varying probabilities. A discrete first order hidden Markov model for a single sequence is characterized by the following:

- *Observed state sequence* $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ with observed states $m \in \{1, \ldots, M\}$.
- *Hidden state sequence* $\mathbf{z} = (z_1, z_2, \ldots, z_T)$ with hidden states $s \in \{1, \ldots, S\}$.
- *Transition 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:

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

  We only consider homogeneous HMMs, where the transition probabilities a_{sr} are constant over time.

- *Emission matrix* $B = \{b_s(m)\}$ of size $S \times M$, where $b_s(m)$ is the probability of the hidden state s emitting the observed state m:

$$b_s(m) = P(y_t = m | z_t = s); \quad s \in \{1, \ldots, S\}, m \in \{1, \ldots, M\}.$$

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

$$\pi_s = P(z_1 = s); \quad s \in \{1, \ldots, S\}.$$

The (first order) Markov assumption states that the hidden state transition probability at time t only depends on the hidden state at the previous time point $t-1$:

$$P(z_t | z_{t-1}, \ldots, z_1) = P(z_t | z_{t-1}). \tag{1}$$

Also, the observation at time t is only dependent on the current hidden state, not on previous hidden states or observations:

$$P(y_t | y_{t-1}, \ldots, y_1, z_t, \ldots, z_1) = P(y_t | z_t). \tag{2}$$

For a more detailed description of hidden Markov models, see e.g. Rabiner (1989), MacDonald and Zucchini (1997), and Durbin *et al.* (1998).

HMM for multiple sequences

We can also fit the same HMM for multiple subjects; instead of one observed sequence $\mathbf{y}$ we have N sequences as $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^\top$, where the observations $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ of each subject i take values in the observed state space. Observed sequences are assumed to be mutually independent given the hidden states. The observations are assumed to be generated by the same model, but each subject has its own hidden state sequence.

HMM for multichannel sequences

In the case of multichannel sequence data, such as the example described in section 2.1, for each subject i there are C parallel sequences. Observations are now of the form y_{itc}, $i = 1, \ldots, N; \quad t = 1 \ldots, T; \quad c = 1 \ldots, C$, so that our complete data is $Y = \{Y^1, \ldots, Y^C\}$. In **seqHMM**, multichannel data are handled as a list of C data frames of size $N \times T$. We also define Y_i as all the observations corresponding to subject i.

We apply the same latent structure for all channels. In such a case the model has one transition matrix A but several emission matrices $B_1, \ldots, B_C$, one for each channel. We assume that the observed states in different channels at a given time point t are independent of each other given the hidden state at t, i.e., $P(\mathbf{y}_{it}|z_{it}) = P(y_{it1}|z_{it}) \cdots P(y_{itC}|z_{it})$.

Sometimes the independence assumption does not seem theoretically plausible. For example, even conditioning on a hidden state representing a general life stage, are marital status and parenthood truly independent? On the other hand, given a person's religious views, could their opinions on abortion and gay marriage be though as independent?

If the goal is to use hidden Markov models for prediction or simulating new sequence data, the analyst should carefully check the validity of independence assumptions. However, if the goal is merely to describe structures and compress information, it can be useful to accept the independence assumption even though it is not completely reasonable in a theoretical sense. When using multichannel sequences, the number of observed states is smaller, which leads to a more parsimonious representation of the model and easier inference of the phenomenon. Also due to the decreased number of observed states, the number of parameters of the model is decreased leading to the improved computational efficiency of model estimation.

The multichannel approach is particularly useful if some of the channels are only partially observed; combining missing and non-missing information into one observation is usually problematic. One would have to decide whether such observations are coded completely missing, which is simple but loses information, or whether all possible combinations of missing and non-missing states are included, which grows the state space larger and makes the interpretation of the model more difficult. In the multichannel approach the data can be used as it is.

Missing data

Missing observations are handled straightforwardly in the context of HMMs. When observation y_{itc} is missing, we gain no additional information regarding hidden states. In such a case, we set the emission probability $b_s(y_{itc}) = 1$ for all $s \in 1, \ldots, S$. Sequences with varying lengths are handled by setting missing values before and/or after the observed states.

Log-likelihood and parameter estimation

The unknown transition, emission and initial probabilities are commonly estimated via maximum likelihood. The log-likelihood for multiple multichannel sequences is written as

$$\log L = \sum_{i=1}^{N} \log P\left(Y_i|\mathcal{M}\right), \tag{3}$$

where Y_i are the observed sequences in channels $c = 1, \ldots, C$ for subject i and $\mathcal{M}$ describes the model and its parameters $\{\pi, A, B_1, \ldots, B_C\}$. The probability of the observation sequence

of subject i given the model is

$$\begin{aligned} P(Y_i|\mathcal{M}) &= \sum_{\text{all } z} P\left(Y_i|z, \mathcal{M}\right) P\left(z|\mathcal{M}\right) \\ &= \sum_{\text{all } z} P(z_1|\mathcal{M}) P(\mathbf{y}_{i1}|z_1, \mathcal{M}) \prod_{t=2}^{T} P(z_t|z_{t-1}, \mathcal{M}) P(\mathbf{y}_{it}|z_t, \mathcal{M}) \\ &= \sum_{\text{all } z} \pi_{z_1} b_{z_1}(y_{i11}) \cdots b_{z_1}(y_{i1C}) \prod_{t=2}^{T} \left[a_{z_{t-1}z_t} b_{z_t}(y_{it1}) \cdots b_{z_t}(y_{itC})\right], \end{aligned} \tag{4}$$

where the hidden state sequences $z = (z_1, \ldots, z_T)$ take all possible combinations of values in the hidden state space $\{1, \ldots, S\}$ and where $\mathbf{y}_{it}$ are the observations of subject i at t in channels $1, \ldots, C$; π_{z_1} is the initial probability of the hidden state at time $t = 1$ in sequence z; $a_{z_{t-1}z_t}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t; and $b_{z_t}(y_{itc})$ is the probability that the hidden state of subject i at time t emits the observed state at t in channel c.

For direct numerical maximization (DNM) of the log-likelihood, any general-purpose optimization routines such as BFGS or Nelder–Mead can be used (with suitable reparameterizations). Another common estimation method is the expectation–maximization (EM) algorithm, also known as the Baum–Welch algorithm in the HMM context. The EM algorithm rapidly converges close to a local optimum, but compared to DNM, the converge speed is often slow near the optimum.

The probability (4) is efficiently calculated using the forward part of the *forward–backward algorithm* (Baum and Petrie 1966; Rabiner 1989, see appendix B). The backward part of the algorithm is needed for the EM algorithm, as well as for computation of analytical gradients for derivative based optimization routines.

The estimation process starts by giving initial values to the estimates. Good starting values are needed for finding the optimal solution in a reasonable time. In order to reduce the risk of being trapped in a poor local maximum, a large number of initial values should be tested.

Inference on hidden states

Given our model and observed sequences, we can make several interesting inferences regarding the hidden states. Forward probabilities $\alpha_{it}(s)$ (Rabiner 1989) are defined as the joint probability of hidden state s at time t and the observation sequences $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{it}$ given the model $\mathcal{M}$, whereas backward probabilities $\beta_{it}(s)$ are defined as the joint probability of hidden state s at time t and the observation sequences $\mathbf{y}_{i(t+1)}, \ldots, \mathbf{y}_{iT}$ given the model $\mathcal{M}$.

From forward and backward probabilities we can compute the *posterior probabilities* of states, which give the probability of being in each hidden state at each time point, given the observed sequences of subject i. These are defined as

$$P(z_{it} = s|Y_i, \mathcal{M}) = \frac{\alpha_{it}\beta_{it}}{P\left(Y_i|\mathcal{M}\right)}. \tag{5}$$

Posterior probabilities can be used to find the locally most probable hidden state at each time point, but the resulting sequence is not necessarily globally optimal. To find the single best hidden state sequence $\hat{z}_i(Y_i) = \hat{z}_{i1}, \hat{z}_{i2}, \ldots, \hat{z}_{iT}$ for subject i, we maximize $P(z|Y_i, \mathcal{M})$ or,

equivalently, $P(z, Y_i|\mathcal{M})$. A dynamic programming method, the *Viterbi algorithm* (Rabiner 1989, see appendix C), is used for solving the problem.

Model comparison

Models with the same number of parameters can be compared with the log-likelihood. For choosing between models with a different number of hidden states, we need to take account of the number of parameters. We define the Bayesian information criterion (BIC) as

$$BIC = -2\log(L_d) + p\log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{C}\sum_{c=1}^{C}\mathrm{I}(y_{itc} \text{ observed})\right), \tag{6}$$

where L_d is computed using equation 3, p is the number of estimated parameters, I is the indicator function, and the summation in the logarithm is the size of the data. If data are completely observed, the summation is simplified to $N \times T$. Missing observations in multichannel data may lead to non-integer data size.

2.3. Clustering by mixture hidden Markov models

There are many approaches for finding and describing clusters or latent classes when working with HMMs. A simple option is to group sequences beforehand (e.g. using sequence analysis and some clustering method), after which one HMM is fitted for each cluster. This approach is simple in terms of HMMs. Models with a different number of hidden states and initial values are explored and compared one cluster at a time. HMMs are used for compressing information and comparing different clustering solutions, e.g. finding the best number of clusters. The problem with this solution is that it is, of course, very sensitive to the original clustering and the estimated HMMs might not be well suited for borderline cases.

Instead of fixing sequences into clusters, it is possible to fit one model for the whole data and determine clustering during modeling. Now sequences are not in fixed clusters but get assigned to clusters with certain probabilities during the modeling process. In this section we expand the idea of HMMs to mixture hidden Markov models (MHMMs). This approach was formulated by van de Pol and Langeheine (1990) as a mixed Markov latent class model and later generalized to include time-constant and time-varying covariates by Vermunt, Tran, and Magidson (2008) (who named the resulting model as mixture latent Markov model, MLMM). The MHMM presented here is a variant of MLMM where only time-constant covariates are allowed. Time-constant covariates deal with unobserved heterogeneity and they are used for predicting cluster memberships of subjects.

Mixture hidden Markov model

Assume that we have a set of models $\mathcal{M} = \{\mathcal{M}^1, \ldots, \mathcal{M}^K\}$, where $\mathcal{M}^k = \{\pi^k, A^k, B_1^k, \ldots, B_C^k\}$ for $k = 1, \ldots, K$. For each subject Y_i, denote $P(\mathcal{M}^k) = w_k$ as the prior probability that the observation sequences of subject i belongs to the submodel/cluster $\mathcal{M}^k$. Now the log-

likelihood is extended from equation (3) as

$$\begin{aligned}\log L &= \sum_{i=1}^{N} \log P(Y_i|\mathcal{M}) \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(\mathcal{M}^k) \sum_{\text{all } z} P\left(Y_i|z, \mathcal{M}^k\right) P\left(z|\mathcal{M}^k\right) \right] \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \sum_{\text{all } z} \pi^k_{z_1} b^k_{z_1}(y_{i11}) \cdots b^k_{z_1}(y_{i1C}) \prod_{t=2}^{T} \left[a^k_{z_{t-1}z_t} b^k_{z_t}(y_{it1}) \cdots b^k_{z_t}(y_{itC}) \right] \right].\end{aligned} \tag{7}$$

Compared to the usual hidden Markov model, there is an additional summation over the clusters in equation (7), which seems to make the computations less straightforward than in the non-mixture case. Fortunately, by redefining MHMM as a special type HMM allows us to use standard HMM algorithms without major modifications. We combine the K submodels into one large hidden Markov model consisting of $\sum_{k=1}^{K} S_k$ states, where the initial state vector contains elements of the form $w_k \pi^k$. Now the transition matrix is block diagonal

$$A = \begin{pmatrix} A^1 & 0 & \cdots & 0 \\ 0 & A^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A^K \end{pmatrix}, \tag{8}$$

where the diagonal blocks $A^k, k = 1, \ldots, K$, are square matrices containing the transition probabilities of one cluster. The off-diagonal blocks are zero matrices, so transitions between clusters are not allowed. Similarly, the emission matrices for each channel contain stacked emission matrices B^k.

Covariates and cluster probabilities

Covariates can be added to MHMM to explain cluster memberships as in latent class analysis. The prior cluster probabilities now depend on the subject's covariate values $\mathbf{x}_i$ and are defined as multinomial distribution:

$$P(\mathcal{M}^k|\mathbf{x}_i) = w_{ik} = \frac{e^{\beta_k \mathbf{x}_i}}{1 + \sum_{j=2}^{K} e^{\beta_j \mathbf{x}_i}}. \tag{9}$$

The first cluster is set as the reference by fixing $\beta_1 = 0$. Note that by convention we use β when referring to regression coefficients. It is not to be mixed with backward probabilities, which are usually given the same notation.

As in MHMM without covariates, we can still use standard HMM algorithms with a slight modification; we now allow initial state probabilities to vary between subjects. Of course, we also need to estimate the coefficients β. For direct numerical maximization the modifications are straightforward. In the EM algorithm, regarding the M-step for β, `seqHMM` uses Newton's method with analytical gradients and Hessian which are straightforward to compute given all other model parameters. This Hessian can also be used for computing the conditional standard errors of coefficients. For unconditional standard errors, which take account of

possible correlation between the estimates of β and other model parameters, the Hessian is computed using finite difference approximation of the Jacobian of the analytical gradients.

The posterior cluster probabilities $P(\mathcal{M}^k|Y_i, \mathbf{x}_i)$ are obtained as

$$\begin{aligned} P(\mathcal{M}^k|Y_i, \mathbf{x}_i) &= \frac{P(Y_i|\mathcal{M}^k, \mathbf{x}_i)P(\mathcal{M}^k|\mathbf{x}_i)}{P(Y_i|\mathbf{x}_i)} \\ &= \frac{P(Y_i|\mathcal{M}^k, \mathbf{x}_i)P(\mathcal{M}^k|\mathbf{x}_i)}{\sum_{j=1}^K P(Y_i|\mathcal{M}^j, \mathbf{x}_i)P(\mathcal{M}^j|\mathbf{x}_i)} = \frac{L_k^i}{L^i}, \end{aligned} \tag{10}$$

where L^i is the likelihood of the complete MHMM for subject i, and L_k^i is the likelihood of cluster k for subject i. These are straightforwardly computed from forward probabilities. Posterior cluster probabilities are used e.g. for computing classification tables.

2.4. Important special cases

The hidden Markov model is not the only important special case of the mixture hidden Markov model. Here we cover some of the most important special cases that are included in the **seqHMM** package.

Markov model

The Markov model (MM) is a special case of the HMM, where there is no hidden structure. It can be regarded as an HMM where the hidden states correspond to the observed states perfectly. Now the number of hidden states matches the number of the observed states. The emission probability $P(y_{it}) = 1$ if $z_t = y_{it}$ and 0 otherwise, i.e., the emission matrices are identity matrices. Note that for building Markov models the data must be in a single-channel format.

Mixture Markov model

Like MM, the mixture Markov model (MMM) is a special case of the MHMM, where there is no hidden structure. The likelihood of the model is now of the form

$$\begin{aligned} \log L &= \sum_{i=1}^N \log P(\mathbf{y}_i|\mathbf{x}_i, \mathcal{M}^k) = \sum_{i=1}^N \log \sum_{k=1}^K P(\mathcal{M}^k|\mathbf{x}_i)P(\mathbf{y}_i|\mathbf{x}_i, \mathcal{M}^k) \\ &= \sum_{i=1}^N \log \sum_{k=1}^K P(\mathcal{M}^k|\mathbf{x}_i)P(y_{i1}|\mathbf{x}_i, \mathcal{M}^k) \prod_{t=2}^T P(y_{it}|y_{i(t-1)}, \mathbf{x}_i, \mathcal{M}^k). \end{aligned} \tag{11}$$

Again, the data must be in a single-channel format.

Latent class model

Latent class models (LCM) are another class of models that are often used for longitudinal research. Such models have been called, e.g., (latent) growth models, latent trajectory models, or longitudinal latent class models (Vermunt *et al.* 2008; Collins and Wugalter 1992). These models assume that dependencies between observations can be captured by a latent class, i.e., a time-constant variable which we call cluster in this paper.

The **seqHMM** includes a function for fitting an LCM as a special case of MHMM where there is only one hidden state for each cluster. The transition matrix of each cluster is now reduced to a scalar 1 and the likelihood is of the form

$$\begin{aligned} \log L &= \sum_{i=1}^{N} \log P(Y_i|\mathbf{x}_i, \mathcal{M}^k) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} P(\mathcal{M}^k|\mathbf{x}_i) P(Y_i|\mathbf{x}_i, \mathcal{M}^k) \\ &= \sum_{i=1}^{N} \log \sum_{k=1}^{K} P(\mathcal{M}^k|\mathbf{x}_i) \prod_{t=1}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, \mathcal{M}^k). \end{aligned} \tag{12}$$

For LCMs, the data can consist of multiple channels, i.e., the data for each subject consists of multiple parallel sequences. It is also possible to use **seqHMM** for estimating LCMs for non-longitudinal data with only one time point, e.g. to study multiple questions in a survey.

3. Package features

The purpose of the **seqHMM** package is to offer tools for the whole HMM analysis process from sequence data manipulation and description to model building, evaluation, and visualization. Naturally, **seqHMM** builds on other packages, especially the **TraMineR** package designed for sequence analysis. For constructing, summarizing, and visualizing sequence data, **TraMineR** provides many useful features. First of all, we use the **TraMineR**'s `stslist` class as the sequence data structure of **seqHMM**. These state sequence objects have attributes such as color palette and alphabet, and they have specific methods for plotting, summarizing, and printing. Many other **TraMineR**'s features for plotting or data manipulation are also used in **seqHMM**.

On the other hand, **seqHMM** extends the functionalities of **TraMineR**, e.g. by providing easy-to-use plotting functions for multichannel data and a simple function for converting such data into single-channel representation.

Other significant packages include the **igraph** package (Csardi and Nepusz 2006), which is used for drawing graphs of HMMs, and the **nloptr** package (Ypma, Borchers, and Eddelbuettel 2014; Johnson 2014), which is used in direct numerical optimization of model parameters. The computationally intensive parts of the package are written in C++ with the help of the **Rcpp** (Eddelbuettel and François 2011; Eddelbuettel 2013) and **RcppArmadillo** (Eddelbuettel and Sanderson 2014) packages.

In addition of using C++ for major algorithms, **seqHMM** also supports parallel computation via the OpenMP interface by dividing computations for subjects between threads. The user can choose the number of parallel threads (typically the number of cores) to use for the specific task, using the `threads` argument where available.

Table 3 shows the functions and methods available in the **seqHMM** package. The package includes functions for estimating and evaluating HMMs and MHMMs as well as visualizing data and models. There are some functions for manipulating data and models, and for simulating model parameters or sequence data given a model. In the next sections we discuss the usage of these functions more thoroughly.

As the straightforward implementation of the forward–backward algorithm poses a great risk of under- and overflow, typically forward probabilities are scaled so that there should be no

Table 1: Functions and methods in the **seqHMM** package

Usage	Functions/methods
Model construction	`build_hmm`, `build_mhmm`, `build_mm`, `build_mmm`, `build_lcm`, `simulate_initial_probs`, `simulate_transition_probs`, `simulate_emission_probs`
Model estimation	`fit_model`
Model visualization	`plot`, `ssplot`, `mssplot`
Model inference	`logLik`, `BIC`, `summary`
State inference	`hidden_paths`, `posterior_probs`, `forward_backward`
Data visualization	`ssplot`, `ssp + plot`, `ssp + gridplot`
Data and model manipulation	`mc_to_sc`, `mc_to_sc_data`, `trim_model`, `separate_mhmm`
Data simulation	`simulate_hmm`, `simulate_mhmm`

underflow. Although scaling is often sufficient for forward algorithm, it can still result in an overflow problem in the backward algorithm. This is especially true in the case of global optimization algorithms which can search infeasible areas of the parameter space. Thus, `seqHMM` also supports computation on the logarithmic scale in most of the algorithms, which further reduces the numerical unstabilities. On the other hand, as there is a need to back-transform to the natural scale during the algorithms, the log-space approach is somewhat slower than the scaling approach. Therefore, the default option is to use the scaling approach, which can be changed to the log-space approach by setting the `log_space` argument to `TRUE` e.g. in `fit_model`.

3.1. Building and fitting models

A model is first constructed using an appropriate build function. As Table 3 illustrates, several such functions are available: `build_hmm` for hidden Markov models, `build_mhmm` for mixture hidden Markov models, `build_mm` for Markov models, `build_mmm` for mixture Markov models, and `build_lcm` for latent class models.

Build functions check that the data and matrices are of the right form and create an object of class `hmm` (for HMMs and MMs) or `mhmm` (for MHMMs, MMMs, and LCMs). For the latter, covariates can be omitted or added with the usual `formula` argument using symbolic formulas familiar from e.g. the `lm` function. Even though missing observations are allowed in sequence data, covariates must be completely observed.

After a model is constructed, model parameters are estimated with the `fit_model` function. MMs, MMMs, and LCMs are handled internally as their more general counterparts, except in the case of `print` methods, where some reduntant parts of the model are not printed.

In all models, initial zero probabilities are regarded as structural zeroes and only positive probabilities are estimated. Thus it is easy to construct e.g. a left-to-right model by defining the transition probability matrix as an upper triangular matrix.

The `fit_model` function provides three estimation steps: 1) EM algorithm, 2) global DNM, and 3) local DNM. The user can call for one method or any combination of these steps, but

should note that they are performed in the above-mentioned order. At the first step starting values are based on the model object given to `fit_model`. Results from a former step are then used as starting values in the latter. Exceptions to this are some global optimization algorithms, which do not use initial values (because of this, performing just the local DNM step can lead to a better solution than global DNM with a small number of iterations).

