Advances in Life Course Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Advances in Life Course Research

journal homepage: www.elsevier.com/locate/alcr

Holistic analysis of the life course: Methodological challenges and new perspectives

Raffaella Piccarreta[a,*], Matthias Studer[b,c]

[a] Department of Decision Sciences, Bocconi University, via Rontgen 1, Milan, Italy
[b] Geneva School of Social Sciences
[c] NCCR LIVES—Overcoming Vulnerability: Life Course Perspectives, University of Geneva, Bd du Pont-d'Arve 28, CH-1205, Geneva, Switzerland

ARTICLE INFO

Keywords:
Sequence analysis
Trajectories
Cluster analysis
Mixed latent Markov models
Multistate models

ABSTRACT

We survey state-of-the-art approaches to study trajectories in their entirety, adopting a holistic perspective, and discuss their strengths and weaknesses. We begin by considering sequence analysis (SA), one of the most established holistic approaches. We discuss the inherent problems arising in SA, particularly in the study of the relationship between trajectories and covariates. We describe some recent developments combining SA and Event History Analysis, and illustrate how weakening the holistic perspective—focusing on sub-trajectories—might result in a more flexible analysis of life courses. We then move to some model-based approaches (included in the broad classes of multistate and of mixture latent Markov models) that further weaken the holistic perspective, assuming that the difficult task of predicting and explaining trajectories can be simplified by focusing on the collection of observed transitions.

Our goal is twofold. On one hand, we aim to provide social scientists with indications for informed methodological choices and to emphasize issues that require consideration for proper application of the described approaches. On the other hand, by identifying relevant and open methodological challenges, we highlight and encourage promising directions for future research.

1. Introduction

In many frameworks, for instance, in Event History Analysis (EHA), the analysis of life courses and of their dynamics is based on the study of focal events or transitions. Instead, the adoption of a holistic approach implies regarding life courses as meaningful units (e.g., "careers" or trajectories). Such a perspective aims to account for complex time-related interdependencies underlined in life course research (e.g., the "life course cube", Bernardi, Huinink, & Settersten, 2018, in this special issue). The goal is to undertake a joint study of events, their duration, and the transitions experienced by individuals over a prolonged period.

Following the idea of multidimensionality of the life course (Bernardi et al., 2018), this approach allows one to consider that "single events should not be isolated from each other," but rather need "to be understood in their continuity" (Aisenbrey & Fasang, 2010, p. 421). Many important transitions in the life course, such as the transition to adulthood (Shanahan, 2000) or the professional integration of unemployed cannot be described by a single instantaneous event, and should be conceptualized as part of a process that takes time and that can be relevant by itself. Similar considerations hold for "turning points," defined as "alterations or deflection in a long-term pathway or trajectory that was initiated at an earlier point in time" (Sampson & Laub, 2005, p.16). A holistic perspective aims to identify such processes, turning points, and transitions within trajectories by taking an overall approach, and to achieve a deeper understanding of the phenomenon. The goal of this work is to review state-of-the-art approaches for the study of trajectories, and to help social scientists make informed methodological choices. Therefore, we discuss the strengths and weaknesses of each approach, identify some relevant and unresolved methodological challenges, and highlight promising, and much needed, directions for further research.

Table 1 offers a summary of the methods considered in this paper, together with an illustration of their most relevant and distinguishing

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion; CTA, Competing Trajectory Analysis; EHA, event history analysis; HMM, hidden Markov model; LCA, latent class analysis; MHMM, mixture hidden Markov model; MSM, multistate model; OMA, optimal matching analysis; SA, sequence analysis; SAMM, sequence analysis multistate model

* Corresponding author at: Department of Decision Sciences, Bocconi University, via Rontgen 1, Milan, Italy.
E-mail addresses: raffaella.piccarreta@unibocconi.it (R. Piccarreta), Matthias.Studer@unige.ch (M. Studer).

hiips://doi.org/10.1016/j.alcr.2018.10.004
Received 27 November 2017; Received in revised form 7 October 2018; Accepted 10 October 2018
1040-2608/ © 2018 Elsevier Ltd. All rights reserved.

Please cite this article as: Piccarreta, R., Advances in Life Course Research, https://doi.org/10.1016/j.alcr.2018.10.004

ARTICLE IN PRESS

Table 1
Summarized characteristics of the approaches described in the work.[a]

	Standard SA	Combination of SA and EHA			SCM	Mixed Latent Markov Model[b]		
		CTA	SAMM	SHA		LCA	HMM	MHMM
Scientific culture	Algorithmic modeling	Combination of Algorithmic and Stochastic data modeling culture			Stochastic data modeling	Stochastic data modeling		
Theoretical concept	Trajectory	Start of the processes	Transition to sub-sequences	Previous trajectory and event	Transitions/Durations	Trajectory	Transitions (latent states)	Trajectory and Transitions (latent states)
Goal: identify	Groups of equivalent trajectories	Probability of equivalent start of the processes	Probability of transition to equivalent sub-sequences	Association between previous trajectory and upcoming event	Probability of transitions, duration of states	Groups of equivalent trajectories	Probability of transitions	Groups of equivalent latent trajectories
Assumptions on the sequences generating process	None	None	Semi-Markov	None	Semi-Markov	Independence (conditional to the latent class)	First-order Markov Time-homogeneous transition rates	
Multiple domains	Yes, Joint SA	No			No	Yes		
Inference on whole trajectory (baseline covariates)	ANOVA-like approaches (testing differences) Multinomial regression (cluster membership)	No			No	Yes	No	Yes
Inference on trajectory's unfolding (time-varying covariates)	Description of mutual association using joint SA	Yes, on the start of the process	Yes (association over a medium-term period)	Yes	Yes	No	Yes	Yes
Censored Trajectories	No	Yes			Yes	Yes		
Missing data	No	No			Yes (only in trajectories)	Yes (only in trajectories)		

[a] Aisenbrey and Fasang (2010, p. 424) propose a similar table to compare EHA and SA. Our table extends the original one both with respect to the models taken into account and with respect to the distinctive aspects considered in the comparison.
[b] The characteristics of the considered models refer to the specific formulations considered in the paper.

ARTICLE IN PRESS

features.

Following Breiman (2001), the first distinction concerns the "culture" underlying the different approaches. Sequence Analysis (SA) refers to the data mining or algorithmic culture, which aims to efficiently recover the most relevant patterns in data without any assumption on the data-generating process or any predefined judgment about the relevant features of the life course. Therefore, SA is an *exploratory* data-driven approach, based on the idea that transitions and changes within sequences might have medium-term effects on future evolution, and analogously, that it is not possible to simplify how past experience impacts the trajectory's subsequent unfolding. This allows the unveiling of different types of *temporal interdependencies* (Bernardi et al., 2018), such as "anticipative" or "path dependence" mechanisms (e.g., steps facilitating or hindering the experience of specific events), or a mix of the two. This clearly comes at a cost, because considering the trajectories as a whole and without any simplifying assumption on their unfolding mechanism, poses some problems with respect to the possibility of handling trajectories only partially observed and/or of studying the impact of time-varying covariates on life courses.

On the other hand, Event History analysis (EHA), multistate and mixed latent Markov models are rooted in statistical culture, based on the assumption of a data generating mechanism with specific characteristics (Aisenbrey & Fasang, 2010). This allows using statistical inference to draw conclusions about the structure of the data or the relationships between covariates and trajectories. Nonetheless, the simplifying assumptions at the basis of such models are not necessarily well suited for the data at hand, and may lead to unreliable results when violated.

Between these two alternative views, other approaches have been recently introduced in the literature on life courses which —combining SA and EHA—exploit and somehow reconcile the two cultures.

Other than for their scientific tradition, the considered methods differ in their primary object of interest and in their goals. Some approaches, specifically SA and Latent Class Analysis (LCA), focus on whole trajectories, and aim to uncover groups of similar sequences, and possibly, explain their relations with baseline covariates (observed prior to the moment when the trajectory starts). Instead, multistate models and hidden Markov Models (HMM) focus on instantaneous transitions within the life course, and on factors that might explain the probability of experiencing them. Again, some methods hold an intermediate position between the two opposite perspectives. Mixture Hidden Markov Models (MHMM) allow identifying groups of similar trajectories and studying transitions simultaneously. Models combining SA and EHA focus on transitions to sub-trajectories; thus, they adopt a medium-term perspective on changes within life courses.

We review these approaches, their strengths and weaknesses, and discuss the aspects that require consideration for their proper application. Section 2 reviews the SA framework and some recent proposals combining EHA and SA. In Section 3, we discuss model-based approaches used in the literature to analyze life courses. We conclude with a discussion and some remarks in Section 4.

2. Sequence analysis

Since its introduction in the social sciences by Abbott (1995), SA has been increasingly used in life-course research to study processes that are coded as sequences, that is, as the ordered collection of the states experienced over a period, typically observed at regular intervals. This coding is close to the concept of trajectory used in the life-course paradigm, defined as the sequence of roles and social statuses (Bernardi et al., 2018; Elder, Kirkpatrick Johnson, & Crosnoe, 2003). Rooted in the data-mining culture, SA provides a holistic perspective on trajectories by considering them as the main statistical units. The main goal of SA is to describe trajectories and identify their most salient and distinctive features.