In order to reduce the risk of being trapped in a poor local optimum, a large number of initial values should be tested. The **seqHMM** package strives to automize this. One option is to run the EM algorithm multiple times with more or less random starting values for transition or emission probabilities or both. These are called for in the `control_em` argument. Although not done by default, this method seems to perform very well as the EM algorithm is relatively fast compared to DNM.

Another option is to use the multilevel single-linkage method (MLSL) (Rinnooy Kan and Timmer 1987a,b). It draws multiple random starting values and performs local optimization from each starting point. The LDS modification uses low-discrepancy sequences instead of random numbers as starting points and should improve the convergence rate (Kucherenko and Sytsko 2005).

By default, the `fit_model` function uses the EM algorithm with a maximum of 1000 iterations and skips the local and global DNM steps. For the local step, the L-BFGS algorithm (Nocedal 1980; Liu and Nocedal 1989) is used by default. Setting `global_step = TRUE`, the function performs MSLS-LDS with the L-BFGS as the local optimizer. In order to reduce the computation time spent on non-global optima, the convergence tolerance of the local optimizer is set relatively large, so again local optimization should be performed at the final step. For DNM steps (2 and 3), any optimization method available in the **nloptr** package can be used.

There are some theoretical guarantees that the MLSL method finds all local optima in a finite number of local optimizations. Of course, it might not always succeed in a reasonable time. Also, it requires setting boundaries for the parameter space, which is not always straightforward. In DNM steps the transition, emission, and initial probabilities are estimated using unconstrained reparameterization using the softmax function (a generalization of the logistic function), but good boundaries are essential for the efficient use of the MLSL algorithm. If the boundaries are too strict, the global optimum cannot be found; if too wide, the probability of finding the global optimum is decreased. The `fit_model` function uses starting values or results from the preciding estimation step to adjust the boundaries. EM can help in setting good boundaries, but in some cases it can also lead to worse results. For finding the best solution, it is advisable to try a couple of different settings; e.g. randomized EM, EM followed by MLSL, a couple of EM iterations followed by MLSL, and only MLSL.

State and model inference

In `seqHMM`, forward and backward probabilities are computed using the `forward_backward` function, either on the logarithmic scale or in the form of scaled probabilities, depending on the argument `log_space`. Posterior probabilities are obtained from the `posterior_probs` function. In `seqHMM`, the most probable paths are computed with the `hidden_paths` function. For details of the Viterbi and the forward–backward algorithm, see e.g. Rabiner (1989).

The `seqHMM` package provides the `logLik` method for computing the log-likelihood of a model. The method returns an object of class `logLik` which is compatible with the generic information

criterion functions `AIC` and `BIC` of R. When constructing the `hmm` and `mhmm` objects via model building functions, the number of observations and the number of parameters of the model are stored as attributes `nobs` and `df` which are extracted by the `logLik` method for the computation of information criteria. The number of model parameters defined from the initial model by taking account of the parameter redundancy constraints (stemming from sum-to-one constraints of transition, emission, and initial state probabilities) and by defining all zero probabilities as structural, fixed values.

The `summary` method automatically computes some features for the MHMM, MMM, and the latent class model, e.g. standard errors for covariates and prior and posterior cluster probabilities for subjects. A `print` method for this summary shows an output of the summaries: estimates and standard errors for covariates, log-likelihood and BIC, and information on most probable clusters and prior probabilities.

3.2. Visualizing sequence data

Good graphical presentations of data and models are useful during the whole analysis process from the first glimpse into the data to the model fitting and presentation of results. The **TraMineR** package provides nice plotting options and summaries for simple sequence data, but at the moment there is no easy way of plotting multichannel data. We propose to use a so-called *stacked sequence plot* (ssp), where the channels are plotted on top of each other so that the same row in each figure matches the same subject. Figure 1 illustrates an example of a stacked sequence plot with the ten first sequences of the `biofam` data set. The code for creating the figure is shown in section 4.1.

The `ssplot` function is the simplest way of plotting multichannel sequence data in `seqHMM`. It can be used to illustrate state distributions or sequence index plots. The former is the default option, since index plots can take a lot of time and memory if data are large. Figure 2 illustrates a default plot which the user can modify in many ways (see the code in section 4.1). More examples are shown in the documentation pages of the `ssplot` function.

Another option is to define function arguments with the `ssp` function and then use previously saved arguments for plotting with a simple `plot` method. It is also possible to combine several `ssp` figures into one plot with the `gridplot` function. Figure 3 illustrates an example of such a plot showing sequence index plots for women and men (see the code in section 4.1). Sequences are ordered in a more meaningful order using multidimensional scaling scores of observations (computed from sequence dissimilarities). After defining the plot for one group, a similar plot for other groups is easily defined using the `update` function.

The `gridplot` function is useful for showing different features for the same subjects or the same features for different groups. The user has a lot of control over the layout, e.g. dimensions of the grid, widths and heights of the cells, and positions of the legends.

We also provide a function `mc_to_sc_data` for the easy conversion of multichannel sequence data into a single channel representation. Plotting combined data is often useful in addition to (or instead of) showing separate channels.

3.3. Visualizing hidden Markov models

For the easy visualization of the model structure and parameters, we propose plotting HMMs as directed graphs. Such graphs are easily called with the `plot` method, with an object of

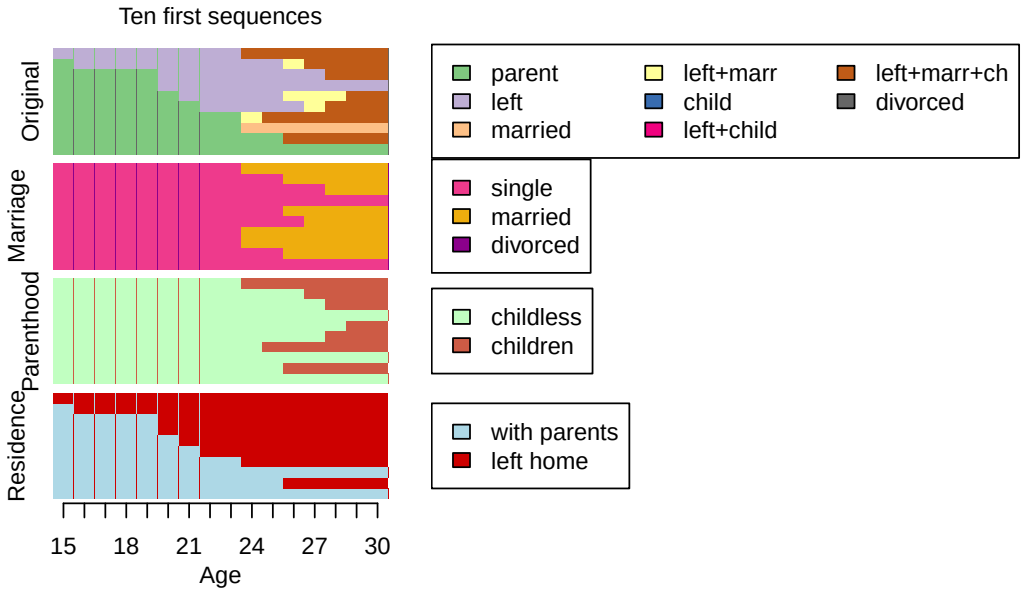


Figure 1: Stacked sequence plot of the first ten individuals in the `biofam` data plotted with the `ssplot` function. The top plot shows the original sequences, and the three bottom plots show the sequences in the separate channels for the same individuals. The sequences are in the same order in each plot, i.e., the same row always matches the same individual.

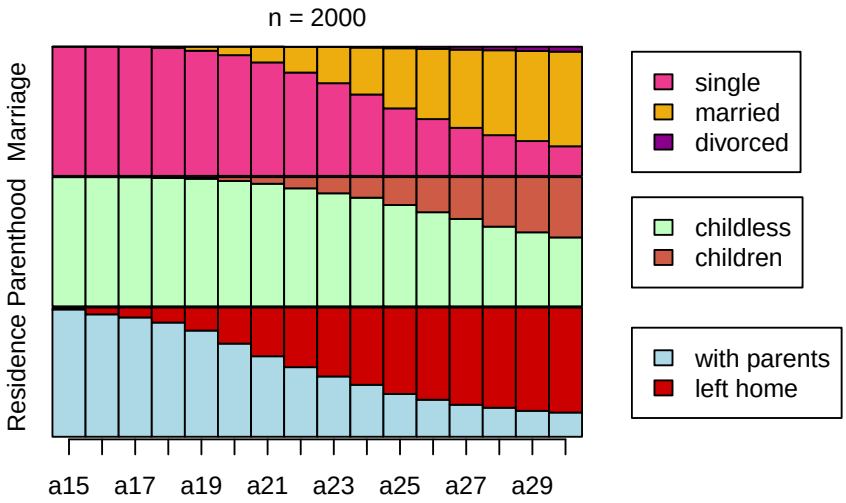


Figure 2: Stacked sequence plot of annual state distributions in the three-channel `biofam` data. This is the default output of the `ssplot` function.

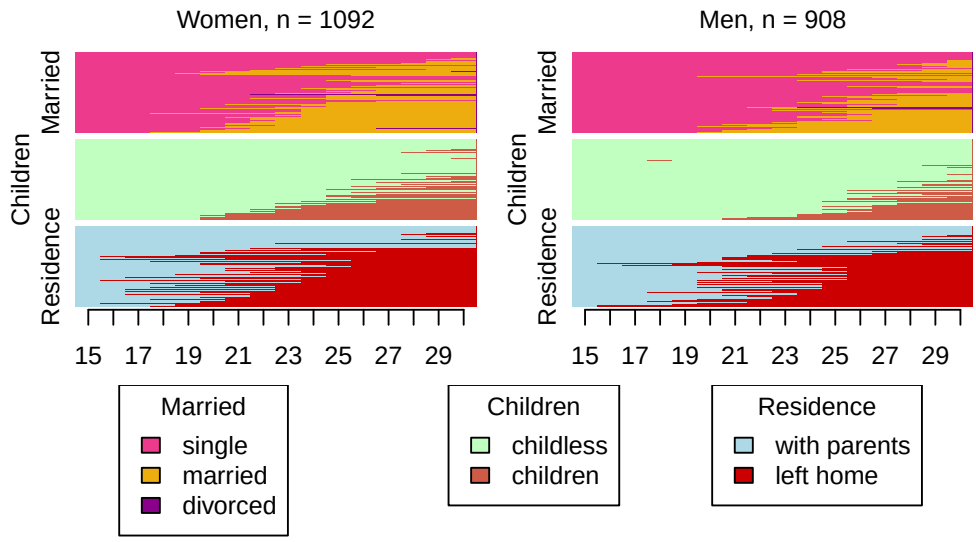


Figure 3: Showing state distribution plots for women and men in the `biofam` data. Two figures were defined with the `ssp` function and then combined into one figure with the `gridplot` function.

class `hmm` as an argument. Figure 4 illustrates a five-state HMM. The code for producing the plot is shown in section 4.4.

Hidden states are presented with pie charts as vertices (or nodes), and transition probabilities are shown as edges (arrows, arcs). By default, the higher the transition probability, the thicker the stroke of the edge. Emitted observed states are shown as slices in the pies. For gaining a simpler view, observations with small emission probabilities (less than 0.05 by default) can be combined into one category. Initial state probabilities are given below or next to the respective vertices. In the case of multichannel sequences, the data and the model are converted into a single-channel representation with the `mc_to_sc` function.

A simple default plot is easy to call, but the user has a lot of control over the layout. Figure 5 illustrates another possible visualization of the same model. The code is shown in section 4.4.

For defining the colors, the plotting functions use `colorpalette` data, which is a list of ready-made color palettes with 1–200 distinct colors. It is provided in the package, so the user can easily modify colors in the plots. See also the `RColorBrewer` package (Neuwirth 2014) for more color palettes with distinct colors. The `plot_colors` function is provided for the easy visualization of color palettes.

The `ssplot` function (see section 3.2) also accepts an object of class `hmm`. The user can easily choose to plot observations, most probable paths of hidden states, or both. The function automatically computes hidden paths if the user does not provide them.

Figure 6 shows observed sequences with the most probable paths of hidden states given the model. Sequences are sorted according to multidimensional scaling scores computed from

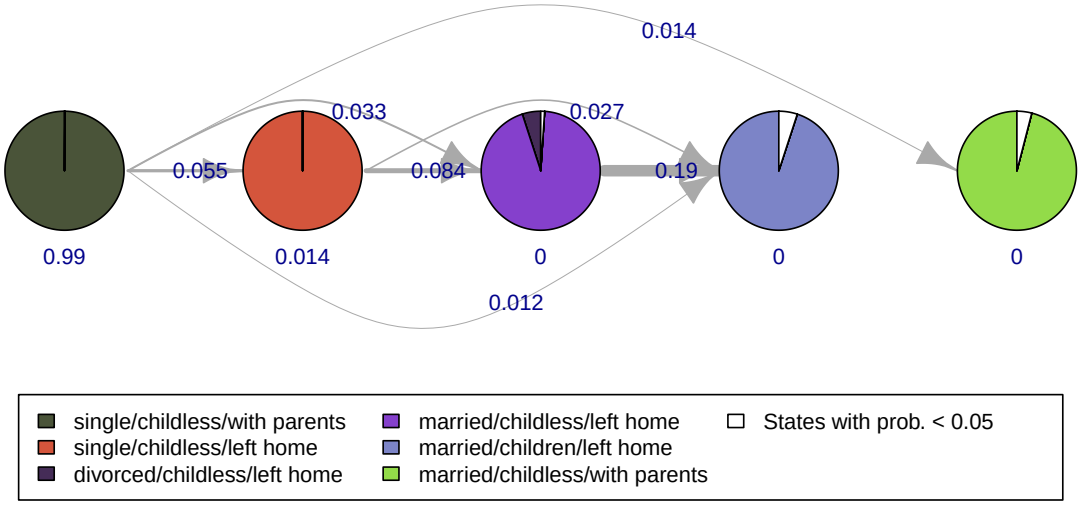


Figure 4: Illustrating a hidden Markov model as a directed graph. Pies represent five hidden states, with slices showing emission probabilities of combinations of observed states. States with emission probability less than 0.05 are combined into one slice. Edges show the transtion probabilities. Initial probabilities of hidden states are given below the pies.

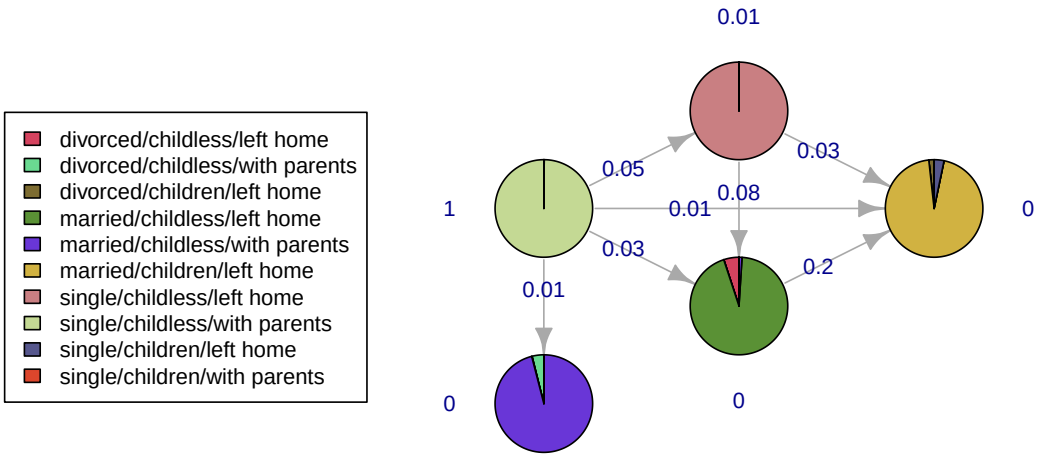


Figure 5: Another version of the hidden Markov model of Figure 4 with a different layout and modified labels, legends, and colors. All observed states are shown.

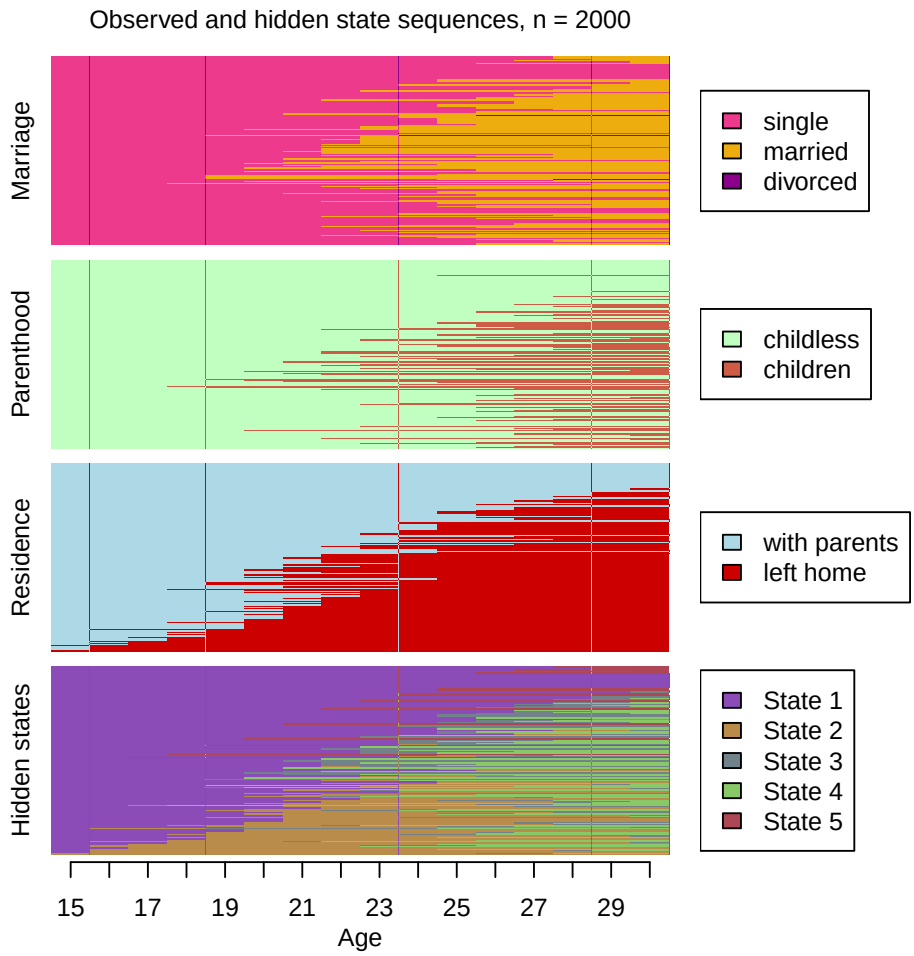


Figure 6: Using the `ssplot` function for an `hmm` object makes it easy to plot the observed sequences together with the most probable paths of hidden states given the model.

hidden paths. The code for creating the plot is shown in section 4.4.

The `plot` method works for `mhmm` objects as well. The user can choose between an interactive mode, where the model for each (chosen) cluster is plotted separately, and a combined plot with all models in one plot. The equivalent to the `ssplot` function for MHMMs is `mssplot`. It plots stacked sequence plots separately for each cluster. If the user asks to plot more than one cluster, the function is interactive by default.

4. Examples with life course data

In this section we show examples of using the **seqHMM** package. We start by constructing and visualizing sequence data, then show how HMMs are built and fitted for single-channel and multichannel data, then move on to clustering with MHMMs, and finally illustrate how

to plot HMMs.

Throughout the examples we use the same `biofam` data described in section 2.1. We use both the original single-channel data and a three-channel modification named `biofam3c`, which is included in the **seqHMM** package. For more information on the conversion, see the documentation of the `biofam3c` data.

4.1. Sequence data

Before getting to the estimation, it is good to get to know the data. We start by loading the original `biofam` data as well as the three-channel version of the same data, `biofam3c`. We convert the data into the `stslist` form with the `seqdef` function. We set the starting age at 15 and set the order of the states with the `alphabet` argument (for plotting). Colors of the states can be modified and stored as an attribute in the `stslist` object – this way the user only needs to define them once.

```
R> library("seqHMM")
R>
R> data("biofam", package = "TraMineR")
R> biofam_seq <- seqdef(
+    biofam[, 10:25], start = 15,
+    labels = c("parent", "left", "married", "left+marr", "child",
+               "left+child", "left+marr+ch", "divorced"))
R>
R> data("biofam3c")
R> marr_seq <- seqdef(biofam3c$married, start = 15,
+    alphabet = c("single", "married", "divorced"))
R> child_seq <- seqdef(biofam3c$children, start = 15,
+    alphabet = c("childless", "children"))
R> left_seq <- seqdef(biofam3c$left, start = 15,
+    alphabet = c("with parents", "left home"))
R>
R> attr(marr_seq, "cpal") <- c("violetred2", "darkgoldenrod2", "darkmagenta")
R> attr(child_seq, "cpal") <- c("darkseagreen1", "coral3")
R> attr(left_seq, "cpal") <- c("lightblue", "red3")
```

Here we show codes for creating Figures 2, 1, and 3. Such plots give a good glimpse into multichannel data.

Figure 2: Plotting state distributions

We start by showing how to call the simple default plot of Figure 2 in section 3.3. By default the function plots state distributions (`type = "d"`). Multichannel data are given as a list where each component is an `stslist` corresponding to one channel. If names are given, those will be used as labels in plotting.

```
R> ssplot(list("Marriage" = marr_seq, "Parenthood" = child_seq,
+    "Residence" = left_seq))
```

Figure 1: Plotting sequences

Figure 1 with the whole sequences requires modifying more arguments. We call for sequence index plots (`type = "I"`) and sort sequences according to the first channel (the original sequences), starting from the beginning. We give labels to y and x axes and modify the positions of y labels. We give a title to the plot but omit the number of subjects, which by default is printed. We set the proportion of the plot given to legends and the number of columns in each legend.

```
R> ssplot(list(biofam_seq[1:10,], marr_seq[1:10,], child_seq[1:10,],
+    left_seq[1:10,]),
+    sortv = "from.start", sort.channel = 1, type = "I",
+    ylab = c("Original", "Marriage", "Parenthood", "Residence"),
+    xtlab = 15:30, xlab = "Age", title = "Ten first sequences",
+    title.n = FALSE, legend.prop = 0.63, ylab.pos = c(1, 1.5),
+    ncol.legend = c(3, 1, 1, 1))
```

Figure 3: Plotting sequence data in a grid

For using the `gridplot` function, we first need to specify the `ssp` objects of the separate plots. Here we start by defining the first plot for women with the `ssp` function. It stores the features of the plot, but does not draw anything. We want to sort sequences according to multidimensional scaling scores. These are computed from optimal matching dissimilarities for observed sequences. Any dissimilarity method available in `TraMineR` can be used instead of the default (see the documentation of the `seqdef` function for more information). We want to use the same legends for the both plots, so we remove legends from the `ssp` objects.

Since we are going to plot to two similar figures, one for women and one for men, we can pass the first `ssp` object to the `update` function. This way we only need to define the changes and omit everything that is similar.

These two `ssp` objects are then passed on to the `gridplot` function. Here we make a 2×2 grid, of which the bottom row is for the legends, but the function can also automatically determine the number of rows and columns and the positions of the legends.

```
R> ssp_f <- ssp(
+    list(marr_seq[biofam3c$covariates$sex == "woman",],
+      child_seq[biofam3c$covariates$sex == "woman",],
+      left_seq[biofam3c$covariates$sex == "woman",]),
+    type = "I", sortv = "mds.obs", withlegend = FALSE,
+    title = "Women", ylab.pos = c(1, 2, 1),
+    ylab = c("Married", "Children", "Residence"), xtlab = 15:30)
R>
R> ssp_m <- update(ssp_f, title = "Men",
+    x = list(marr_seq[biofam3c$covariates$sex == "man",],
+      child_seq[biofam3c$covariates$sex == "man",],
+      left_seq[biofam3c$covariates$sex == "man",]))
R>
```

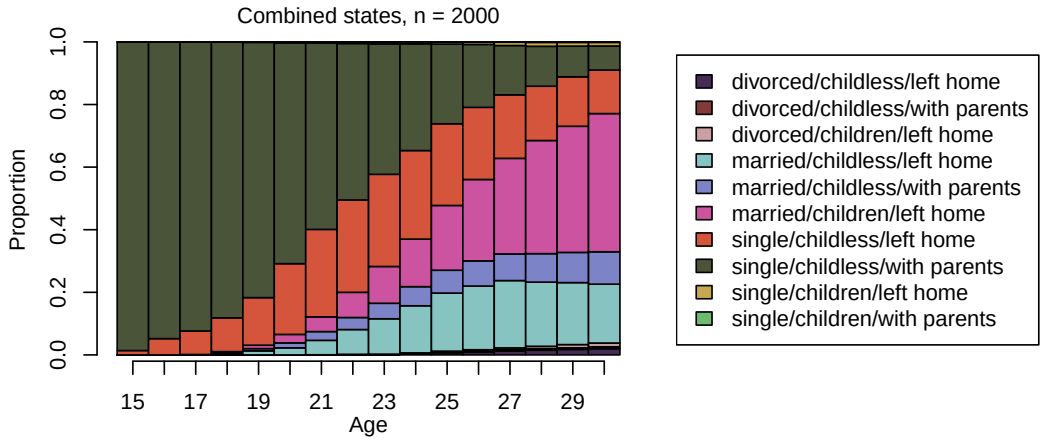


Figure 7: Three-channel `biofam3c` data converted into single-channel data.

```
R> gridplot(list(ssp_f, ssp_m), ncol = 2, nrow = 2, byrow = TRUE,
+    legend.pos = "bottom", legend.pos2 = "top", row.prop = c(0.65, 0.35))
```

Figure 7: Converting multichannel data to single-channel

When working with multiple channels, it is useful to look at the combined data as well. The `mc_to_sc_data` converts the data into a single-channel representation. At each time point of each subject, the states in each channel are combined into one. Note that here the number of combined observations (10 states) is larger than in the original data (8 states), because we have split the original divorced state into three.

Also single-channel data can be plotted with the `ssplot` function. Figure 7 illustrates the state distributions of the combined data. Here we ask to show the y-axis, which by default is omitted for gaining a less cluttered output in stacked plots.

```
R> sc_data <- mc_to_sc_data(list(marr_seq, child_seq, left_seq))
R>
R> ssplot(sc_data, type = "d", ylab = "Proportion", yaxis = TRUE,
+    xtlab = 15:30, xlab = "Age", title = "Combined states",
+    legend.prop = 0.4)
```

4.2. Hidden Markov models

We start by showing how to fit an HMM for single-channel `biofam` data.

First we set starting values for initial, transition, and emission probabilities. Here the hidden states are regarded as more general life stages, during which individuals are more likely to meet certain observable life events. We expect that the life stages are somehow related to age, so constructing starting values from the observed state frequencies by age group seems like an option worth a try (these are easily computed using the `seqstatf` function in **TraMineR**).

We construct a model with four hidden states using age groups 15–18, 19–21, 22–24, 25–27 and 28–30.

The `fit_model` function uses the probabilities given by the initial model as starting values when estimating the parameters. Only positive probabilities are estimated; zero values are fixed to zero. Thus, the amount of 0.1 is added to each value in case of zero-frequencies in some categories (at this point we do not want to fix any parameters to zero). Each row is divided by its sum, so that the row sums equal to 1.

```
R> sc_init <- c(0.9, 0.06, 0.02, 0.01, 0.01)
R>
R> sc_trans <- matrix(
+    c(0.80, 0.10, 0.05, 0.03, 0.02,
+      0.02, 0.80, 0.10, 0.05, 0.03,
+      0.02, 0.03, 0.80, 0.10, 0.05,
+      0.02, 0.03, 0.05, 0.80, 0.10,
+      0.02, 0.03, 0.05, 0.05, 0.85),
+    nrow = 5, ncol = 5, byrow = TRUE)
R>
R> sc_emiss <- matrix(NA, nrow = 5, ncol = 8)
R> sc_emiss[1,] <- seqstatf(biofam_seq[, 1:4])[, 2] + 0.1
R> sc_emiss[2,] <- seqstatf(biofam_seq[, 5:7])[, 2] + 0.1
R> sc_emiss[3,] <- seqstatf(biofam_seq[, 8:10])[, 2] + 0.1
R> sc_emiss[4,] <- seqstatf(biofam_seq[, 11:13])[, 2] + 0.1
R> sc_emiss[5,] <- seqstatf(biofam_seq[, 14:16])[, 2] + 0.1
R> sc_emiss <- sc_emiss / rowSums(sc_emiss)
```

The model is initialized with the `build_hmm` function. It checks that the data and matrices are of the right form and creates an object of class `hmm`. Markov models are constructed in a similar way using the `build_mm` function, only emission probabilities are omitted.

```
R> sc_initmod <- build_hmm(observations = biofam_seq, initial_probs = sc_init,
+    transition_probs = sc_trans, emission_probs = sc_emiss)
```

We then use the `fit_model` function for parameter estimation. Here we estimate the model using the default options of the EM step.

```
R> sc_fit <- fit_model(sc_initmod)
```

The fitting function returns the estimated model, its log-likelihood, and information on the optimization steps.

```
R> sc_fit$logLik

[1] -16781.99
```

```
R> sc_fit$model

Initial probabilities :
State 1 State 2 State 3 State 4 State 5
  0.986   0.000   0.014   0.000   0.000

Transition probabilities :
         to
from      State 1 State 2 State 3 State 4 State 5
  State 1   0.786   0.175  0.0391 0.00000  0.0000
  State 2   0.000   0.786  0.0751 0.07568  0.0631
  State 3   0.000   0.000  0.8898 0.08342  0.0267
  State 4   0.000   0.000  0.0000 0.78738  0.2126
  State 5   0.000   0.000  0.0000 0.00136  0.9986

Emission probabilities :
           symbol_names
state_names 0 1       2     3       4      5     6      7
    State 1 1 0 0.00000 0.000 0.00000 0.0000 0.000 0.0000
    State 2 1 0 0.00000 0.000 0.00000 0.0000 0.000 0.0000
    State 3 0 1 0.00000 0.000 0.00000 0.0000 0.000 0.0000
    State 4 0 0 0.00195 0.992 0.00581 0.0000 0.000 0.0000
    State 5 0 0 0.21508 0.000 0.00000 0.0246 0.713 0.0474
```

As a multichannel example we fit a 5-state model for the 3-channel data. Emission probabilities are now given as a list of three emission matrices, one for each channel. The `alphabet` function from the **TraMineR** package can be used to check the order of the observed states – the same order is used in the `build` functions. Here we construct a left-to-right model where transitions to earlier states are not allowed, so the transition matrix is upper-triangular. This seems like a valid option from a life-course prespective. Also, in the previous single-channel model of the same data the transition matrix was estimated almost upper triangular. We also give names for channels – these are used when printing and plotting the model.

We estimate model parameters using the local step with the default L-BFGS algorithm using parallel computation with 4 threads.

```
R> mc_init <- c(0.9, 0.05, 0.02, 0.02, 0.01)
R>
R> mc_trans <- matrix(
+    c(0.80, 0.10, 0.05, 0.03, 0.02,
+      0,    0.90, 0.05, 0.03, 0.02,
+      0,       0, 0.90, 0.07, 0.03,
+      0,       0,    0, 0.90, 0.10,
+      0,       0,    0,    0,    1),
+    nrow = 5, ncol = 5, byrow = TRUE)
R>
R>
```

```
R> mc_emiss_marr <- matrix(
+    c(0.90, 0.05, 0.05,
+      0.90, 0.05, 0.05,
+      0.05, 0.90, 0.05,
+      0.05, 0.90, 0.05,
+      0.30, 0.30, 0.40),
+    nrow = 5, ncol = 3, byrow = TRUE)
R>
R> mc_emiss_child <- matrix(
+    c(0.9, 0.1,
+      0.9, 0.1,
+      0.1, 0.9,
+      0.1, 0.9,
+      0.5, 0.5),
+    nrow = 5, ncol = 2, byrow = TRUE)
R>
R> mc_emiss_left <- matrix(
+    c(0.9, 0.1,
+      0.1, 0.9,
+      0.1, 0.9,
+      0.1, 0.9,
+      0.5, 0.5),
+    nrow = 5, ncol = 2, byrow = TRUE)
R>
R> mc_initmod <- build_hmm(
+    observations = list(marr_seq, child_seq, left_seq),
+    initial_probs = mc_init, transition_probs = mc_trans,
+    emission_probs = list(mc_emiss_marr, mc_emiss_child, mc_emiss_left),
+    channel_names = c("Marriage", "Parenthood", "Residence"))
R>
R> # For CRAN vignette: load the estimated model object for speed-up
R> data("hmm_biofam")
R> # mc_fit <- fit_model(mc_initmod, em_step = FALSE, local_step = TRUE,
R> #   threads = 4)
```

We store the model as a separate object for the ease of use and then compute BIC.

```
R> # Vignette: already loaded hmm_biofam
R> # hmm_biofam <- mc_fit$model
R> BIC(hmm_biofam)

[1] 28842.7
```

4.3. Clustering and mixture hidden Markov models

When fitting mixture hidden Markov models, the starting values are given as lists, with one component per cluster. For multichannel data, emission probabilities are given as a list of

lists. Here we fit a model for two clusters with 5 and 4 hidden states. For the cluster with five states we use the same starting values as for the multichannel HMM described earlier. Covariates are defined with the usual `formula` and `data` arguments.

Here we fit a model using 100 random restarts of the EM algorithm followed by the local L-BFGS method. Again we use parallel computation.

```
R> mc_init2 <- c(0.9, 0.05, 0.03, 0.02)
R>
R> mc_trans2 <- matrix(
+    c(0.85, 0.05, 0.05, 0.05,
+      0,    0.90, 0.05, 0.05,
+      0,       0, 0.95, 0.05,
+      0,       0,    0,    1),
+    nrow = 4, ncol = 4, byrow = TRUE)
R>
R> alphabet(marr_seq)

[1] "single"   "married"  "divorced"

R> mc_emiss_marr2 <- matrix(
+    c(0.90, 0.05, 0.05,
+      0.90, 0.05, 0.05,
+      0.05, 0.85, 0.10,
+      0.05, 0.80, 0.15),
+    nrow = 4, ncol = 3, byrow = TRUE)
R>
R> alphabet(child_seq)

[1] "childless" "children"

R> mc_emiss_child2 <- matrix(
+    c(0.9, 0.1,
+      0.5, 0.5,
+      0.5, 0.5,
+      0.5, 0.5),
+    nrow = 4, ncol = 2, byrow = TRUE)
R>
R> alphabet(left_seq)

[1] "with parents" "left home"

R> mc_emiss_left2 <- matrix(
+    c(0.9, 0.1,
+      0.5, 0.5,
+      0.5, 0.5,
+      0.5, 0.5),
```

```
+    nrow = 4, ncol = 2, byrow = TRUE)
R>
R>
R> init_mhmm <- build_mhmm(
+    observations = list(marr_seq, child_seq, left_seq),
+    initial_probs = list(mc_init, mc_init2),
+    transition_probs = list(mc_trans, mc_trans2),
+    emission_probs = list(list(mc_emiss_marr, mc_emiss_child, mc_emiss_left),
+      list(mc_emiss_marr2, mc_emiss_child2, mc_emiss_left2)),
+    formula = ~sex + birthyr, data = biofam3c$covariates,
+    cluster_names = c("Cluster 1", "Cluster 2"),
+    channel_names = c("Marriage", "Parenthood", "Residence"))
R>
R> # Vignette: One thread and less restarts
R> set.seed(1001)
R> mhmm_fit <- fit_model(
+    init_mhmm, local_step = TRUE, threads = 1,
+    control_em = list(restart = list(times = 10)))
R> mhmm <- mhmm_fit$model
```

The `summary` method automatically computes some features for an MHMM, e.g. standard errors for covariates and prior and posterior cluster probabilities for subjects. A `print` method shows some summaries of these: estimates and standard errors for covariates (see section 2.3), log-likelihood and BIC, and information on most probable clusters and prior probabilities. Parameter estimates for transitions, emissions, and initial probabilities are omitted by default. The classification table shows mean probabilities of belonging to each cluster by the most probable cluster (defined from posterior cluster probabilities). A good model should have values close to 1 on the diagonal.

```
R> summary(mhmm, conditional_se = FALSE)

Covariate effects :
Cluster 1 is the reference.

Cluster 2 :
             Estimate  Std. error
(Intercept)   99.3274    12.52341
sexwoman       0.1767     0.14137
birthyr       -0.0522     0.00646

Log-likelihood: -12965.93   BIC: 26575.01

Means of prior cluster probabilities :
Cluster 1 Cluster 2
    0.857     0.143
```

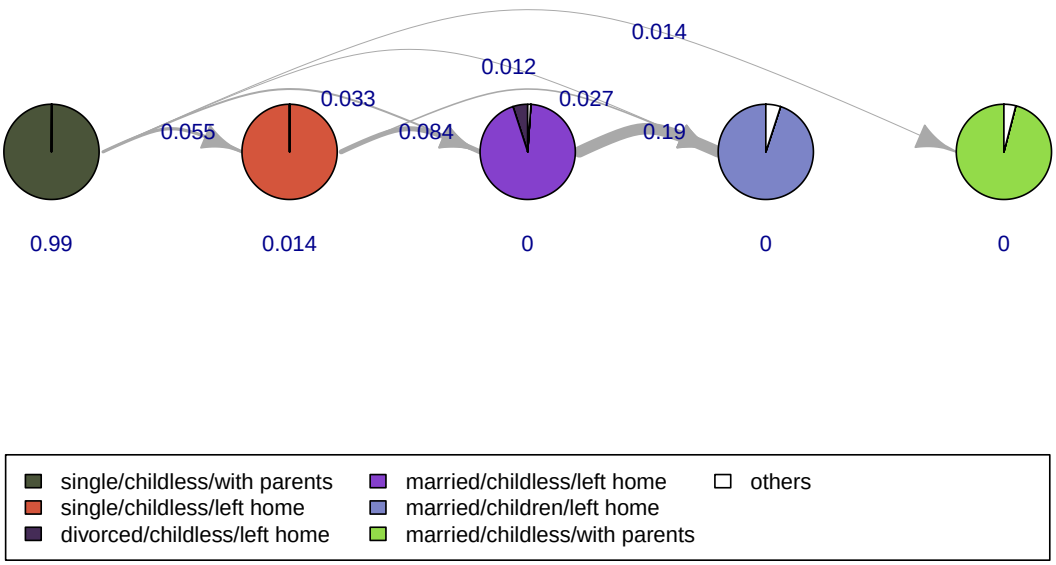


Figure 8: A default plot of a hidden Markov model.

```
Most probable clusters :
            Cluster 1  Cluster 2
count            1748        252
proportion      0.874      0.126

Classification table :
Mean cluster probabilities (in columns) by the most probable cluster (rows)

          Cluster 1 Cluster 2
Cluster 1    0.9784    0.0216
Cluster 2    0.0125    0.9875
```

4.4. Visualizing hidden Markov models

The figures in section 3.3 illustrate the five-state multichannel HMM fitted in section 4.2. A basic HMM graph is easily called with the `plot` method.

```
R> plot(hmm_biofam)
```

A simple default plot is a convenient way of visualizing the models during the analysis process, but for publishing it is often better to modify the plot to get an output that best illustrates

the structure of the model in hand. Figure 4 and Figure 5 show two variants of the same model.

Figure 4: HMM plot with modifications

In Figure 4 we draw larger vertices, control the distances of initial probabilities (vertex labels), set the curvatures of the edges, give a more descriptive label for the combined slices and give less space for the legend.

```
R> plot(hmm_biofam, vertex.size = 50, vertex.label.dist = 1.5,
+    edge.curved = c(0, 0.6, -0.8, 0.6, 0, 0.6, 0),
+    legend.prop = 0.3, combined.slice.label = "States with prob. < 0.05")
```

Figure 5: HMM plot with a different layout

Here we position the vertices using given coordinates. Coordinates are given in a two-column matrix, with x coordinates in the first column and y coordinates in the second. Arguments `xlim` and `ylim` set the lengths of the axes, and `rescale = FALSE` prevents rescaling the coordinates to the $[-1, 1] \times [-1, 1]$ interval (the default). We modify the positions of initial probabilities, fix edge widths to 1, reduce the size of the arrows in edges, position legend on top of the figure, and print labels in two columns in the legend. Parameter values are shown with one significant digit. All emission probabilities are shown regardless of their value (`combine.slices = 0`).

New colors are set from the ready-defined `colorpalette` data. The **seqHMM** package uses these palettes when determining colors automatically, e.g. in the `mc_to_sc` function. Since here there are 10 combined states, the default color palette is number 10. To get different colors, we choose the ten first colors from palette number 14.

```
R> plot(hmm_biofam, layout = matrix(c(1, 2, 2, 3,  1,
+                                      0, 0.5, -0.5, 0, -1), ncol = 2),
+    xlim = c(0.5, 3.5), ylim = c(-1.5, 1), rescale = FALSE,
+    vertex.label.pos = c("left", "top", "bottom", "right", "left"),
+    vertex.size = 50, edge.curved = FALSE, edge.width = 1,
+    edge.arrow.size = 1, withlegend = "left", legend.prop = 0.4,
+    label.signif = 1, combine.slices = 0,
+    cpal = colorpalette[[30]][c(14:5)])
```

Figure 6: `ssplot` *for an HMM object*

Plotting observed and hidden state sequences is easy with the `ssplot` function: the function accepts an `hmm` object instead of (a list of) `stslist`s. If hidden state paths are not provided, the function automatically computes them when needed.

```
R> ssplot(hmm_biofam, plots = "both", type = "I", sortv = "mds.hidden",
+    xtlab = 15:30, xlab = "Age",
+    title = "Observed and hidden state sequences")
```

4.5. Visualizing mixture hidden Markov models

Objects of class `mhmm` have similar plotting methods to `hmm` objects. The default way of visualizing a model is to plot in an interactive mode, where the model for each cluster is plotted separately. Another option is a combined plot with all models in one plot, although it can be difficult to fit several graphs and legends in one figure.

Figure 9 illustrates the MHMM fitted in section 4.3. By setting `interactive = FALSE` and `nrow = 2` we plot graphs in a grid with two rows. The rest of the arguments are similar to basic HMM plotting and apply for all the graphs.

```
R> plot(mhmm, interactive = FALSE, nrow = 2, legend.prop = 0.45,
+    vertex.size = 50, vertex.label.cex = 1.3, cex.legend = 1.3,
+    edge.curved = 0.65, edge.label.cex = 1.3, edge.arrow.size = 0.8)
```

The equivalent of the `ssplot` function for `mhmm` objects is `mssplot`. It shows data and/or hidden paths one cluster at a time. The function is interactive if more than one cluster is plotted (thus omitted here). Subjects are allocated to clusters according to the posterior cluster probabilities.

```
R> mssplot(mhmm, ask = TRUE)
```

If the user wants more control than the default `mhmm` plotting functions offer, they can use the `separate_mhmm` function to convert a `mhmm` object into a list of separate `hmm` objects. These can then be plotted as any `hmm` objects, e.g. use `ssp` and `gridplot` for plotting sequences and hidden paths of each cluster into the same figure.