Several powerful visualization techniques, such as the chronogram, the index plot (Scherer, 2001) and its extensions (see e.g., Fasang & Liao, 2014; Piccarreta, 2017; Piccarreta & Lior, 2010), the decorated parallel coordinate plot (Bürgin & Ritschard, 2014), are available for effective exploration of trajectories.

Typically, SA proceeds by grouping similar trajectories, obtaining a *typology* identifying typical temporal patterns in sequences. Indeed, individual trajectories usually have some small and negligible differences (e.g., slightly different duration of the visited states, or small spells in different states). The construction of a typology of sequences is designed to ignore such differences, and to unveil homogeneous groups of trajectories that are distinct from one another. Sometimes, the resulting data-driven *types* match theoretically expected ideal-types (in a Weberian sense, see Abbott & Hrycak, 1990) of processes or trajectories in data.

Technically, SA proceeds by calculating dissimilarities among life courses, based on criteria that properly account for the most relevant observed differences in timing ("when"), sequencing ("in what order"), and duration ("how long") of those states experienced by individuals throughout a period (see e.g., Studer & Ritschard, 2016). Cluster analysis is then used to obtain a typology based on these dissimilarities.

The original SA framework has been enriched and extended in several directions.

In their seminal works, Abbott and coauthors (Abbott, 1983; Abbott & Forrest, 1986; Abbott & Hrycak, 1990) first addressed the problem of measuring dissimilarity between sequences by extending to social science Optimal Matching (OM), an edit distance—originally developed in the field of information theory and computer science (Levenshtein, 1965)—based on the effort needed to transform one sequence into another. Since then, many different dissimilarity measures have been considered and introduced in the literature, assigning different importance to the trajectory features (timing, sequencing and duration) when assessing dissimilarities between two life courses (see Studer & Ritschard, 2016, for a theoretical and empirical review and comparison of alternative proposals). The unavoidable dependence of SA results on the dissimilarity measure should not be considered disadvantageous; rather, it guides the researcher in defining the career aspects worthy of being distinguished, in line with a specific research question. Therefore, this choice should be fully motivated from the theoretical viewpoint (see Studer & Ritschard, 2016, for some guidelines on this choice). In addition, the performance of the chosen criterion for the data at hand should be evaluated to ensure it actually highlights the expected career characteristics.

The development of "joint" or "multichannel" SA (see Gauthier, Widmer, Bucher, & Notredame, 2010; Pollock, 2007; and Piccarreta, 2017, for a review) allows considering several life domains simultaneously when a set of trajectories is available for each individual (e.g., describing the evolution of work activities, partnership, and parenthood over time). Technically, joint SA includes any approach leading to the definition of a dissimilarity measure based on the information arising from all the domains taken into account. As in the single domain case, such dissimilarities can be exploited to obtain—via cluster analysis—a *joint typology* of the set of sequences, describing the most typical *combinations of patterns* observed across the domains.

Focus on the whole trajectory does not exclude the interest or need for inference, or for procedures to verify whether individuals with specific characteristics experience significantly different careers. Making inferences about sequences is not easy, as their relationship with covariates is typically complex and manifold. Within the standard SA framework, the focus has been mainly on the possible association between a set of covariates and the *whole trajectory*. This is usually achieved by estimating a multinomial regression where the typology is the dependent variable (see e.g., McVicar & Anyadike-Danes, 2002).

The standard SA framework provides researchers with a rich collection of tools to analyze trajectories. Nonetheless, in the following, we illustrate some specific aspects, often overlooked in applied research, which should be carefully considered to obtain reliable results.

ARTICLE IN PRESS

2.1. The standard SA framework: assessment of quality and reliability of results

The application of SA requires the preliminary coding of the trajectories (or processes) as sequences and the choice of a dissimilarity criterion. Typically, cluster analysis is applied to obtain a typology identifying the most typical patterns in data. Multinomial regression is then used to relate the obtained types to factors that supposedly influence the probability of experiencing different types of trajectories. While very common, this path of analysis can raise some issues that, if neglected, might undermine the quality and reliability of the obtained results. In this section, we review such criticisms, while providing general guidelines to tackle them.