5. Conclusion

Hidden Markov models are useful in various longitudinal settings with categorical observations. They can be used for accounting measurement error in the observations (e.g. drug use as in Vermunt *et al.* 2008), for detecting true unobservable states (e.g. different periods of the bipolar disorder as in Lopez 2008), and for compressing information accross several types of observations. The life course example of this paper serves as a simple illustration of such a problem, where hidden states are regarded as general life stages during which individuals are more likely to encounter certain life events.

The **seqHMM** package is designed for analyzing categorical sequences with hidden Markov models and mixture hidden Markov models, as well as their restricted variants Markov models, mixture Markov models, and latent class models. It can handle many types of data from a single sequence to multiple multichannel sequences. Covariates can be included in MHMMs to explain cluster membership. The package also offers versatile plotting options for sequence data and HMMs, and can easily convert multichannel sequence data and models into single-channel representations.

Parameter estimation in (M)HMMs is often very sensitive to starting values. To deal with that, **seqHMM** offers several fitting options with global and local optimization using direct numerical estimation and the EM algorithm.

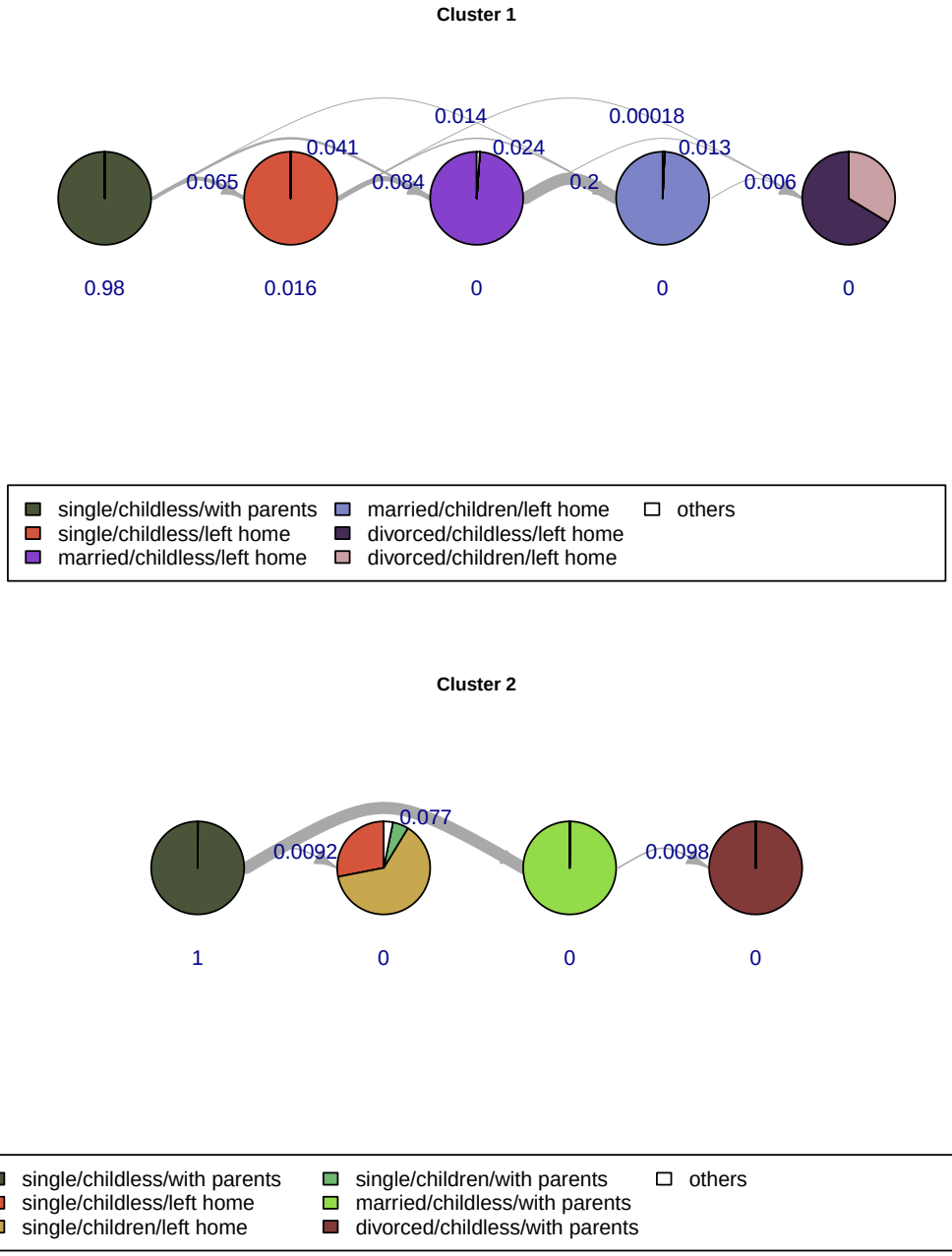


Figure 9: Plotting submodels of an MHMM with the `plot` method.

Almost all intensive computations are done in C++. The package also supports parallel computation.

Especially combined with the **TraMineR** package, **seqHMM** is designed to offer tools for the whole analysis process from data preparation and description to model fitting, evaluation, and visualization. In future we could develop MHMMs to deal with time-varying covariates and add an option to incorporate sampling weights for model estimation. Also, the computational efficiency of the restricted variants of (M)HMMs, such as latent class models, could be improved by taking account of the restricted structure of those models in EM and log-likelihood computations.

A. Notations

Symbol	Meaning
Y_i	Observation sequences of subject $i, i = 1 \ldots, N$
$\mathbf{y}_{it}$	Observations of subject i at time $t, t = 1, \ldots, T$
y_{itc}	Observation of subject i at time t in channel $c, c = 1, \ldots, C$
$m_c \in \{1, \ldots, M_c\}$	Observed state space for channel c
z_{it}	Hidden state at time t for subject i
$s \in \{1, \ldots, S\}$	Hidden state space
$A = \{a_{sr}\}$	Transition matrix of size $S \times S$
$a_{sr} = P(z_t = r \vert z_{t-1} = s)$	Transition probability between hidden states s and r
$B_c = \{b_s(m_c)\}$	Emission matrix of size $S \times M_c$ for channel c
$b_s(m_c) = P(y_{itc} = m_c \vert z_{it} = s)$	Emission probability of observed state m_c in channel c given hidden state s
$b_s(\mathbf{y}_{it}) = b_s(y_{it1}) \cdots b_s(y_{itC})$	Joint emission probability of observations at time t in channels $1, \ldots, C$ given hidden state s
$\pi = (\pi_1, \ldots, \pi_S)$	Vector of initial probabilities
$\pi_s = P(z_1 = s)$	Initial probability of hidden state s
$\hat{z}_i(Y_i)$	The most probable hidden state sequence for subject i
$\mathbf{x}_i$	Covariates of subject i
$\mathcal{M}_k, k = 1, \ldots, K$	Submodel for cluster k (latent class/cluster)
w_{ik}	Probability of cluster k for subject i
β_k	Regression coefficients for cluster k
$\{\pi^k, A^k, B_1^k, \ldots, B_C^k, \beta_k\}$	Model parameters for cluster k

B. Forward–Backward Algorithm

The *forward variable*

$$\alpha_{it}(s) = P(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{it}, z_t = s \vert \mathcal{M})$$

is the joint probability of partial observation sequences for subject i until time t and the hidden state s at time t given the model M. Let us denote $b_s(\mathbf{y}_{it}) = b_s(y_{it1}) \cdots b_s(y_{itC})$, the joint emission probability of observations at time t in channels $1, \ldots, C$ given hidden state s. The forward variable can be solved inductively:

1. Initialization
   $\alpha_{i1}(s) = \pi_s b_s(\mathbf{y}_{i1}), i = 1, \ldots, N, s = 1, \ldots, S$

2. Induction
$\alpha_{i(t+1)}(r) = \left[\sum_{s=1}^{S} \alpha_{it}(s)a_{sr}\right] b_r(\mathbf{y}_{i(t+1)}), t = 1, \ldots, T-1, r = 1, \ldots, S$

3. Termination
$P(Y_i|\mathcal{M}) = \sum_{s=1}^{S} \alpha_{iT}(s)$

The *backward variable*

$$\beta_{it}(s) = P(\mathbf{y}_{i(t+1)}, \ldots, \mathbf{y}_{iT}|z_t = s, \mathcal{M})$$

is the joint probability of the partial observation sequence after time t and hidden state s at time t given the model parameters M. (By convention we use the notion β for the backward variable. This is not to be confused with the regression coefficients in the mixture HMM.) This can also be solved inductively:

1. Initialization
$\beta_{iT}(s) = 1, i = 1, \ldots, N, s = 1, \ldots, S$

2. Induction
$\beta_{i(t+1)}(s) = \left[\sum_{r=1}^{S} a_{sr}\right] b_s(\mathbf{y}_{i(t+1)})\beta_{i(t+1)}(r), t = T-1, \ldots, 1, s = 1, \ldots, S$

C. Viterbi Algorithm

We define the score

$$\delta_{it}(s) = \max_{z_{i1}z_{i2}\cdots z_{i(t-1)}} P(z_{i1} \cdots z_{it} = s, \mathbf{y}_{i1} \cdots \mathbf{y}_{it}|\mathcal{M}), \tag{13}$$

which is the highest probability of the hidden state sequence up to time t ending in state s. By induction we have

$$\delta_{i(t+1)}(r) = \left[\max_s \delta_{it}(s)a_{sr}\right] \cdot b_r(\mathbf{y}_{i(t+1)}). \tag{14}$$

We collect the arguments maximizing equation (14) in an array $\psi_{it}(r)$ to keep track of the best hidden state sequence. The full Viterbi algorithm can be stated as follows:

1. Initialization
$\delta_{i1}(s) = \pi_s b_s(\mathbf{y}_{i1}), s = 1, \ldots, S$
$\psi_{i1}(s) = 0$

2. Recursion
$\delta_{it}(r) = \max_{s=1,\ldots,S}(\delta_{i(t-1)}(s)a_{sr})b_h(\mathbf{y}_{it})$,
$\psi_{it}(s) = \arg\max_{s=1,\ldots,S}(\delta_{i(t-1)}(s)a_{sr}), s = 1, \ldots, S; t = 2, \ldots, T$

3. Termination
$\hat{P} = \max_{s=1,\ldots,S}(\delta_{iT}(s))$
$\hat{z}_{iT} = \arg\max_{s=1,\ldots,S}(\delta_{iT}(s))$

4. Sequence backtracking
$\hat{z}_{it} = \psi_{i(t+1)}(\hat{s}_{i(t+1)}), t = T-1, \ldots, 1.$

To avoid underflow an error due to multiplying many small probabilities, the Viterbi algorithm can be computed in log space, i.e., calculating $\log(\delta_{it}(s))$.

References

Aisenbrey S, Fasang A (2010). "New Life for Old Ideas: The "Second Wave" of Sequence Analysis – Bringing the "Course" Back Into the Life Course." *Sociological Methods & Research*, **38**(3), 420–462. doi:10.1177/0049124109357532.

Bartolucci F, Pandolfi S (2015). ***LMest**: Latent Markov Models with and without Covariates*. R package version 2.1, URL http://CRAN.R-project.org/package=LMest.

Baum LE, Petrie T (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains." *The Annals of Mathematical Statistics*, **67**(6), 1554–1563. doi:10.1214/aoms/1177699147.

Blanchard P, Bühlmann F, Gauthier JA (eds.) (2014). *Advances in Sequence Analysis: Theory, Method, Applications*. Springer New York Heidelberg Dordrecht London. doi:10.1007/978-3-319-04969-4.

Collins LM, Wugalter SE (1992). "Latent Class Models for Stage-Sequential Dynamic Latent Variables." *Multivariate Behavioral Research*, **27**(1), 131–157. doi:10.1207/s15327906mbr2701_8.

Csardi G, Nepusz T (2006). "The **igraph** Software Package for Complex Network Research." *InterJournal*, **Complex Systems**(1695). URL http://igraph.org.

Durbin R, Eddy S, Krogh A, Mitchison G (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eddelbuettel D (2013). *Seamless R and C++ Integration with **Rcpp***. Springer, New York. ISBN 978-1-4614-6867-7.

Eddelbuettel D, François R (2011). "**Rcpp**: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08.

Eddelbuettel D, Sanderson C (2014). "RcppArmadillo: Accelerating R with High-Performance C++ Linear Algebra." *Computational Statistics and Data Analysis*, **71**, 1054–1063. doi:10.1016/j.csda.2013.02.005.

Elzinga CH, Studer M (2014). "Spell Sequences, State Proximities, and Distance Metrics." *Sociological Methods & Research*, pp. 3–47. doi:10.1177/0049124114540707.

Gabadinho A, Ritschard G, Müller NS, Studer M (2011). "Analyzing and Visualizing State Sequences in R with **TraMineR**." *Journal of Statistical Software*, **40**(4), 1–37. doi:10.18637/jss.v040.i04.

Gauthier JA, Widmer ED, Bucher P, Notredame C (2009). "How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data." *Sociological Methods & Research*, **38**(1), 197–231. doi:10.1177/0049124109342065.

Gauthier JA, Widmer ED, Bucher P, Notredame C (2010). "Multichannel Sequence Analysis Applied to Social Science Data." *Sociological Methodology*, **40**(1), 1–38. doi:10.1111/j.1467-9531.2010.01227.x.

Halpin B (2010). "Optimal Matching Analysis and Life-Course Data: The Importance of Duration." *Sociological Methods & Research*, **38**(3), 365–388. doi:10.1177/0049124110363590.

Himmelmann L (2010). *HMM – Hidden Markov Models*. R Package Version 1.0, URL http://CRAN.R-project.org/package=HMM.

Hollister M (2009). "Is Optimal Matching Suboptimal?" *Sociological Methods & Research*, **38**(2), 235–264. doi:10.1177/0049124109346164.

Jackson CH (2011). "Multi-State Models for Panel Data: The **msm** Package for R." *Journal of Statistical Software*, **38**(8), 1–29. doi:10.18637/jss.v038.i08.

Johnson SG (2014). *The NLopt Nonlinear Optimization Package*. URL http://ab-initio.mit.edu/nlopt.

Kucherenko S, Sytsko Y (2005). "Application of Deterministic Low-Discrepancy Sequences in Global Optimization." *Computational Optimization and Applications*, **30**(3), 297–318. doi:10.1007/s10589-005-4615-1.

Lesnard L (2010). "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns." *Sociological Methods & Research*, **38**(3), 389–419. doi:10.1177/0049124110362526.

Liu DC, Nocedal J (1989). "On the Limited Memory BFGS Method for Large Scale Optimization." *Mathematical programming*, **45**(1-3), 503–528. doi:10.1007/BF01589116.

Lopez A (2008). *Markov Models for Longitudinal Course of Youth Bipolar Disorder*. ProQuest, Ann Arbor, MI. URL http://d-scholarship.pitt.edu/6524/1/LopezAdrianaApril23.pdf.

MacDonald IL, Zucchini W (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*, volume 110. CRC Press, Boca Raton, FL.

McVicar D, Anyadike-Danes M (2002). "Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **165**(2), 317–334. doi:10.1111/1467-985X.00641.

Müller NS, Studer M, Ritschard G (2007). "Classification de Parcours de Vie à l'Aide de l'Optimal Matching." *XIVe Rencontre de la Société francophone de classification (SFC 2007)*, pp. 157–160.

Neuwirth E (2014). ***RColorBrewer*: *ColorBrewer* Palettes**. R Package Version 1.1-2, URL http://CRAN.R-project.org/package=RColorBrewer.

Nocedal J (1980). "Updating Quasi-Newton Matrices with Limited Storage." *Mathematics of computation*, **35**(151), 773–782. doi:10.1090/S0025-5718-1980-0572855-7.

O'Connell J, Højsgaard S (2011). "Hidden Semi Markov Models for Multiple Observation Sequences: The **mhsmm** Package for R." *Journal of Statistical Software*, **39**(4), 1–22. doi:10.18637/jss.v039.i04.

Rabiner L (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE*, **77**(2), 257–286. doi:10.1109/5.18626.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rinnooy Kan A, Timmer G (1987a). "Stochastic Global Optimization Methods Part I: Clustering Methods." *Mathematical programming*, **39**(1), 27–56. doi:10.1007/BF02592070.

Rinnooy Kan A, Timmer G (1987b). "Stochastic Global Optimization Methods Part II: Multi-Level Methods." *Mathematical Programming*, **39**(1), 57–78. doi:10.1007/BF02592071.

Turner R, Liu L (2014). ***hmm.discnp**: Hidden Markov Models with Discrete Non-Parametric Observation Distributions*. R Package Version 0.2-3, URL http://CRAN.R-project.org/package=hmm.discnp.

van de Pol F, Langeheine R (1990). "Mixed Markov Latent Class Models." *Sociological Methodology*, **20**, 213–247. doi:10.2307/271087.

Vermunt JK, Tran B, Magidson J (2008). *Latent Class Models in Longitudinal Research*, pp. 373–385. Handbook of Longitudinal Research: Design, Measurement, and Analysis. Elsevier, Burlington, MA.

Visser I, Speekenbrink M (2010). "**depmixS4**: An R-package for hidden Markov models." *Journal of Statistical Software*, **36**(7), 1–21. doi:10.18637/jss.v036.i07.

Ypma J, Borchers HW, Eddelbuettel D (2014). ***nloptr**: R interface to **NLopt***. R Package Version 1.0.4, URL http://CRAN.R-project.org/package=nloptr.

Affiliation:

Satu Helske
Department of Mathematics and Statistics
P.O.Box 35 (MaD)
FI-40014 University of Jyvaskyla
Finland
E-mail: satu.helske@jyu.fi
URL: http://users.jyu.fi/~samahels/

IV

Helske, S., Helske, J. and Eerola, M. (2016) Analysing complex life sequence data with hidden Markov modelling. In G. Ritschard & M. Studer (eds), *Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016*, pp 209–240.

Analysing Complex Life Sequence Data with Hidden Markov Modelling

Satu Helske, Jouni Helske, and Mervi Eerola

Abstract When analysing complex sequence data with multiple channels (dimensions) and long observation sequences, describing and visualizing the data can be a challenge. Hidden Markov models (HMMs) and their mixtures (MHMMs) offer a probabilistic model-based framework where the information in such data can be compressed into hidden states (general life stages) and clusters (general patterns in life courses).

We studied two different approaches to analysing clustered life sequence data with sequence analysis (SA) and hidden Markov modelling. In the first approach we used SA clusters as fixed and estimated HMMs separately for each group. In the second approach we treated SA clusters as suggestive and used them as a starting point for the estimation of MHMMs.

Even though the MHMM approach has advantages, we found it to be unfeasible in this type of complex setting. Instead, using separate HMMs for SA clusters was useful for finding and describing patterns in life courses.

1 Introduction

In social science applications, sequence analysis (SA) has gained more and more interest since its introduction in the mid-80s. It is now central to the life course perspective where it has been used to understand various trajectories and crucial transitions (Gauthier et al., 2014).

Satu Helske
University of Jyvaskyla, PO Box 35, FI-40014, University of Jyväskylä, Finland, e-mail: satu.helske@jyu.fi

Jouni Helske
University of Jyvaskyla, Finland

Mervi Eerola
University of Turku, Finland

Often the goal in SA is to find a typology of life sequences described as categorical time series data. Dissimilarities between each pair of sequences is determined using some criterion. Common choices have been optimal matching (McVicar and Anyadike-Danes, 2002) and Hamming distances (Hamming, 1950; Lesnard, 2010), but many modifications to these and also more fundamentally different methods have been developed (see, e.g., Aisenbrey and Fasang, 2010; Elzinga and Studer, 2014). Usually these dissimilarities are then grouped using cluster analysis such as Ward's agglomerative algorithm.

Life course data often consists of not only one sequence per subject, but multiple parallel sequences, one for each life domain of interest. We refer to *complex sequence data* for data which consist of multiple subjects and long multichannel (multidimensional) sequences.

One option for studying such data is to combine the sequences of each subject time point by time point by extending the state space of observations. This approach is simple if the number of possible combinations is moderate, but the combined state space grows rapidly as the number of domains and/or states grows. Multichannel sequence analysis (Gauthier et al., 2010) has been used for computing pairwise dissimilarities and finding clusters in complex sequence data (see, e.g., Eerola and Helske, 2016; Müller et al., 2012; Spallek et al., 2014). However, the dissimilarities are largely affected by the chosen dissimilarity metric and the cluster allocation may not be well suited to borderline cases. Also, describing, visualizing, and comparing such data is difficult. We use hidden Markov modelling for gaining a probabilistic descriptions of complex sequence data.

Hidden Markov models (HMMs) have been widely used in biological sequence analysis (Durbin et al., 1998) and speech recognition (Rabiner, 1989). Typically, the interest is in one long time series or another type of sequence. In social sciences this approach has been called latent Markov modelling. Typically, the data consists of a few measurements for multiple subjects.

Mixture hidden Markov model (MHMM) is a generalization of the HMM. There we assume that the data consists of latent subpopulations with different model structures. In the context of social sciences, the mixture hidden Markov model approach was formulated by van de Pol and Langeheine (1990) as the mixed Markov latent class model and later generalized to include time-constant and time-varying covariates by Vermunt et al. (2008) (who named the resulting model as the mixture latent Markov model, MLMM).

Multidimensional responses are included in the formulation of the MLMM but, to our knowledge, there are no empirical studies with complex life sequence data. Few studies use (M)HMMs for multichannel social science data. Helske and Helske (2016) have illustrated HMMs and MHMMs for multichannel data but do not conduct actual analyses with real data. Bartolucci et al. (2007) have studied criminal trajectories using HMMs with multiple binary sequences per subject. The data were large in the number of subjects (684 000 individuals), but sequences were short (6 age categories) and they had fixed groups (men and women) instead of latent clusters. Crayen et al. (2012) have used a hierarchical MLMM for two-channel categorical sequences to model dynamics of mood regulation of university students

during one week. The sequences were longer (56 time points) but the number of subjects is moderate (164) and they used only three states in both channels. In their hierarchical model there were two parallel latent structures; one between the days and the other within the days.

We study two approaches to analysing complex sequence data. The first is to use sequence analysis and cluster analysis for finding a few sets of clusters and then, separately for each cluster, to estimate an HMM. In this approach, hidden Markov modelling is used to compress and describe life course information within the clusters and to help choosing the number of clusters.

The second approach is to estimate a mixture model. Now the clustering is not fixed but we get a probability of each individual belonging to each cluster. For large data, estimating the MHMM with the maximum likelihood can be a complex and time-consuming task unless the set of candidate models is restricted. We study the option of using SA clusters and simple HMMs as a starting point for mixture modelling.

2 Interpretation of hidden Markov models for life sequences

One rationale behind using the HMM approach for life sequence analysis was the attempt to identify similar life course patterns based on similar hidden state trajectories. The similarity of hidden state sequences can be attributed to both external factors, which are common to groups of populations, or to internal behavioural similarities between individuals with similar features. Finding hidden dynamics is thus important for analysing and grouping life courses and also for understanding relationships between factors that are measured. The significance of hidden states in life sequence data is dependent on the chosen structure of the model. The goals of our analysis were two-folded:

1. to group individuals with similar life course patterns (clusters) and
2. to compress information in observed states across life domains to capture patterns and dynamics within a group (hidden states)

The aim was to find hidden states that compress the information across several life domains into more general life stages. 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). We restricted to left-to-right models where transitions back to previous hidden states are not possible. Such representation makes it easier to comprehend the overall dynamics within a group and is also natural from the life course perspective: even though individuals may be in similar states at different times, the second time has a different history compared to the first time. E.g., there could be a group where, at some points of their lives, individuals are married with children, then divorced for a while, and later again married with children (but with the history of having experienced a divorce).

3 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 to life courses of an age cohort born in 1955–1959. Only individuals who were born in Germany or moved there before age 14 were included.

The data consisted of monthly life statuses of 1731 individuals in three life domains (career, 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. Using the monthly time scale allowed for detecting also smaller fluctuations in life courses, e.g. recurrent transitions between unemployment and employment.

3.1 Sequences

The sequences in three life domains were constructed as follows:

Career with 4 states:

- Studying (in school, vocational training, or vocational preparation)
- Employed (full-time or part-time)
- Unemployed
- Else (parental leave, military or non-military service, voluntary work, or other gap in 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 have at least some information from each individual.

4 Hidden Markov models

In the context of hidden Markov models, observed states are determined via a Markov process of hidden states. These hidden states cannot be observed directly, but only through the sequence(s) of observations, since hidden states generate ("emit") observations on varying probabilities.

Assume we have multichannel sequence data for N individuals with C parallel sequences of length T. Naturally, the following applies for single-channel data (subjects with one sequence only) by setting $C = 1$. Let us denote the observation in channel $c,\ c = 1,\ldots,C$, of individual $i,\ i = 1,\ldots,N$, at time $t,\ t = 1,\ldots,T$, with y_{itc} and the corresponding hidden state with z_{it}. A discrete first order hidden Markov model $\mathscr{M}$ is characterized by the following parameters:

- Initial probability of hidden state s:

$$\pi_s = P(z_{i1} = s);\ s \in \{1,\ldots,S\}, \text{for all } i = 1,\ldots,N.$$

- Transition probability from hidden state s to hidden state r:

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

- Emission probability of observed state m_c in channel c given the hidden state s:

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

The (first order) Markov assumption states that the hidden state transition probability at time t only depends on the hidden state at the previous time point $t-1$:

$$P(z_{it} | z_{i(t-1)},\ldots,z_{i1}) = P(z_{it} | z_{i(t-1)}). \quad (2)$$

Also, the observed states at time t are independent of all other observations and hidden states given the hidden state at t. For multichannel sequence data, we assume the same latent structure applies for all channels, i.e., the hidden state at time t for individual i generates the observed state y_{itc} in all channels c. Observations $y_{it1},\ldots,y_{itC}$ are assumed independent of each other given the hidden state z_{it}, i.e.,

$P(\mathbf{y}_{it}|z_{it}) = P(y_{it1}|z_{it}) \cdots P(y_{itC}|z_{it})$. Fig. 1 illustrates an HMM with a hidden state sequence and two channels.

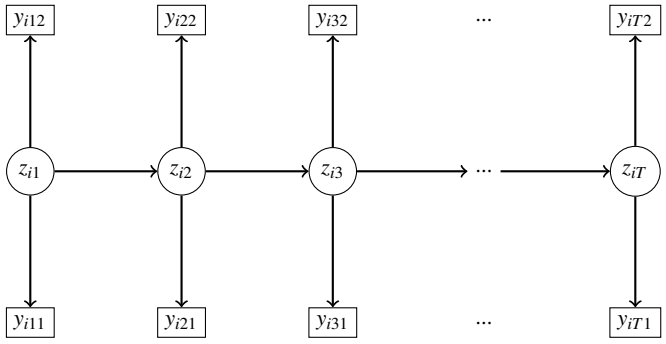


Fig. 1 Illustration of the hidden Markov model structure for two-channel sequence data for individual i with hidden states $z_{i1} \ldots, z_{iT}$ and observed states $y_{i1c} \ldots, y_{iTc}, c = 1, 2$.

The log-likelihood for the HMM is written as

$$\log L = \sum_{i=1}^{N} \log P\left(Y_i|\mathscr{M}\right), \tag{3}$$

where Y_i are the observed sequences in channels $1, \ldots, C$ for subject i and $\mathscr{M}$ describes the model and its parameters $\{\pi, A, B_1, \ldots, B_C\}$, where $A = \{a_{sr}\}$ is a matrix of transition probabilities and $B_c = \{b_s(m_c)\}$ is a matrix of emission probabilities for channel c. The probability of observation sequences for subject i given the model is

$$\begin{aligned} P(Y_i|\mathscr{M}) &= \sum_{\text{all } z} P\left(Y_i|z, \mathscr{M}\right) P\left(z|\mathscr{M}\right) \\ &= \sum_{\text{all } z} P(z_1|\mathscr{M}) P(\mathbf{y}_{i1}|z_1, \mathscr{M}) \prod_{t=2}^{T} P(z_t|z_{t-1}, \mathscr{M}) P(\mathbf{y}_{it}|z_t, \mathscr{M}) \\ &= \sum_{\text{all } z} \pi_{z_1} b_{z_1}(y_{i11}) \cdots b_{z_1}(y_{i1C}) \prod_{t=2}^{T} \left[a_{z_{t-1}z_t} b_{z_t}(y_{it1}) \cdots b_{z_t}(y_{itC})\right], \end{aligned} \tag{4}$$

where the hidden state sequences $z = (z_1, \ldots, z_T)$ take all possible combinations of values in the hidden state space $\{1, \ldots, S\}$ and where $\mathbf{y}_{it}$ are the observations of subject i at t in channels $1, \ldots, C$; π_{z_1} is the initial probability of the hidden state at time $t = 1$ in sequence z; $a_{z_{t-1}z_t}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t; and $b_{z_t}(y_{itc})$ is the probability that the hidden state of subject i at time t emits the observed state at t in channel c.

4.1 Mixture hidden Markov model

The mixture hidden Markov model is, by definition, a mixture of simple hidden Markov models. We assume that the population consists of subpopulations of individuals (latent classes or clusters) with different life patterns. Respectively, the mixture model consists of varying submodels that characterize the clusters. Transitions from one cluster to another are not allowed.

Assume that we have a set of HMMs $\mathscr{M} = \{\mathscr{M}^1, \ldots, \mathscr{M}^K\}$, where $\mathscr{M}^k = \{\pi^k, A^k, B_1^k, \ldots, B_C^k\}$ for clusters $k = 1, \ldots, K$. We denote $P(\mathscr{M}^k) = w_k$ as the prior probability that an arbitrary observation sequence is generated by the submodel $\mathscr{M}^k$ such that $\sum_{k=1}^K w_k = 1$.

The log-likelihood of the MHMM is of the form

$$\begin{aligned}
\log L &= \sum_{i=1}^{N} \log P(Y_i|\mathscr{M}) \\
&= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(\mathscr{M}^k) \sum_{\text{all } z} P\left(Y_i|z, \mathscr{M}^k\right) P\left(z|\mathscr{M}^k\right) \right] \\
&= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \sum_{\text{all } z} \pi_{z_1}^k b_{z_1}^k(y_{i11}) \cdots b_{z_1}^k(y_{i1C}) \prod_{t=2}^{T} \left[a_{z_{t-1}z_t}^k b_{z_t}^k(y_{it1}) \cdots b_{z_t}^k(y_{itC}) \right] \right].
\end{aligned} \tag{5}$$

For more detailed description of MHMMs, see Helske and Helske (2016) or Vermunt et al. (2008).

4.2 Model estimation

The log-likelihoods of (4) and (5) are efficiently calculated with the *forward–backward algorithm* (Baum and Petrie, 1966; Rabiner, 1989). A common maximum likelihood estimation method is the Baum–Welch algorithm, i.e., the expectation–maximization (EM) algorithm in the HMM context.

The Baum–Welch algorithm requires starting values for model parameters. In order to reduce the risk of being trapped in a poor local optimum, a large number of initial values should be tested. Simpler models with few parameters are fast to estimate; therefore, it is possible to fit the model numerous times with varying random starting values for finding the model with the best likelihood. When the model is large, estimation is more time-consuming and good starting values for model parameters are useful or even essential.

The most probable path of hidden states for each subject given their observations and the model can be computed using the *Viterbi algorithm* (see, e.g., Rabiner, 1989). This path maximizes the probability of $P(z|Y_i, \mathscr{M})$.

The forward–backward algorithm can also be used for computing posterior cluster probabilities (the probability that subject i belongs to a certain cluster) for MHMMs. These can be used for classifying subjects into different groups.

4.3 Model comparison

Models with the same number of parameters can be compared with the value of the log-likelihood function. For choosing between models with a different number of hidden states, we need to take account of the number of parameters.

Bayesian information criterion (BIC) is the usual criterion for comparing (M)HMMs. We define it as

$$BIC = -2\log(L) + p\log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{C}\sum_{c=1}^{C}\mathrm{I}(y_{itc}\text{ observed})\right), \tag{6}$$

where L is given in equation 3, p is the number of estimated parameters, I is the indicator function, and the summation in the logarithm is the size of the data. If data are completely observed, the summation is simplified to $N \times T$. The smaller the BIC, the better the model.

When computing the log-likelihood for the combined model with fixed SA clusters we simply sum the log-likelihoods of the cluster-wise HMMs. BIC of the combined model is determined as

$$BIC = -2 \times \sum_{k=1}^{K}\log(L_k) + \sum_{k=1}^{K} p_k \log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{C}\sum_{c=1}^{C}\mathrm{I}(y_{itc}\text{ observed})\right), \tag{7}$$

where L_k is the likelihood of the HMM of cluster k, p_k is the number of estimated parameters in the HMM for cluster k, and the summation in the logarithm is the size of the full data set.

5 Visualizing sequence data and models

Visualization is an important tool throughout the analysis process from the first glimpses into the data to presenting the results. As an example, we consider the data and the HMM for one of the preliminary clusters described “Long education and later family” (from the ten-cluster solution).

Fig. 2 illustrates a five-state HMM with the following life stages:

1. Single and (mostly) studying
2. Cohabiting, separated, or divorced; studying or employed
3. Married, studying or employed
4. Married with children, non-employed

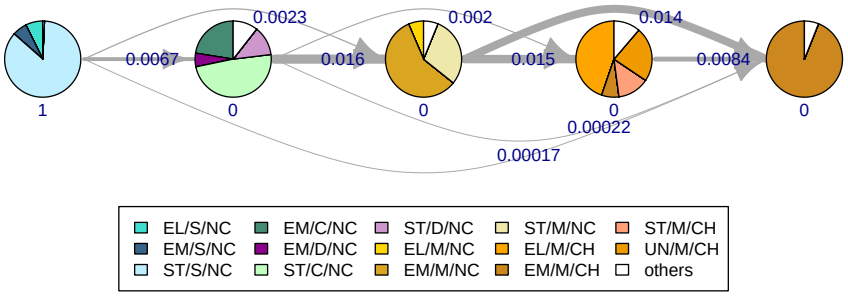


Fig. 2 Illustrating the hidden Markov model for the cluster of individuals with long education and later family. Pies present five hidden states, with slices showing the emission probabilities of combinations of observed states. States with emission probability less than 0.05 are combined into one slice for easier interpretation. The edges show the transtion probabilities – the thicker the edge, the higher the probability. Initial probabilities of the hidden states are given below the pies. The descriptions of the combined states show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren).

5. Married with children, employed

The hidden states are described by the most probable emitted observations, but there are also less probable states that are omitted from the plot for readability. E.g., the second state also emits marriages with a small probability—from the most probable hidden state paths in Fig. 3 we can see that these are marriages which end in divorce relatively fast. We could interpret that the second hidden state describes a life stage of searching for a partner before forming a long-lasting marriage.

All subjects start from the first state at age 15. At the start of the follow-up they are all single and mostly studying. The most common transition is to the second state, but the third state is quite probable also. Due to the monthly data, the transition probabilities are small—individuals usually spend years in each state.

Most individuals move to the third hidden state which describes childless marriage. It is the hidden state where individuals spend the least time on average. Transitions to the fourth and the fifth hidden state are almost as common. These both describe parenthood; some move out of workforce for a while or until the end of the follow-up, while some continue working.

6 Analysis

Estimating a large MHMM for complex sequence data can be difficult and time-consuming unless the structure of the model is fixed or known, even approxim-

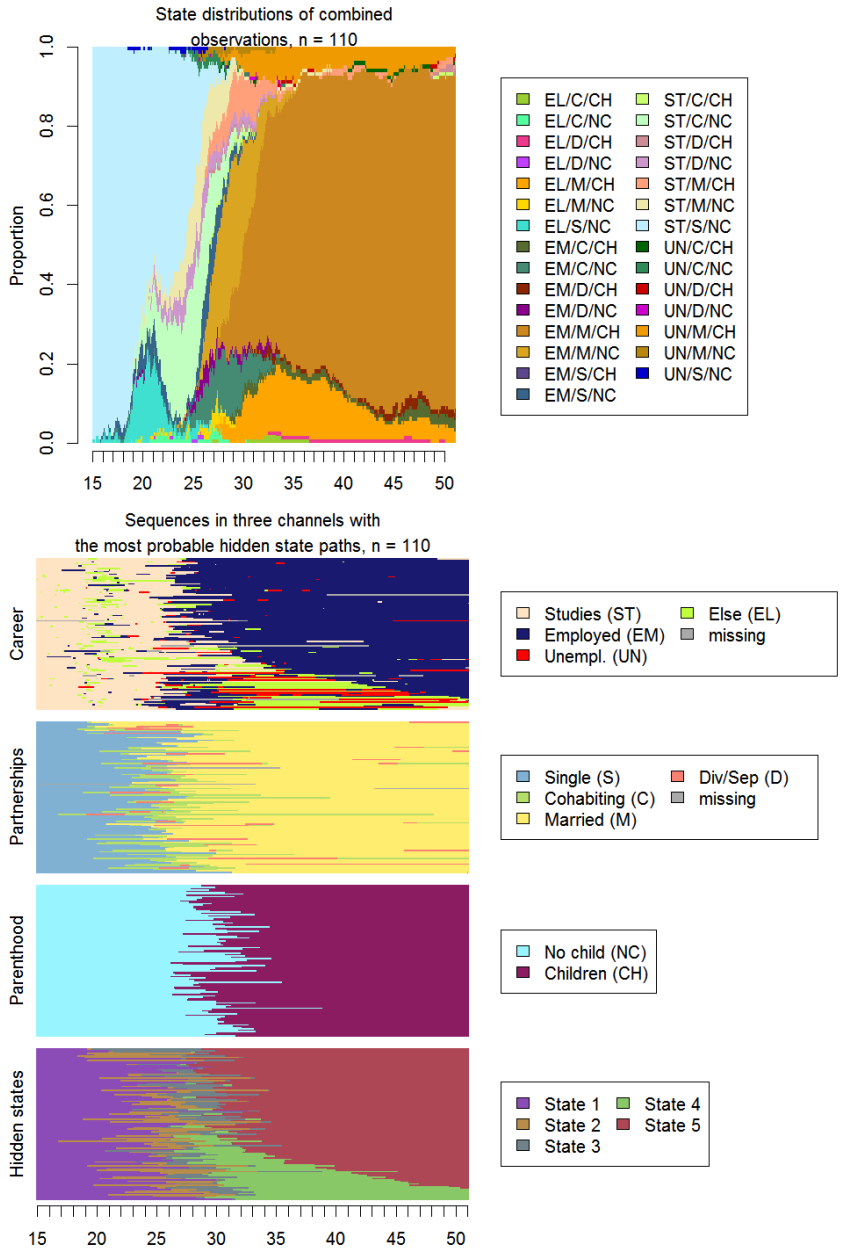


Fig. 3 State distributions of combined observations (top) and sequences of observations in each channel as well as the most probable paths of hidden states (bottom). Sequences are ordered by multidimensional scaling scores. States 1–5 correspond to the hidden states presented in Fig. 1.

ately. In other cases, the set of candidate models must be somehow restricted. In this case we had 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? As transitions were frequent in some of the trajectories and infrequent in others, it was clear that some of the clusters should contain more hidden states than others, leading to an unfeasible large number of possible model structures.

We compared two different approaches for the analysis of complex sequence data, of which both were conducted in a stepwise manner. The first two steps applied for both approaches, whereas step 3 was different (denoted as 3a and 3b). More detailed descriptions of the analysis process are given in the following sections.

1. **Sequence analysis**. Computing the dissimilarities between the subjects with the Hamming distance. Using Ward's hierarchical method for clustering individuals with similar life courses. Choosing a set of reasonable clustering solutions for preliminary analysis.
2. **Hidden Markov models.** Separately for each SA cluster, fitting simple HMMs with a different number of hidden states. Choosing the best model for each preliminary cluster.

3a. **Combined HMMs.** Constructing a combined model from separate HMMs (from step 2), keeping parameters fixed. Computing the likelihood and BIC for combined models with 7–12 clusters for determining the number of clusters. Computing the most probable path of hidden states for each individual.

3b. **Mixture hidden Markov models.** For each clustering solution (7–12 clusters), estimating an MHMM by using parameters of the corresponding HMMs (from step 2) as starting values. Computing the likelihood and BIC of the MHMMs for determining the number of clusters. Computing the most probable path of hidden states for each individual.

6.1 Step 1: Sequence analysis and preliminary clustering

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

6.1.1 Sequence dissimilarities

We compared a few dissimilarity metrics that are suitable for multichannel data: optimal matching (OM), generalized Hamming distance (HAM), and dynamic Hamming distance (DHD) (Lesnard, 2010). We chose the generalized Hamming distance with theory-driven substitution costs (see Table 1). The metric compares observed states time point by time point and gives a cost for mismatches. It generally works relatively well in a problem where timing is important and also here resulted in meaningful clusters with high goodness-of-fit (see Sect. 6.1.2).

Table 1 Substitution costs for Hamming distances.

Career status	→ ST	→ EM	→ UN	→ EL	→ *
Studying (S) →	0	3	2	1	0
Employed (EM) →	3	0	2	2	0
Unemployed (UN) →	2	2	0	1	0
Else (EL) →	1	2	1	0	0
Missing (*) →	0	0	0	0	0

Partnership status	→ S	→ C	→ M	→ D	→ *
Single (S) →	0	2	2	3	0
Cohabiting (C) →	2	0	1	2	0
Married (M) →	2	1	0	2	0
Divorced/sep. (D) →	3	2	2	0	0
Missing (*) →	0	0	0	0	0

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

6.1.2 Cluster analysis

Ward's method was chosen for clustering since 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). We chose six clustering solutions with 7–12 clusters for further examination. The choice was based on the dendrogram and interpretability of the clusters. Ward's method is agglomerative, so when two smaller clusters are merged, all other clusters remain the same. This means that within the six sets of clustering results there were only $7+2+2+2+2+2=17$ distinct clusters (see Fig. 4 for an illustration).

Table 2 shows the goodness-of-fit statistics for different clustering results and dissimilarity metrics, as measured by the proportion of the variation explained by the clusters (pseudo coefficient of determination (R^2); see Studer et al., 2011). Here, generalized Hamming distances resulted in meaningful clusters with a relatively high goodness-of-fit. OM resulted in clusters with as high goodness-of-fit while DHD resulted in somewhat lower values of R^2 (though not by much). OM clusters were similar to HAM clusters in many ways but had more variation in the timings of first transitions into employment, partnerships, and parenthood.