2.1.1. Robustness of cluster-analysis results

The typology obtained with cluster analysis is often used to identify the most relevant temporal patterns in data. Typically, data-driven types are individuated by summarizing the features of sequences within the same cluster. This can be done by associating to each cluster its *medoid*, that is, the trajectory most similar to all the others in the cluster. When interpreting a typology, deviating sequences are usually ignored, given that descriptions of the social world require a certain degree of simplification and that the deviations of trajectories from the types can be considered as the reflections of different realizations of the same underlying process (Abbott, 1995; Studer, 2013). Nonetheless, this is only reasonable and trustworthy when a reliable partition has been obtained.

Indeed, cluster analysis always produces a grouping of sequences, even when there is not a "natural" or "relevant" partition (see, among others, Abbott & Tsay, 2000, and Levine, 2000). Furthermore, different clustering algorithms might lead to different partitions, potentially with strong differences when there are no well-separated groups of cases within the data. Therefore, assessing clustering quality is crucial: it can guide and support the identification of the most suitable typology, and therefore, the strength of the conclusions drawn on its basis. Too often, such evaluation is disregarded in SA studies; however, it should be a part of standard and routine procedures.

Cluster evaluation can be conducted at different levels. To assess the *global* quality of a partition (see Studer, 2013, for a review), we avoid criteria based on assumptions (such as multivariate normality), which would not be encountered when analyzing sequences and their associated dissimilarities. Among the available criteria, the well-known R^2 measures the amount of heterogeneity within the whole sample, which is accounted for by the clusters. The *average silhouette width* summarizes the individual silhouette coefficients, based on the comparisons between the closeness of a sequence to its own cluster and its closeness to others (Kaufman & Rousseeuw, 1990). The *stability* of a partition (Hennig, 2007, 2008), refers to its resistance to small data perturbations or to the clustering algorithm used.

However, a good global quality partition does not necessarily imply high levels of internal cohesion for *all* clusters. In fact, some clusters might be very well defined, while others could include outliers, that is, sequences showing very distinguishing features and deviating consistently from the others. Therefore, it is also necessary to evaluate the quality of each cluster, for example, by using the aforementioned average silhouette width or stability measure, which can be calculated separately for each cluster.

In addition, a typology might perfectly summarize some sequences and not others. For instance, some sequences could be on the borderline, lying between two different clusters.

Low levels of internal homogeneity, and in some cases, weak degrees of cluster membership, are not a crucial concern when clusters are used to describe the most relevant patterns, *and* when borderline or outlying cases do not influence the typology description. To verify that this is the case, one could define "robust" medoids by focusing on the most central sequences in the clusters, and excluding critical sequences. Critical sequences can be identified based on their average dissimilarity to other sequences in their cluster, or on their deviation from the medoid, or on their individual silhouette coefficient (Kaufman & Rousseeuw, 1990; Studer, 2013). Some algorithms, such as partitioning around the medoids or fuzzy clustering (Everitt, Landau, Leese, & Stahl, 2011; Kaufman & Rousseeuw, 1990), directly provide such information. The partitioning around medoids algorithm assigns sequences to clusters based on dissimilarity to the clusters' medoids. Fuzzy clustering calculates for each sequence the degree to which it belongs to each cluster (see Studer, 2018, for a discussion on the use of fuzzy clustering in SA).

When several domains are studied using joint SA, the quality of a partition—obtained based on joint dissimilarities—should be evaluated both at a joint level and in relation to each specific domain. Indeed, if the trajectories in different domains are not *all* interrelated (see Piccarreta, 2017, and Piccarreta & Elzinga, 2013, for a discussion on association criteria), the obtained clusters typically will satisfactorily describe the characteristics of only some domains. In fact, cluster analysis will be driven by the more interconnected domains, whose trajectories evolve coherently, and/or by the less turbulent domains, whose trajectories are (relatively) more similar, and therefore, more easily grouped. In these situations, the possibly relevant multiple-domains features would be only partially identified by the joint typologies. Therefore, on one hand, a preliminary assessment of the level of association among domains corroborates the suitability of a joint SA. On the other hand, the evaluation of cluster quality also verifies whether a joint typology can be identified and/or which domains are sufficiently connected for a joint analysis.

While the described procedures are preconditions for reliable and reasonable interpretations of an SA typology, their implementation cannot guarantee that the extracted types will properly match those eventually existing in the data. First, the (chosen) clustering algorithms might fail to correctly recover the "true" patterns underlying data (see Warren, Luo, Halpern-Manners, Raymo, & Palloni, 2015, for discussion). Furthermore, while statistical quality should certainly play a key role in the evaluation of a typology, a statistically satisfactory grouping of cases can be sociologically meaningless or insignificant, because data-driven types might not match with the expected ideal types, or because they aggregate or disaggregate sociologically meaningful types in unexpected or questionable ways. Consider, for example, a study on the school-to-work transition, and assume that standard (statistical) criteria lead to the selection of a partition mainly based on the sequence of visited states; thus placing in the same cluster people who attended college and entered the labor market after a period of unemployment. Such a partition might be inadequate from a sociological perspective, because it would not distinguish between individuals entering the labor market soon after finishing studies and individuals who instead experienced long periods of unemployment. To prevent (at least partially) such situations, one should carefully evaluate which trajectories' traits should be regarded as similar from the (adopted) sociological perspective. A coherent choice of the dissimilarity criterion, which clearly influences the results of cluster analysis, would increase the likelihood of obtaining typologies reflecting the (supposedly) relevant types.

To the best of our knowledge, the literature does not yet offer clear indications for assessing the *sociological validity* of a typology, despite the evident relevance of this aspect. We think that defining suitable procedures and guidelines in this direction is one of the most important challenges faced by cluster analysis in general and by SA in particular.

2.1.2. Relationships between trajectories and covariates

In many situations, it is of interest to study the relationships between covariates and entire trajectories. Specifically, attention can be focused on the evaluation of the *significance and strength* of the relationship or on the substantive *interpretation* of the covariates' impact on trajectories. In both cases, to avoid anticipatory analysis (see Hoem & Kreyenfeld, 2006), attention must be limited to *baseline* covariates

ARTICLE IN PRESS

(i.e., measured before the beginning of the trajectory).

To measure the *strength* of the association, Piccarreta and Billari (2007) extend ANOVA concepts and R^2 to dissimilarities and to SA to evaluate the ability of a *categorical* variable to account for total-sample heterogeneity. Studer, Ritschard, Gabadinho, and Müller, (2011) suggest assessing the statistical significance using permutation tests to estimate *p*-values; they further discuss an extension of MANOVA that permits considering the joint impact of several covariates. These ANOVA-like approaches allow one to identify significant relationships between covariates and sequences, and/or to individuate the most relevant covariates (see, e.g., Bonetti, Piccarreta, & Salford, 2013). Nonetheless, they do not provide qualitative indications on the *shape* of the relationships, nor do they allow drawing any substantive interpretation. For instance, one might conclude that men and women experience "significantly" different trajectories, without any insights about *how* they differ. It is a common practice to deduce possible differences by comparing the characteristics of sequences in different groups (e.g., using plots). This is viable only in the ANOVA setting (i.e., one covariate), because in MANOVA, one should consider the differences in the groups induced by one variable *conditioned to* all the others, and very soon, this becomes infeasible (and impossible when continuous covariates are included). Finally, as they are based on qualitative evaluations, these considerations might be highly subjective, and fail at identifying the diverging patterns. Further developments are clearly necessary to define criteria allowing for the appropriate identification of possible structural differences between groups of sequences, within either ANOVA or MANOVA frameworks.

To gain more insights into the sequences-covariates relationship, many studies adopt a two-stage approach. First, they apply cluster analysis to identify the main patterns in the data. Second, they use multinomial regression to relate the probability (or "risk") of experiencing different trajectory types to a set of background variables (see, e.g., McVicar & Anyadike-Danes, 2002). The results of this procedure are much easier to interpret than those obtained using the ANOVA-like approach. In the best case, the covariates' levels can be related to the specific patterns characterizing the different clusters.

Data reduction resulted from cluster analysis demands considerable caution. Indeed, the same response value is assigned to all sequences in the same cluster, completely neglecting the possible within-cluster dispersion. This is not problematic if such dispersion is small and random; however, when it is systematically related to one or more covariates, it could either suggest a false association or mask an existing one (see Studer, 2013, for a discussion). For instance, in a study on the relationship between gender and professional career, sequences within the same cluster might be characterized by similar state sequencing, but by consistent differences in state duration (i.e., structural residual variation). Results can be safely interpreted if gender differences mostly relate to sequencing, because clusters summarize this information. However, one might overlook any relationship related to duration, as the clusters do not account for this residual variation.

When within-cluster dispersion is high, cluster membership cannot be univocally interpreted, and its use as a dependent variable can raise interpretation issues. This procedure could even be misleading or meaningless if sequences within the same cluster are misinterpreted as being *similar in all respects*. Therefore, it is crucially important to preliminarily verify that residual variation does not preclude this approach's reliability.

Besides undertaking evaluations of cluster robustness and quality, one should also carefully identify the characteristics common to the sequences within the same cluster. In this way, cluster membership is well defined and there is no risk of over-interpretation. For instance, in the previous example on gender and professional career, one could stress that the multinomial regression relates the sequencing of states to covariates, *irrespective of the time spent in each state*, because the obtained clusters differ according to their sequencing but not according to their durations in each state.