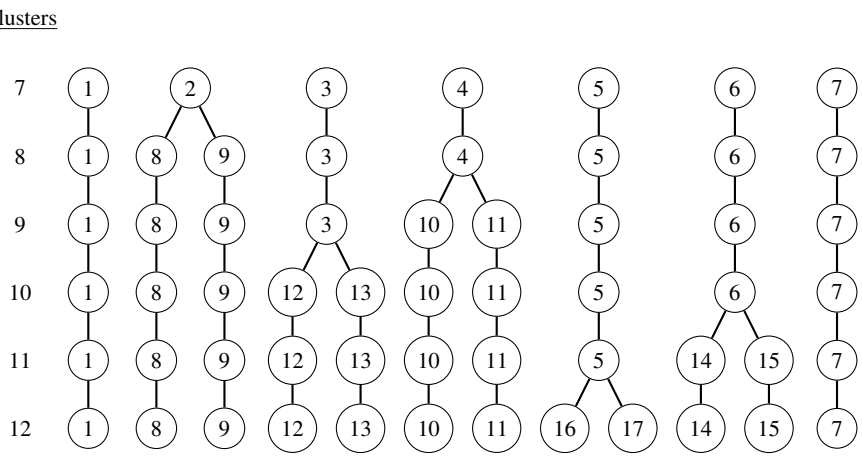


Fig. 4 Clustering structure for Ward's agglomerative method shown for six sets of clustering results with 7–12 clusters.

Table 2 Proportion of variation covered by 7–12 clusters. Clustering was based on different dissimilarity metrics; generalized Hamming distance (HAM), optimal matching (OM), and dynamic Hamming distance (DHD)).

Clusters	HAM	OM	DHD
7	0.38	0.38	0.35
8	0.40	0.40	0.37
9	0.42	0.42	0.38
10	0.43	0.43	0.40
11	0.44	0.44	0.41
12	0.44	0.45	0.42

6.2 Step 2: Simple hidden Markov models for clusters

At the next step, we estimated five HMMs with 4–9 hidden states separately for each of the 16 clusters—fewer hidden states for simpler clusters, more for more complex ones. Since the goal was to find life stages between adolescence and middle age, having too few or too many hidden states was not plausible nor interpretational.

6.2.1 Model estimation

We set starting values for parameters by determining candidate hidden states from observed data and re-estimated the model numerous times by altering these values as follows. At first, we estimated the model 10,000 times with a large variation in starting values. For each re-estimation step we added noise from the $N(0, 0.3^2)$ distribution to the the original starting values (with proper scaling and correction of

signs). The aim of this estimation was to broadly explore the parameter space and to get closer to the global maximum.

To make sure that we were at or near the global optimum, we re-estimated the model by using the model with the highest likelihood as a stating point, now adding noise from the $N(0, 0.15^2)$ distribution. If the model with the highest likelihood was found only a few times, similar estimation was repeated (again using the best model as the new starting point) in order to be fairly certain to have found the global optimum. For clusters with fewer members and models with fewer hidden states, the first estimation step was often enough for finding the (assumed) global maximum.

6.2.2 Model comparison

For each cluster, the HMMs with a different number of hidden states were compared to find the best model to use in the mixture models. 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. BIC has been proven consistent for ergodic stationary HMMs (Whiting and Pickett, 1988), but not to left-to-right HMMs. Here, also BIC consistently chose models with more hidden states and clusters than is interpretational or plausible.

A likely reason for poor performance of information criteria in this problem was that we were comparing models which all were considerably simple compared to the complexity of real life. The goal was to simplify and describe the overall patterns and dynamics in life trajectories, not to find data-generating models.

However, we did 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 were not as profitable. In addition to BIC, the choice of the number of hidden states was based on interpretability of the model and the prevalence of an additional hidden state in the most probable hidden state paths—if a hidden state was "visited" only rarely it was regarded as unnecessary.

6.3 Step 3 a: Combined HMMs

At this step we used the separate cluster-specific HMMs to construct combined models with 7–12 clusters. For each combined model, we computed the likelihood and BIC to determine the best number of clusters.

The combined model with the smallest BIC was used for determining the best number of clusters. Given the best clustering, we computed the most probable paths of hidden states for each individual.

6.4 Step 3 b: Mixture hidden Markov models

At this step we constructed six MHMMs with 7–12 clusters. We used the estimated parameters of respective cluster-wise HMMs as starting values for mixture models. To avoid non-structural zeros in starting values, we added a small amount of 0.001 to each starting value (with proper scaling). We estimated models in a similar manner to the previous step, by using randomized starting values—first with a larger noise and, after getting closer to the optimum, again with a smaller noise.

6.5 Software

Analyses were conducted with the R software (R Core Team, 2015) by using packages TraMineR (Gabadinho et al., 2011) for sequence analysis, cluster (Maechler et al., 2015) for cluster analysis, and seqHMM (Helske and Helske, 2016) for hidden Markov modelling.

7 Results

The number of hidden states per cluster varied between six and eight. We applied both the combined model and the mixture model approach for describing data and determining the best number of clusters.

7.1 Combined model approach

Table 3 shows the BICs for models with 7–12 clusters. The model with eight clusters resulted in smallest BIC (even the highest likelihood) and was chosen as the best model. The model with seven clusters was almost as good; the only difference was that the two childless clusters (see Fig. 6) were combined into one.

Table 3 Number of parameters, log-likelihood, and BIC for combined models with 7–12 clusters. The smallest value of BIC is shown in bold.

Clusters	Parameters	Log-likelihood	BIC
7	533	−369075.7	745059.4
8	**595**	**−364825.9**	**743368.2**
9	643	−370746.2	755208.7
10	705	−368985.0	751686.5
11	767	−368977.5	751671.5
12	800	−373550.3	760817.0

Fig. 5 and Fig. 6 illustrate the HMM structure for each of the eight clusters. More detailed visualizations with observed sequences and most probable hidden state paths are shown in the Appendix.

The clusters were well separated from each other by the timing and occurrence of career and family states. The two largest clusters were characterized by (mostly) short education and family. They differed in the timing of partnership and parenthood transitions which occurred either earlier in life (cluster A with 461 members of which 59% were females) or later (cluster B, 403 members, 54% males) The third largest cluster (cluster C, 266 members, 68% males) mostly consisted of individuals with long education and later family. Another cluster with early family transitions (cluster D, 159 members, 96% females) 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 timing of the partnerships; the larger cluster (cluster E, 177 members, 51% males) had earlier first partnerships while in the smaller cluster (cluster F, 116 members, 59% males) partnerships were delayed or omitted altogether.

The two smallest clusters consisted of single parents (cluster G, 47 individuals, 72% females) or parents living divorced or separated (cluster H, 102 individuals, 61% females).

7.2 Mixture model approach

The estimation of ordinary HMMs can be challenging due to multiple local optima in likelihood surfaces, since typical parameter estimation algorithms often only find these suboptimal solutions. Therefore, multiple starting values for the estimation are needed to ensure that the global optimum is found. The same problem is even more prevalent in complex MHMM settings with a large amount of parameters and mixture components. In addition, when the structure of the model (the number of mixture components and/or hidden states) is unknown, the amount of required computing resources naturally multiplies.

Therefore, even after using allegedly reasonable starting values (from simple HMMs), parallel computation, and extensive computing resources, we were not able reach satisfactory results. With different starting values the estimation always resulted in a different solution, so finding the global optimum would have required an unfeasible amount of computing time and/or resources.

Even though we were not able to find optimal MHMMs, we did study some of the suboptimal solutions. To study the differences of SA and MHMM clusters, we estimated a mixture model by keeping the initial, transition, and emission parameters of the submodels fixed (i.e., estimating only prior cluster probabilities, later referred to as the "non-estimated MHMM"). This approach was similar to the combined model approach, but instead of keeping the cluster memberships fixed we allowed individuals to switch clusters. Each individual was assigned to the cluster with the highest posterior cluster probability given their observed sequences.

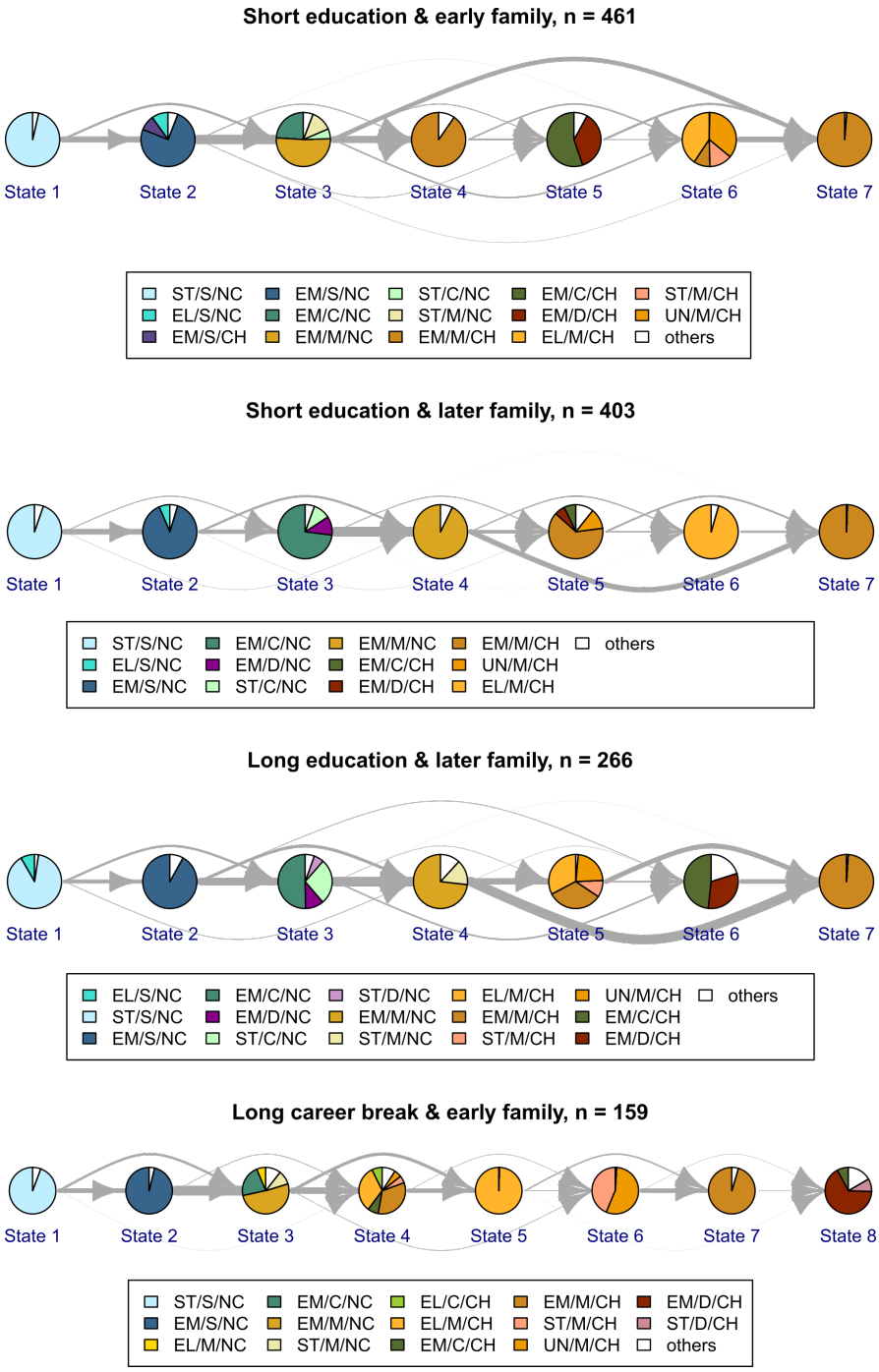


Fig. 5 HMM graphs for the eight cluster solution (clusters A–D). State abbreviations show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren).

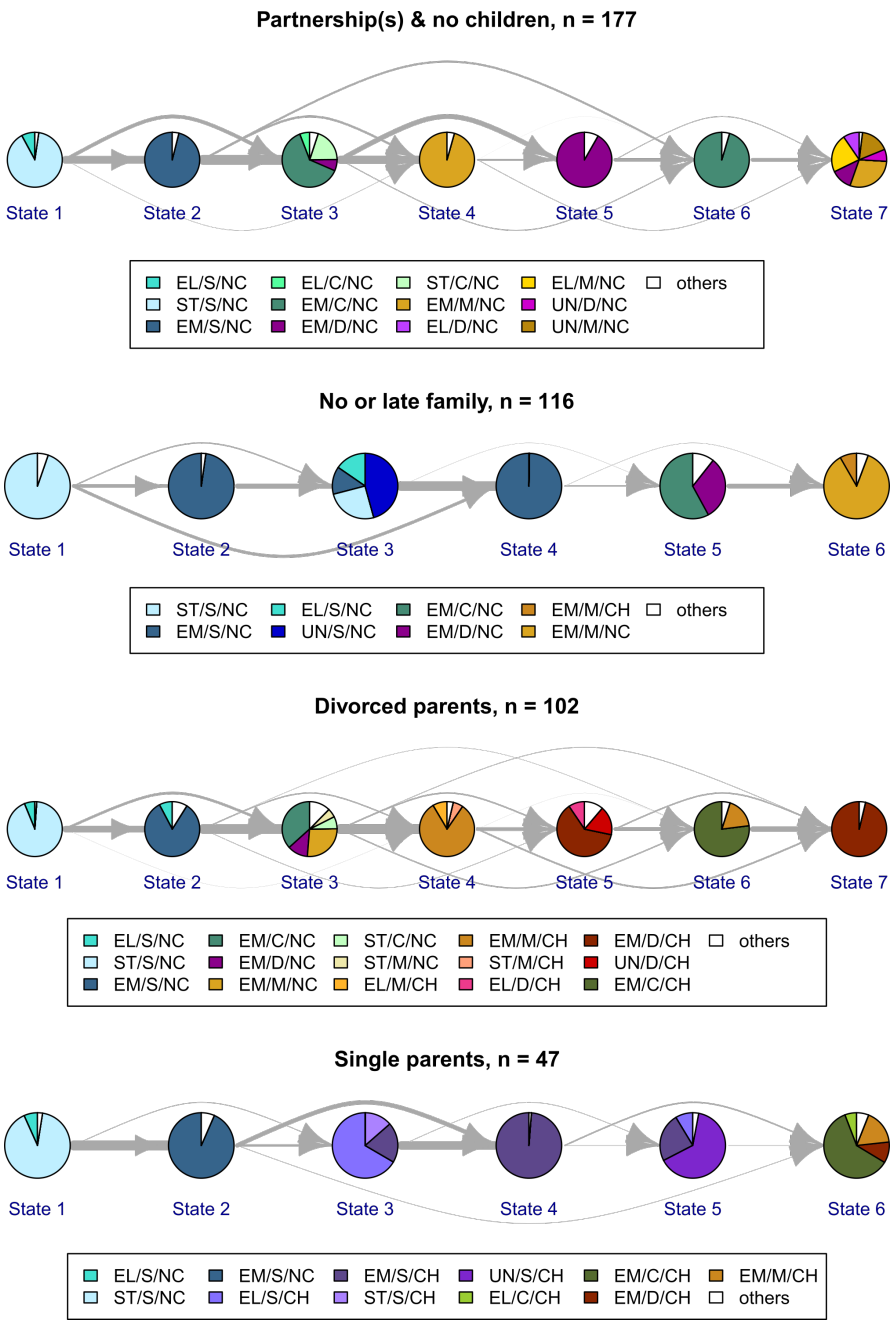


Fig. 6 HMM graphs for the eight cluster solution (clusters E–H). State abbreviations show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren).

Many individuals switched clusters compared to the SA solution (see Table 4). Some cluster were more stable; close to 90% of the members of the SA clusters "Single parents" and "Partners and no children" stayed in the same cluster in the MHMM solution. Others had many switchers; less than half of the members of SA clusters "Short education and early family" and "Long education and later family" stayed in their original clusters in the MHMM solution.

Table 4 Comparison of SA cluster memberships (left) to most probable cluster memberships from the non-estimated MHMM (top). Probabilities of staying in the same cluster are shown in bold.

	MHMM clusters								
SA clusters	A	B	C	D	E	F	G	H	Members
Short educ. & early fam. (A)	**0.32**	0.35	0.15	0.11	0.00	0.00	0.06	0.01	461
Short educ. & later fam. (B)	0.09	**0.64**	0.16	0.09	0.00	0.00	0.03	0.00	403
Long educ. & later fam, (C)	0.06	0.32	**0.43**	0.13	0.00	0.00	0.07	0.00	266
Career break & early family (D)	0.04	0.39	0.03	**0.54**	0.00	0.00	0.00	0.00	159
Partnership(s) & no child (E)	0.00	0.05	0.03	0.00	**0.87**	0.05	0.00	0.01	177
No or late family (F)	0.00	0.03	0.01	0.03	0.32	**0.60**	0.00	0.01	116
Divorced parents (G)	0.04	0.00	0.03	0.16	0.00	0.00	**0.77**	0.00	102
Single parents (H)	0.00	0.00	0.02	0.00	0.00	0.00	0.04	**0.94**	47
Number of cluster members	207	577	260	228	191	79	138	51	1731

If the MHMM parameters were estimated jointly, the differences compared to the SA clusters were even larger (we do not report the findings as we were not able to find the globally optimal model). In both MHMM approaches, the order and occurrence of states were generally more determining for the cluster memberships than the timing and duration of states. Fig. 7 illustrates this difference seen in the cluster "Short education and early family", showing the observed and hidden state sequences of members of the SA cluster and the cluster from the non-estimated MHMM. One can easily see that the variation in the timing of transitions between states (both observed and hidden) is much larger in the MHMM cluster compared to the SA cluster.

8 Discussion

When analysing complex sequence data with multiple channels, describing and visualizing the data can be a challenge. Hidden Markov models and their mixtures offer a probabilistic model-based framework where the information in data can be compressed into hidden states (different life stages) and clusters (general patterns in life courses). Hidden states can capture general life stages that include not only rather stable episodes (as the fifth hidden state of work, marriage, and children in Fig. 2) but also life stages characterized by change (as the second hidden state of searching for a partner in Fig. 2).

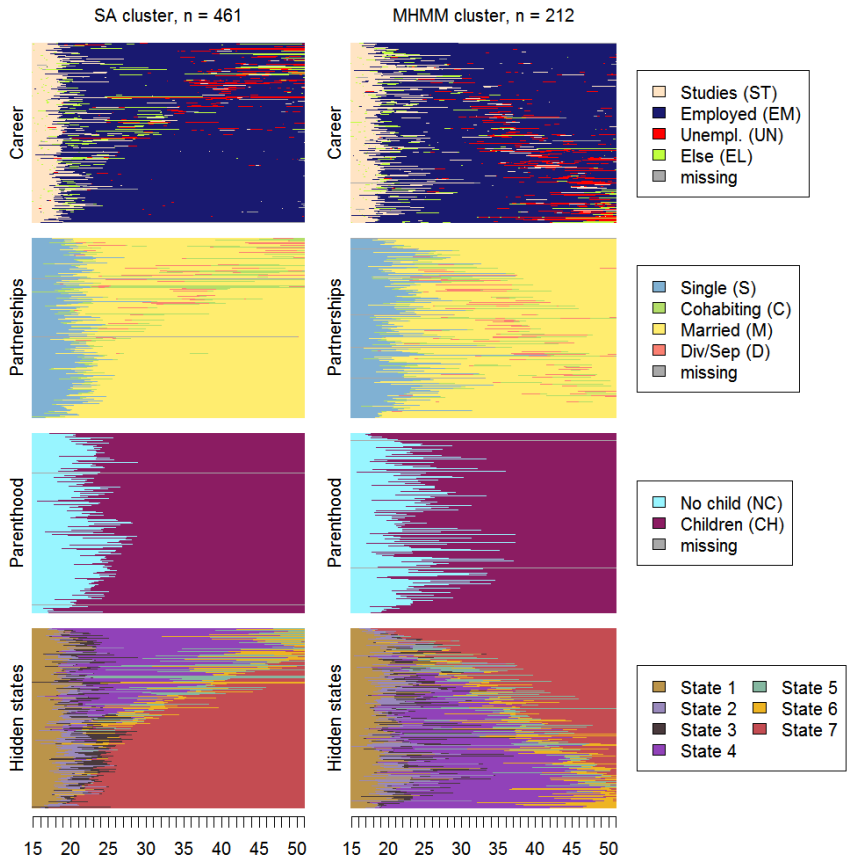


Fig. 7 Comparison of a cluster (Short education and early family) given by SA and the non-estimated MHMM.

Mixture hidden Markov modelling has several advantages. With posterior cluster probabilities we get information on certainty of the clustering for each individual and a measure for the goodness of the classification. We can also extend the model by adding covariates for explaining cluster memberships or transitions between hidden states. The MHMM approach has been used successfully in simpler settings, e.g., for accounting for measurement error and for finding clusters of "movers" and "stayers" between two hidden states.

The downsides of MHMM analysis are related to computational issues. Maximum likelihood estimation of parameters of a complex MHMM is computationally heavy. Due to multimodality of the likelihood surface we need to estimate the model numerous times with different starting values. Also, often the structure of the model (in terms of the number of hidden states and/or clusters) is not known and in general

selecting the best structure is a nontrivial task. Thus, finding the globally optimal MHMM can become unfeasible without constraining the problem.

Using sequence analysis and cluster analysis as a starting point might be useful by providing preliminary classification and by limiting the set of candidate models for a complex MHMM setting. In our study we were not able to reach satisfactory results. Our data was much more complex than in a typical MHMM analysis where sequences often come from panel data with a moderate number of measurement points. The multichannel structure, long sequences, and the relative large number of individuals in our data was a challenging combination for parameter estimation. Also, typically the number of candidate models is rather limited; when HMMs are used for accounting for measurement error, the number of hidden states is known in advance and usually the state space is very limited (e.g., poor/nonpoor or drug user/nonuser). In our study the model structure was unknown and we expected to find several clusters, each with an unknown number of hidden states.