In addition, it is advisable to identify outliers, namely, sequences badly represented by their cluster (this can be done using the criteria illustrated in the previous section). Generally, the lack of a clearly defined type and/or deviations from the type that cannot be considered as unavoidable fluctuations or as intrinsic and expectable differences among structurally similar careers suggest that the partition should be refined in order to obtain a reasonable level of within-cluster homogeneity. If this can be achieved only by considering a relatively large number of clusters, estimation issues might arise. In fact, a too large number of levels for the response variable or a too-low frequency for some of its levels can lead to poor estimation or to low confidence in the parameters estimated with multinomial regressions.

Even when properly applied (in the sense specified above) SA suffers some limitations, preventing its wider adoption by life course researchers. First, the handling of censored and missing data remains an open issue. Second, the focus on the entire career prevents the possibility of studying the relationship between sequences and *time-varying covariates*. Such limitations are discussed in the next subsection.

2.2. The standard SA framework and its limitations

2.2.1. Censoring and missing data

One relevant problem within the standard SA framework relates to the treatment of missing data. Indeed, SA focuses on the *entire trajectory* experienced by individuals over a *specific period*. Nonetheless, some sequences can present missing states at some points. In addition, some trajectories can be right-censored, being observed only for a sub-period (for example, when considering the family formation patterns between 18 and 40 years of age, all the individuals being younger than 40 at the last survey will present partially observed trajectories).

In the presence of missing data, many dissimilarity criteria would treat a missing state as a specific additional state, hence regarding the presence of missing states as an indicator of similarity between trajectories. Therefore, it is advisable not to include partially observed trajectories in the analysis, even if it were reasonable and convenient, at least when the data are missing completely at random (i.e., loosely speaking, when cases with missing values are just a random subset of the data).

Halpin (2016a, 2016b) introduced two interesting proposals to deal with this issue. Halpin (2016a) discusses the use of a "multiple imputation" procedure in conjunction with SA. Halpin (2016b) proposes to treat missing states as "self-different"; missing states are therefore maximally different from any other state, including the missing ones themselves. In this way, the absence of data is no longer a factor of similarity between sequences. Both proposals are reasonable and promising; nonetheless, there is not yet an in-depth analysis or evaluation of their effects on data or results. In particular, the ability of each method to recover a "meaningful" and unbiased typology in the presence of various kinds of missingness (MCAR, MAR, MNAR[1]) has not been discussed. In addition, it would be important to evaluate whether the use of multiple imputation artificially adds structural information.

The problem is even more serious and difficult to handle in the case of censored trajectories. In such situations, the apparently viable strategies of considering sequences of unequal length or using a missing state to code the unobserved part of the trajectories might lead to unsatisfactory results. Indeed, the length of the sequence would typically result as an element of similarity between trajectories. Additionally, normalizing the dissimilarities—as suggested by Levy, Gauthier, and Widmer (2006)—is not a suitable solution, since normalization accounts only for the varying number of features of the sequences and not

[1] MCAR: missing completely at random (missingness is completely random) MAR: Missing at random (missingness is completely random within subgroups of other observed variables), MNAR: Missing not at random (missingness depends on the missing values themselves).

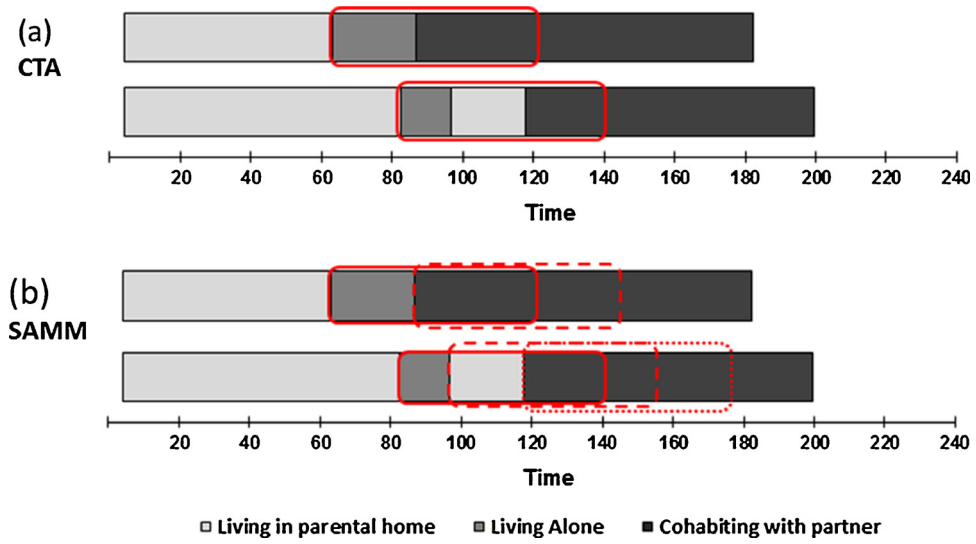

Fig. 1. Extraction of sub-sequences from two sequences of different lengths using (a) CTA and (b) SAMM Note that in (a) only one sub-sequence is extracted following the first exit from the initial state, whereas in (b) more sub-sequences are extracted, following each transition.

for differences in their lengths (see Elzinga & Studer, 2016, for discussion).

The presence of missing states or censored sequences would typically lead to the formation of a typology depending on the observation time or on the amount of missing data; frequently, this does not align with the goals of the analysis. Hence, this problem remains one of SA's major limitations (Aisenbrey & Fasang, 2010) and one of the most important and urgent areas for future research. Indeed, due to the lack of a clear strategy to address the problem, many studies limit their attention to sequences entirely observed for the considered period, thus, disregarding individuals presenting missing states. If missingness relates to individuals' characteristics, this necessarily leads to the systematic exclusion of specific strata of the population. In particular, the most recent cohorts, that are typically only partially observed, are often excluded from the analysis, despite their great research interest.

2.2.2. Time-varying covariates

Focusing on the trajectory as a unique conceptual unit prevents the possibility of analyzing the effect of *time-varying covariates* on its unfolding. However, this perspective is crucial when exploring the many interdependencies advocated by the "life course perspective" (e.g., Bernardi et al., 2018), that might require studying the relation between different life domains (e.g., professional and family formation patterns), or between linked lives (e.g., careers of parents and children). Furthermore, it is often of interest to analyze the multilevel interdependencies in the life course, for example, by evaluating how macro-social indicators (e.g., unemployment rates) relate to the evolution of individual trajectories.

The unique possibility of studying such interdependencies within the standard SA framework is to build a sequence-type representation of the time-varying covariates, and to apply joint SA, obtaining a joint typology describing the most typical combinations of temporal patterns observed across the considered domains. Unfortunately, this approach is far from being satisfactory. Indeed, as already mentioned (see Section 2.1.1), cluster analysis would also produce a joint partition for independent life domains; therefore, the joint typology in itself cannot be regarded as proof of a relationship (see Studer, 2015, for a discussion). A careful preliminary evaluation of the association among domains (Piccarreta, 2017) is necessary to avoid over-interpretation of the results. In addition, the coding of the time-varying covariates might lead to a reduction in their information content (e.g., for numerical continuous variables, an interval recoding will be necessary), and the results might depend on the chosen coding. Furthermore, in joint SA the trajectories are treated symmetrically, so that conclusions can be drawn only on *mutual* association and not on the possible dependence of the response trajectory on the others. Lastly, but perhaps more importantly, joint SA would highlight the relationship between the *entire* trajectories observed across the considered domains, and would not allow exploring how changes in one domain impact the subsequent evolution of the other/s.

In the next section, we review some recent proposals that by combining SA and EHA, permit studying the effect of time-varying covariates, if not on the entire trajectories, at least on sub-trajectories.

2.3. Combinations of event history and sequence analysis

Approaches combining SA and EHA are based on the idea that weakening the holistic perspective allows studying the medium-term unfolding of trajectories. The "Competing Trajectory Analysis" (CTA) and the "Sequence Analysis Multistate Model Procedure" (SAMM) focus on the relation between sub-trajectories of predefined length and time-varying covariates. Instead, "Sequence History Analysis" (SHA) aims to study the relation between the trajectory experienced up to each observation period and a subsequent event.

2.3.1. Competing Trajectory Analysis (CTA)

In CTA, Studer, Liefbroer, and Mooyaart (2018) consider the sub-sequences following the first transition out of an initial state, which in their application is the first event of the transition to adulthood. An example of the extraction of such sub-sequences is given in panel (a) in Fig. 1.

"Typical sub-sequences" are first identified using SA and cluster analysis. The time spent in the initial state is not taken into account, which allows obtaining a detailed typology of the subsequent process, as less information needs to be summarized via cluster analysis. In a second step, the chance to experience one of the identified typical sub-sequences is estimated using a competing risks model. This allows jointly studying the timing of the focal transition and the subsequent (partial) process "type." Time-varying covariates measured at the beginning of the sub-sequence can be accounted for. CTA is useful when all the sequences begin with the same state (e.g., living with parents) and when the duration of the first state is a key aspect of the trajectories.