Instead of using mixture models, we treated the SA clusters as fixed and estimated HMMs separately for each cluster (the combined model approach). With SA we found clusters that were adequately well separated by the timing and duration of life states. Hidden Markov models were used for choosing the number of clusters and for describing the overall dynamics within clusters.

Clusters found using SA and the MHMM were different in several ways. When defining sequence dissimilarities, we considered the timing of the events very important and used Hamming distances. In the MHMM analysis many individuals switched clusters; the order of states was generally more determining than their timing and duration. Further research is needed in order to determine distance metrics that result in SA clusters which capture similar features as HMMs. Metrics that weight the order of states instead of their timing such as the number of matching subsequences or the subsequence vectorial representation metric (Studer and Ritschard, 2016), might produce clustering results that are better suited for the starting point of MHMM estimation. Unfortunately, using these metrics with multichannel data is not a straightforward task.

Another topic for further research is model selection of left-to-right HMMs and MHMMs. In our study, BIC performed poorly. Further theoretical and empirical studies are needed for detecting the reasons for its failure and for discovering selection criteria that are better suited for finding parsimonious HMMs.

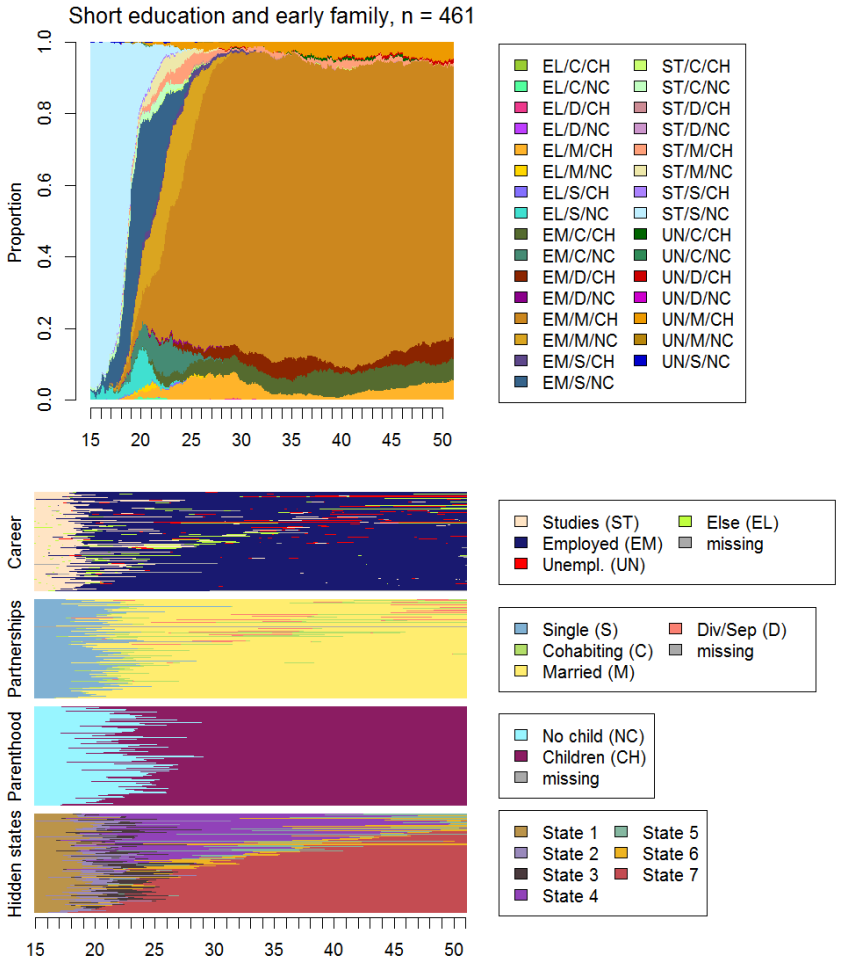
The aim of our study was to describe complex life sequence data. For that goal, SA and the combined HMM approach gave satisfactory results in a reasonable time. We were able to find meaningful clusters and to visualize their complex life course information by using stacked sequence plots, combined state distributions, and HMM graphs.

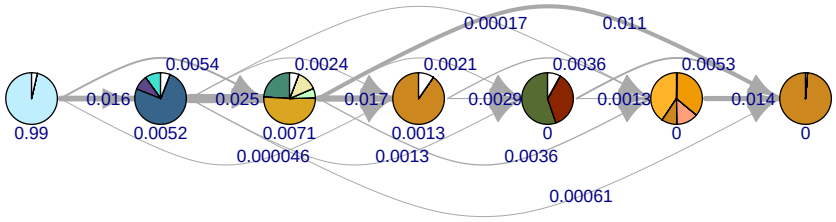
9 Acknowledgements

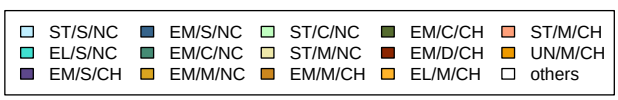
This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

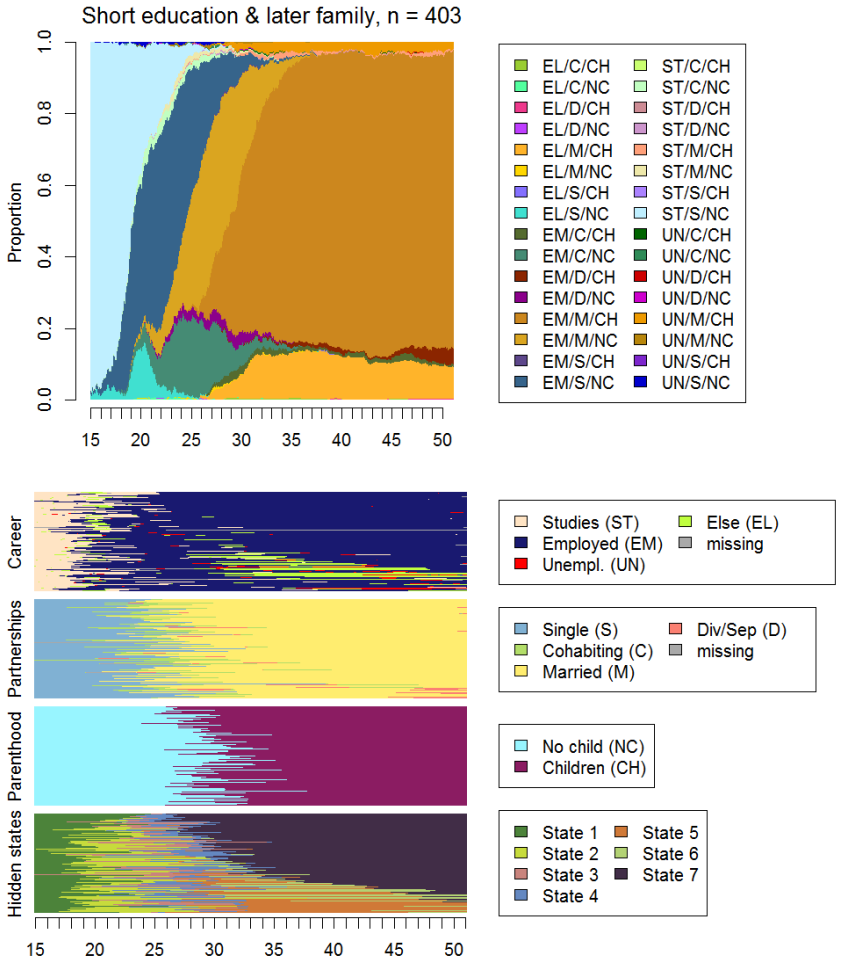
Appendix

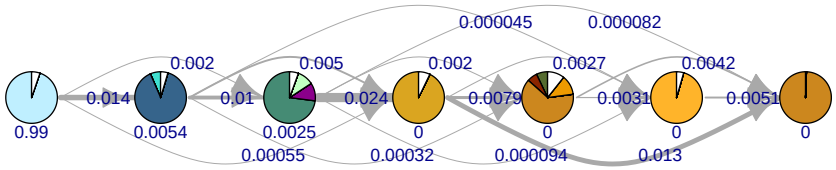
Detailed visualizations of the eight SA clusters and the respective HMMs. Figures show state distributions of combined observations at each time point (top), observed sequences in three life domains and the most probable hidden state paths given the HMM (middle), as well as HMM graphs with initial and transition probabilities (bottom). See Sect. 5 for more information on how to interpret the visualizations.

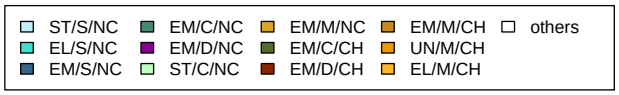

Short education and early family, n = 461
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/CH
EL/C/NC
EL/D/CH
EL/D/NC
EL/M/CH
EL/M/NC
EL/S/CH
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/CH
EM/S/NC
ST/C/CH
ST/C/NC
ST/D/CH
ST/D/NC
ST/M/CH
ST/M/NC
ST/S/CH
ST/S/NC
UN/C/CH
UN/C/NC
UN/D/CH
UN/D/NC
UN/M/CH
UN/M/NC
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
missing
State 1
State 2
State 3
State 4
State 5
State 6
State 7

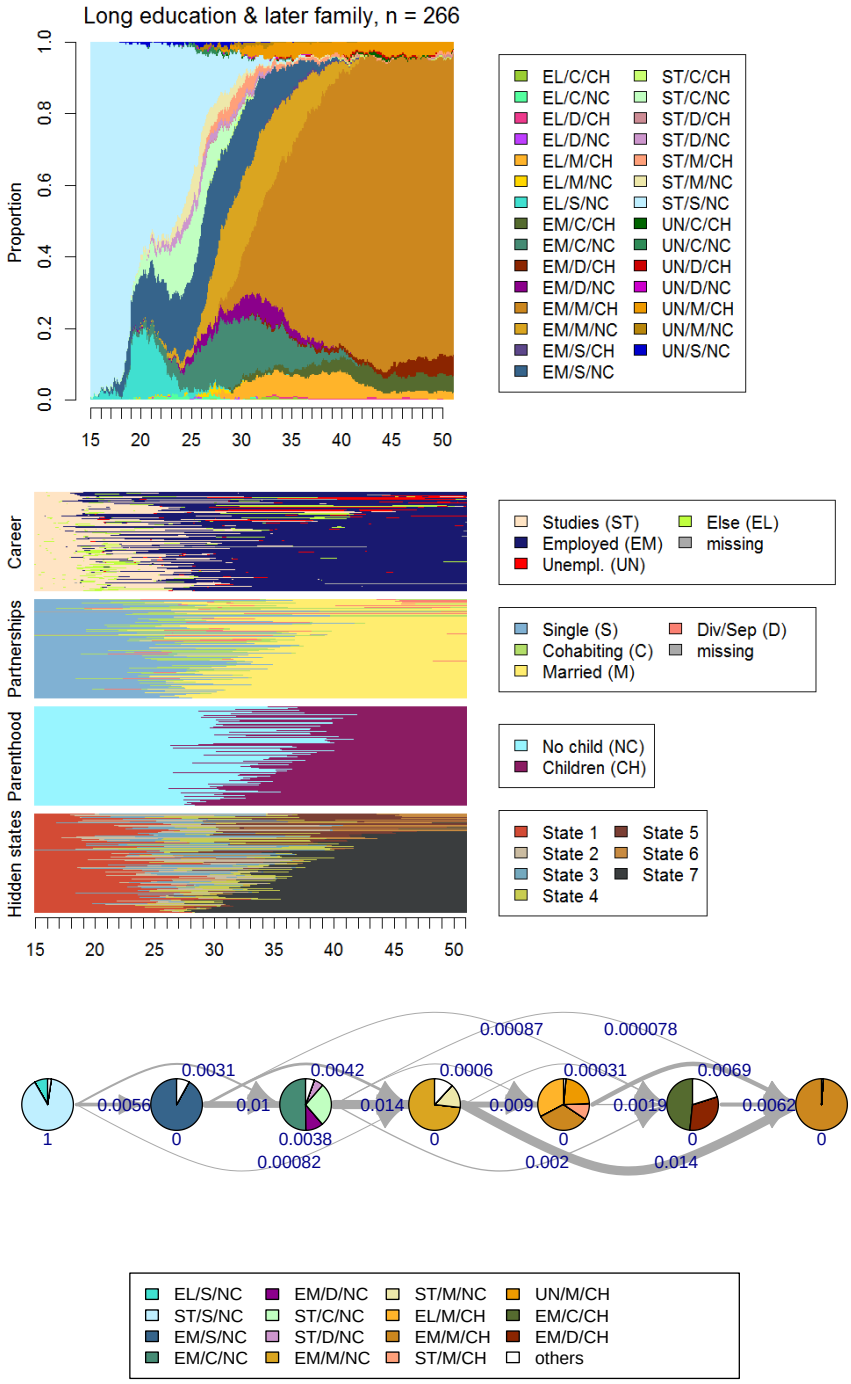

0.99
0.016
0.0054
0.0052
0.025
0.0024
0.0071
0.017
0.0021
0.0013
0.029
0
0.0036
0.0013
0.000046
0.00017
0.011
0.0036
0.0013
0.0053
0
0.014
0
0.00061

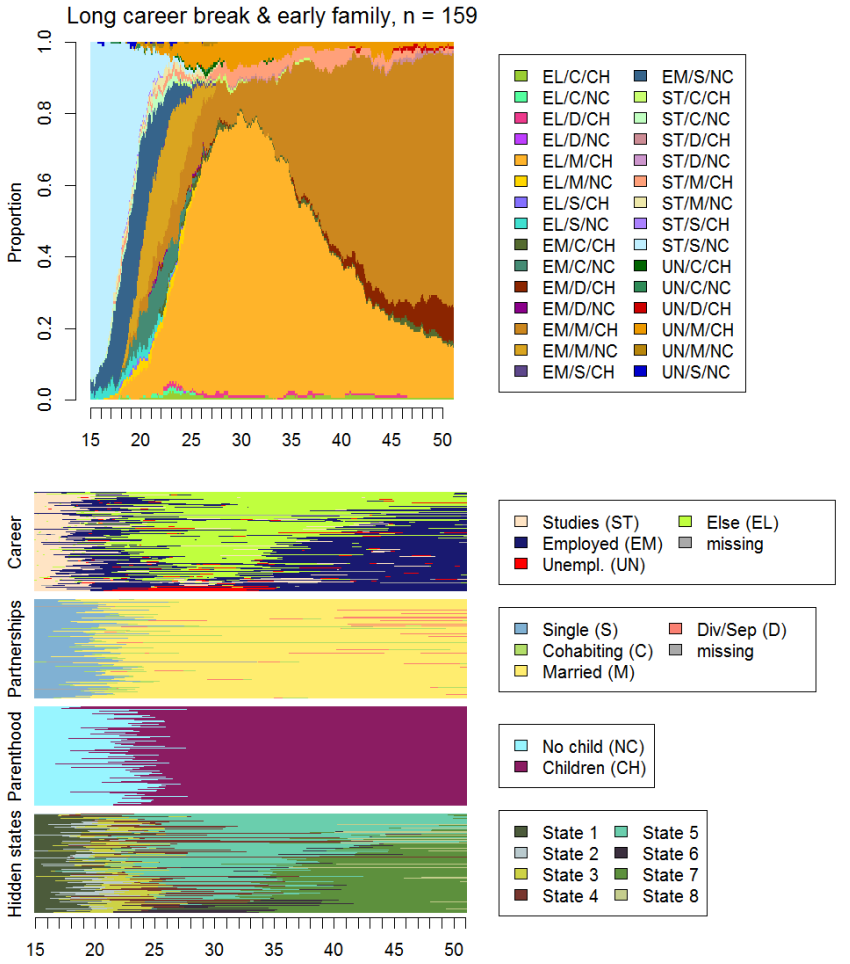

ST/S/NC
EL/S/NC
EM/S/CH
EM/S/NC
EM/C/NC
EM/M/NC
ST/C/NC
ST/M/NC
EM/M/CH
EM/C/CH
EM/D/CH
EL/M/CH
ST/M/CH
UN/M/CH
others

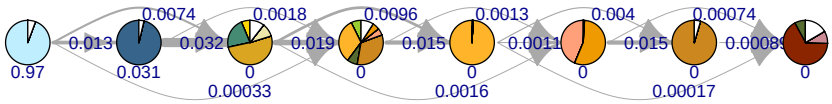

Short education & later family, n = 403
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/CH
EL/C/NC
EL/D/CH
EL/D/NC
EL/M/CH
EL/M/NC
EL/S/CH
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/CH
EM/S/NC
ST/C/CH
ST/C/NC
ST/D/CH
ST/D/NC
ST/M/CH
ST/M/NC
ST/S/CH
ST/S/NC
UN/C/CH
UN/C/NC
UN/D/CH
UN/D/NC
UN/M/CH
UN/M/NC
UN/S/CH
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
State 1
State 2
State 3
State 4
State 5
State 6
State 7


0.99
0.014
0.0054
0.002
0.01
0.0025
0.005
0.00055
0.024
0
0.00032
0.002
0.000045
0.0079
0
0.000094
0.0027
0.000082
0.0031
0
0.013
0.0042
0.0051
0

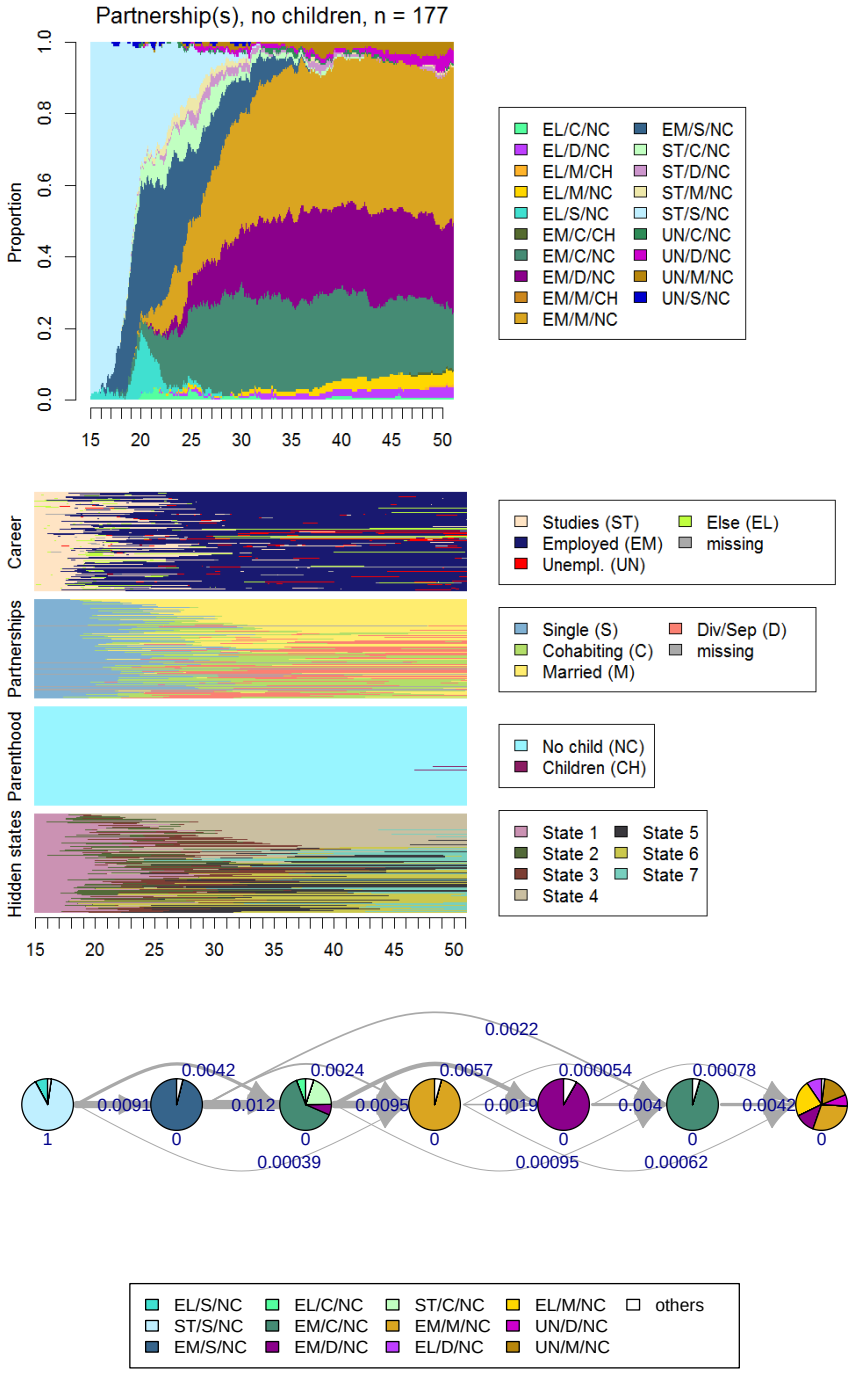

ST/S/NC
EL/S/NC
EM/S/NC
EM/C/NC
EM/D/NC
ST/C/NC
EM/M/NC
EM/C/CH
EM/D/CH
EM/M/CH
UN/M/CH
EL/M/CH
others

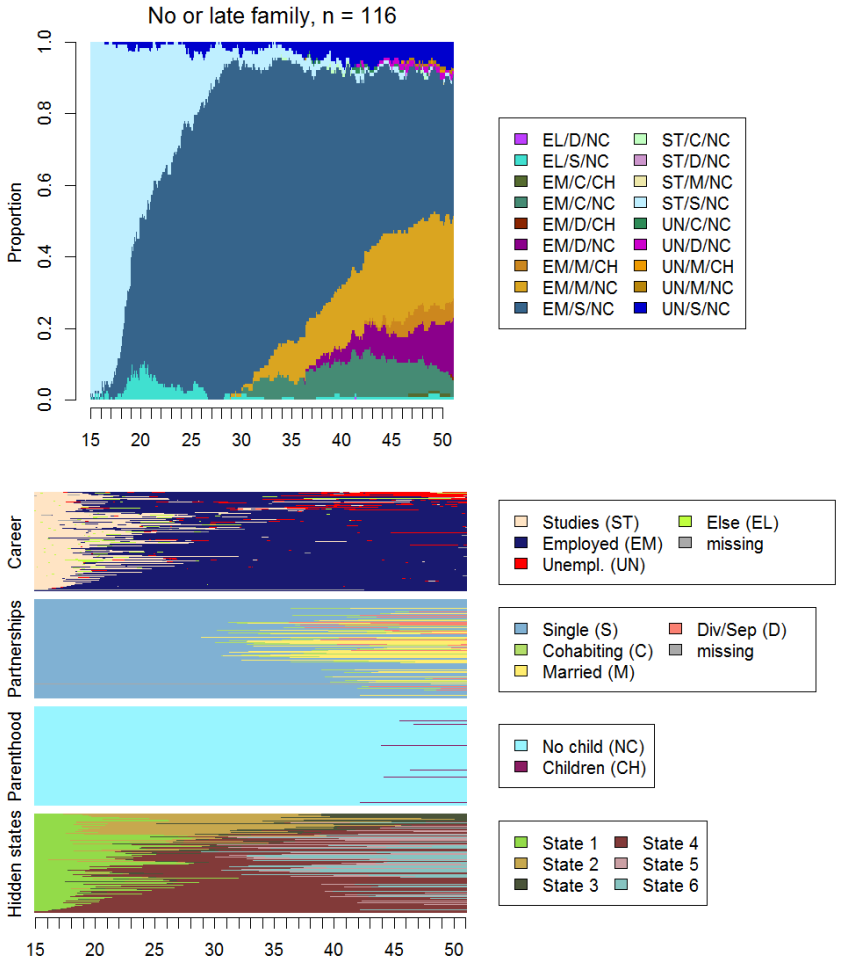

Long education & later family, n = 266
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/CH
EL/C/NC
EL/D/CH
EL/D/NC
EL/M/CH
EL/M/NC
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/CH
EM/S/NC
ST/C/CH
ST/C/NC
ST/D/CH
ST/D/NC
ST/M/CH
ST/M/NC
ST/S/NC
UN/C/CH
UN/C/NC
UN/D/CH
UN/D/NC
UN/M/CH
UN/M/NC
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
State 1
State 2
State 3
State 4
State 5
State 6
State 7
0.00087
0.000078
0.0031
0.0042
0.0006
0.00031
0.0069
0.0056
0,01
0.014
0.009
0.0019
0.0062
1
0
0.0038
0
0
0
0
0.00082
0.002
0.014
EL/S/NC
ST/S/NC
EM/S/NC
EM/C/NC
EM/D/NC
ST/C/NC
ST/D/NC
EM/M/NC
ST/M/NC
EL/M/CH
EM/M/CH
ST/M/CH
UN/M/CH
EM/C/CH
EM/D/CH
others

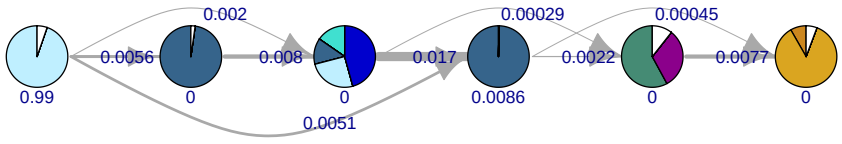

Long career break & early family, n = 159
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/CH
EL/C/NC
EL/D/CH
EL/D/NC
EL/M/CH
EL/M/NC
EL/S/CH
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/CH
EM/S/NC
ST/C/CH
ST/C/NC
ST/D/CH
ST/D/NC
ST/M/CH
ST/M/NC
ST/S/CH
ST/S/NC
UN/C/CH
UN/C/NC
UN/D/CH
UN/M/CH
UN/M/NC
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
State 1
State 2
State 3
State 4
State 5
State 6
State 7
State 8


0.97
0.013
0.0074
0.031
0.032
0.0018
0
0.00033
0.019
0.0096
0
0.015
0.0013
0
0.0016
0.0011
0.004
0
0.015
0.00074
0
0.00017
0.00089
0

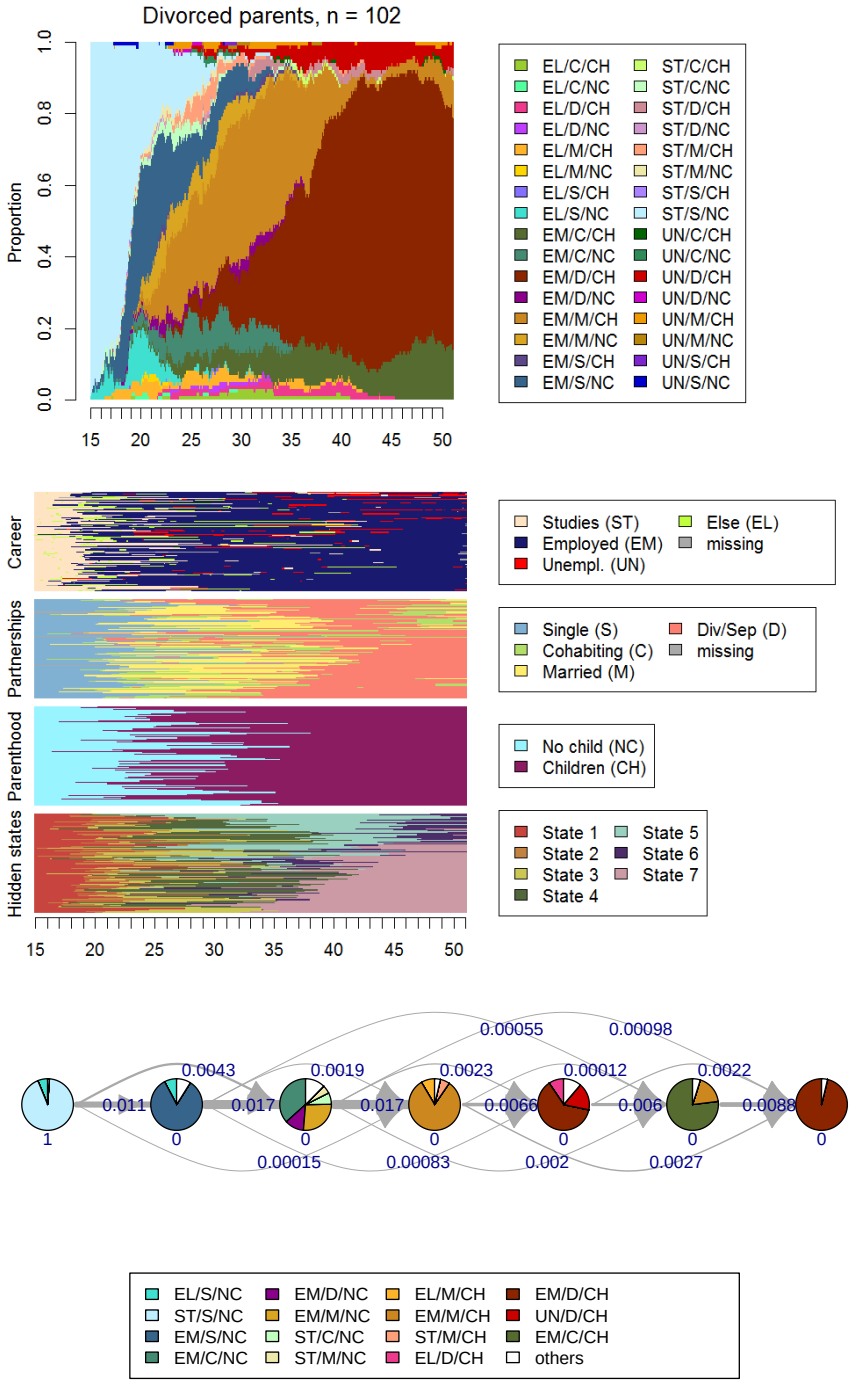

ST/S/NC
EM/S/NC
EL/M/NC
EM/C/NC
EM/M/NC
ST/M/NC
EL/C/CH
EL/M/CH
EM/C/CH
EM/M/CH
ST/M/CH
UN/M/CH
EM/D/CH
ST/D/CH
others

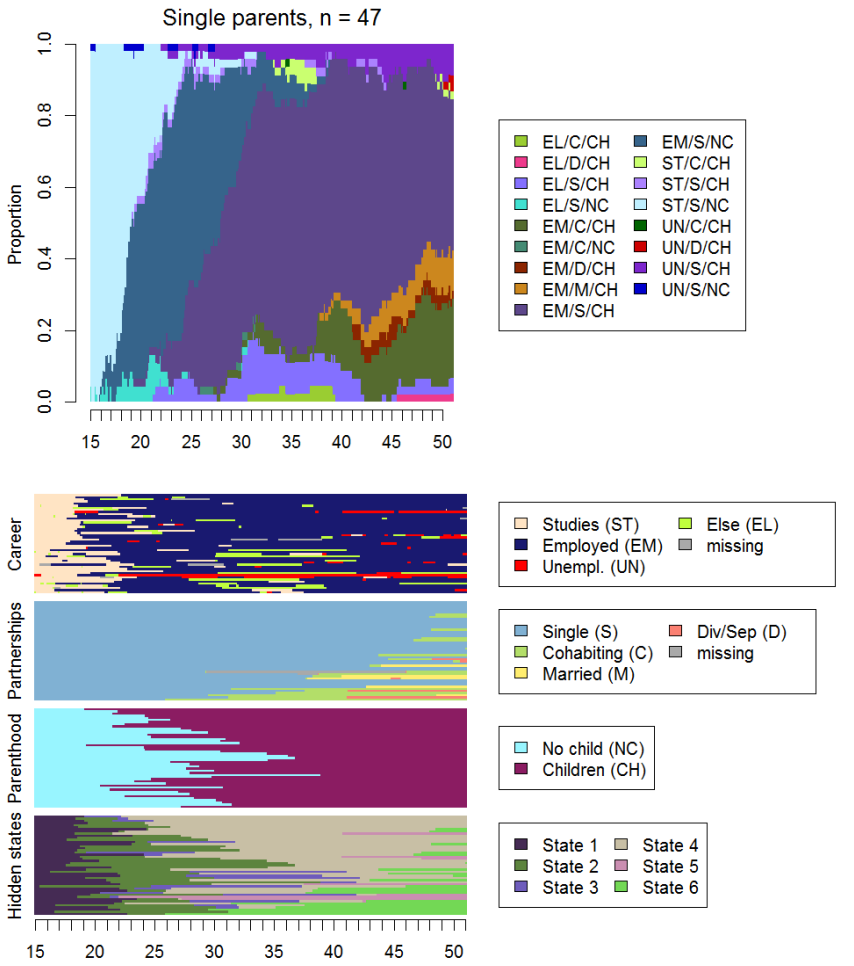

Partnership(s), no children, n = 177
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/NC
EL/D/NC
EL/M/CH
EL/M/NC
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/NC
ST/C/NC
ST/D/NC
ST/M/NC
ST/S/NC
UN/C/NC
UN/D/NC
UN/M/NC
UN/S/NC
Career
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Partnerships
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
Parenthood
No child (NC)
Children (CH)
Hidden states
State 1
State 2
State 3
State 4
State 5
State 6
State 7
0.0022
0.0042
0.0024
0.0057
0.000054
0.00078
0.0091
0.012
0.0095
0.0019
0.004
0.0042
1
0
0
0
0
0
0
0.00039
0.00095
0.00062
EL/S/NC
ST/S/NC
EM/S/NC
EL/C/NC
EM/C/NC
EM/D/NC
ST/C/NC
EM/M/NC
EL/D/NC
EL/M/NC
UN/D/NC
UN/M/NC
others

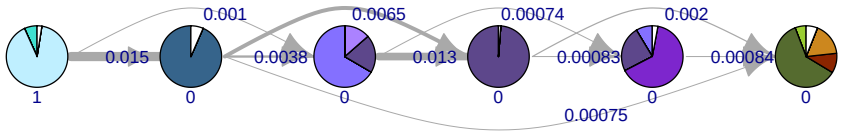

No or late family, n = 116
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/D/NC
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/NC
ST/C/NC
ST/D/NC
ST/M/NC
ST/S/NC
UN/C/NC
UN/D/NC
UN/M/CH
UN/M/NC
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
State 1
State 2
State 3
State 4
State 5
State 6

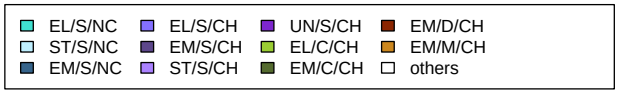

0.99
0.0056
0
0.002
0.008
0
0.0051
0.017
0.0086
0.00029
0.0022
0
0.00045
0.0077
0


ST/S/NC
EM/S/NC
EL/S/NC
UN/S/NC
EM/C/NC
EM/D/NC
EM/M/CH
EM/M/NC
others


Divorced parents, n = 102
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/CH
EL/C/NC
EL/D/CH
EL/D/NC
EL/M/CH
EL/M/NC
EL/S/CH
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/D/NC
EM/M/CH
EM/M/NC
EM/S/CH
EM/S/NC
ST/C/CH
ST/C/NC
ST/D/CH
ST/D/NC
ST/M/CH
ST/M/NC
ST/S/CH
ST/S/NC
UN/C/CH
UN/C/NC
UN/D/CH
UN/D/NC
UN/M/CH
UN/M/NC
UN/S/CH
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
State 1
State 2
State 3
State 4
State 5
State 6
State 7
0.00055
0.00098
0.0043
0.0019
0.0023
0.00012
0.0022
0.011
0.017
0.017
0.0066
0.006
0.0088
1
0
0
0
0
0
0
0.00015
0.00083
0.002
0.0027
EL/S/NC
ST/S/NC
EM/S/NC
EM/C/NC
EM/D/NC
EM/M/NC
ST/C/NC
ST/M/NC
EL/M/CH
EM/M/CH
ST/M/CH
EL/D/CH
EM/D/CH
UN/D/CH
EM/C/CH
others


Single parents, n = 47
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
15
20
25
30
35
40
45
50
EL/C/CH
EL/D/CH
EL/S/CH
EL/S/NC
EM/C/CH
EM/C/NC
EM/D/CH
EM/M/CH
EM/S/CH
EM/S/NC
ST/C/CH
ST/S/CH
ST/S/NC
UN/C/CH
UN/D/CH
UN/S/CH
UN/S/NC
Career
Partnerships
Parenthood
Hidden states
Studies (ST)
Employed (EM)
Unempl. (UN)
Else (EL)
missing
Single (S)
Cohabiting (C)
Married (M)
Div/Sep (D)
missing
No child (NC)
Children (CH)
State 1
State 2
State 3
State 4
State 5
State 6


0.001
0.0065
0.00074
0.002
0.015
0.0038
0.013
0.00083
0.00084
0.00075
1
0
0
0
0
0


EL/S/NC
ST/S/NC
EM/S/NC
EL/S/CH
EM/S/CH
ST/S/CH
UN/S/CH
EL/C/CH
EM/C/CH
EM/D/CH
EM/M/CH
others

References

Aassve, A., Billari, F. C., and 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.

Aisenbrey, S. and Fasang, A. (2010). New life for old ideas: The "second wave" of sequence analysis – bringing the "course" back into the life course. *Sociological Methods & Research*, 38(3):420–462.

Bartolucci, F., Pennoni, F., and 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.

Baum, L. E. and 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., Schneider, T., Kiesl, S. K., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., Prenzel, M. S., et al. (2011). Education as a lifelong process-the German National Educational Panel Study (NEPS). *Age*, 74(73):72.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., and Vermunt, J. K. (2012). Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic Medicine*, 74(4):366–376.

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

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

Elzinga, C. H. and Studer, M. (2014). Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, pages 3–47.

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.

Gauthier, J.-A., Bühlmann, F., and Blanchard, P. (2014). Introduction: Sequence analysis in 2014. In *Advances in Sequence Analysis: Theory, Method, Applications*, pages 1–17. Springer.

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

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160.

Helske, S. and Helske, J. (2016). Mixture hidden Markov models for sequence data: the seqHMM package in R. *Submitted*.

Helske, S., Steele, F., Kokko, K., Räikkönen, E., and 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.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3.

McVicar, D. and Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(2):317–334.

Müller, N. S., Sapin, M., Gauthier, J.-A., Orita, A., and 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.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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

Studer, M. and 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.

Studer, M., Ritschard, G., Gabadinho, A., and Müller, N. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40(3):471–510.

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

Vermunt, J. K., Tran, B., and Magidson, J. (2008). *Latent Class Models in Longitudinal Research*, pages 373–385. Handbook of Longitudinal Research: Design, Measurement, and Analysis. Elsevier, Burlington, MA.

Whiting, R. and Pickett, E. (1988). On model order estimation for partially observed Markov chains. *Automatica*, 24(4):569–572.

124. VIHOLA, MATTI, On the convergence of unconstrained adaptive Markov chain Monte Carlo algorithms. (29 pp.) 2010
125. ZHOU, YUAN, Hajłasz–Sobolev extension and imbedding. (13 pp.) 2010
126. VILPPOLAINEN, MARKKU, Recursive set constructions and iterated function systems: separation conditions and dimension. (18 pp.) 2010
127. JULIN, VESA, Existence, uniqueness and qualitative properties of absolute minimizers. (13 pp.) 2010
128. LAMMI, PÄIVI, Homeomorphic equivalence of Gromov and internal boundaries. (21 pp.) 2011
129. ZAPADINSKAYA, ALEKSANDRA, Generalized dimension distortion under Sobolev mappings. (18 pp.) 2011
130. KEISALA, JUKKA, Existence and uniqueness of $p(x)$-harmonic functions for bounded and unbounded $p(x)$. (56 pp.) 2011
131. SEPPÄLÄ, HEIKKI, Interpolation spaces with parameter functions and L_2-approximations of stochastic integrals. (18 pp.) 2011
132. TUHOLA-KUJANPÄÄ, ANNA, On superharmonic functions and applications to Riccati type equations. (17 pp.) 2012
133. JIANG, RENJIN, Optimal regularity of solutions to Poisson equations on metric measure spaces and an application. (13 pp.) 2012
134. TÖRMÄKANGAS, TIMO, Simulation study on the properties of quantitative trait model estimation in twin study design of normally distributed and discrete event-time phenotype variables. (417 pp.) 2012
135. ZHANG, GUO, Liouville theorems for stationary flows of generalized Newtonian fluids. (14 pp.) 2012
136. RAJALA, TUOMAS, Use of secondary structures in the analysis of spatial point patterns. (27 pp.) 2012
137. LAUKKARINEN, EIJA, On Malliavin calculus and approximation of stochastic integrals for Lévy processes. (21 pp.) 2012
138. GUO, CHANGYU, Generalized quasidisks and the associated John domains. (17 pp.) 2013
139. ÄKKINEN, TUOMO, Mappings of finite distortion: Radial limits and boundary behavior. (14 pp.) 2014
140. ILMAVIRTA, JOONAS, On the broken ray transform. (37 pp.) 2014
141. MIETTINEN, JARI, On statistical properties of blind source separation methods based on joint diagonalization. (37 pp.) 2014
142. TENGVALL, VILLE, Mappings of finite distortion: Mappings in the Sobolev space $W^{1,n-1}$ with integrable inner distortion. (22 pp.) 2014
143. BENEDICT, SITA, Hardy-Orlicz spaces of quasiconformal mappings and conformal densities. (16 pp.) 2014
144. OJALA, TUOMO, Thin and fat sets: Geometry of doubling measures in metric spaces. (19 pp.) 2014
145. KARAK, NIJJWAL, Applications of chaining, Poincaré and pointwise decay of measures. (14 pp.) 2014
146. JYLHÄ, HEIKKI, On generalizations of Evans and Gangbo's approximation method and L^∞ transport. (20 pp.) 2014
147. KAURANEN, AAPO, Space-filling, energy and moduli of continuity. (16 pp.) 2015
148. YLINEN, JUHA, Decoupling on the Wiener space and variational estimates for BSDEs. (45 pp.) 2015
149. KIRSILÄ, VILLE, Mappings of finite distortion on generalized manifolds. (14 pp.) 2015
150. XIANG, CHANG-LIN, Asymptotic behaviors of solutions to quasilinear elliptic equations with Hardy potential. (20 pp.) 2015
151. ROSSI, EINO, Local structure of fractal sets: tangents and dimension. (16 pp.) 2015
152. HELSKE, JOUNI, Prediction and interpolation of time series by state space models. (28 pp.) 2015
153. REINIKAINEN, JAAKKO, Efficient design and modeling strategies for follow-up studies with time-varying covariates. (36 pp.) 2015
154. NUUTINEN, JUHO, Maximal operators and capacities in metric spaces. (22 pp.) 2016
155. BRANDER, TOMMI, Calderón's problem for p-Laplace type equations. (21 pp.) 2016
156. ÄRJE, JOHANNA, Improving statistical classification methods and ecological status assessment for river macroinvertebrates. (30 pp.) 2016

ISBN 978-951-39-6757-4 (print)
ISBN 978-951-39-6758-1 (pdf)
ISSN 1457-8905