2.3.2. Sequence Analysis Multistate Model Procedure (SAMM)

The SAMM procedure was introduced by Studer, Struffolino, and

Fasang (2018) to study the effect of German reunification on employment trajectories among women in East and West Germany. As depicted in panel (b) of Fig. 1, focus is on the sub-sequences of fixed length following any observed transition, so that multiple sub-sequences are possibly extracted from each sequence. Again, SA and cluster analysis are used to identify "typical sub-sequences of medium-term changes." The effect of time-varying covariates on the chance to start each type of sub-sequence and the time spent in each state are then estimated using a multistate model (see Section 3.1 for details). This approach might involve the estimation of several regressions, whose number depends on the number of states in the sequences. However, the authors discuss several strategies for reducing the number of regressions and for simplifying the interpretation of the results.

One of the main advantages of CTA and SAMM is that they can also be applied to censored sequences; this allows including younger individuals in the analysis, and studying their behavior within a limited observation period. This is particularly advantageous in many applications, particularly when studying the transition to adulthood. Furthermore, focus on sub-trajectories allows one to account for a medium-term conception of changes that can describe the individuals' possible anticipations for or awareness of the future. This perspective can be related to the idea of "shadow of the future" (Bernardi et al., 2018).

2.3.3. Sequence History Analysis (SHA)

CTA and SAMM only partially account for the *past trajectory*. Actually, in SAMM, clusters of sub-trajectories are obtained conditional on the state prior to a transition; therefore, the probability of transitioning to a given sub-trajectory is related to the state experienced prior to the transition. Furthermore, indicators summarizing previous experiences can be included among the explanatory variables. However, the sub-trajectory experienced prior to the transition is not fully considered.

Such an aspect is accounted for in SHA, proposed by Rossignon, Studer, Gauthier, and Le Goff (2018). The authors study the probability of experiencing a specific (non-recurrent) event (i.e., leaving the parental home), using a discrete-time EHA model that includes, among the others, a time-varying categorical covariate describing the history experienced up to each time point. Specifically, cluster analysis is applied to partition the trajectories experienced up to each time point into types, and for each individual, the sequence of the visited clusters is registered. Therefore, individuals can change cluster membership over time, even if the same partition is used for all time points. This is a promising approach for studying the "shadow of the past" (Bernardi et al., 2018) on an upcoming event, which can also be extended to include past trajectories related to several life domains among the covariates. However, further work is necessary to extend it to complex situations (e.g., when explanatory trajectories cannot be easily clustered and/or when the clusters cannot be assumed to be constant over time), and to account for recurrent and/or concurrent events.

The methods presented in this section share the advantage of allowing the study of the influence of time-varying covariates at least on sub-sequences. They are clearly well suited in all those cases when future or past sub-sequences are in fact more relevant or more interesting than the entire trajectory. The study of shorter sequences of changes often results in better clustering quality in the SA phase of the procedures. This is surely a positive side effect of this perspective, even if all the considerations made in the previous section for clusters used as inputs in subsequent analyses remain valid. While each procedure is promising in itself, we think that their combination is also worthy of exploration. Indeed, the analysis of the interplay between past trajectories and future sub-sequences will likely shed light on how shadows of the past and of the future intertwine.

3. Model-based analysis of life course trajectories

Alongside the methods grounded in the algorithmic culture, several model-based approaches have been used to study life courses described as sequences. Contrary to SA, these approaches assume the existence of a stochastic process underlying the trajectories' unfolding, and aim to analyze and describe its features. Common to this procedure is the assumption that the complex task of the holistic study of sequences can be efficiently simplified by "decomposing" the whole trajectories into collections of relevant features, such as *transitions* from one state to another and/or *durations* of the visited states. In addition, some simplifying hypotheses are generally made concerning the mechanism relating the past experienced states to the future ones.

Here we focus on two classes of models that permit to study sequence data in a life course perspective, namely, multistate models (MSMs) and latent Markov models. Both approaches allow describing the types of events occurring over time and the relationship between covariates, either baseline or time varying, on the sequence of experienced events and transitions.

Specifically, the class of MSMs includes many popular models (e.g. the competing risk model) that describe the process regulating how individuals move among a finite number of states. Instead, latent Markov models (see Vermunt, Tran, & Magidson, 2008 for a comprehensive review) postulate the existence of a latent and unobservable process, described by a Markov-chain with a finite number of states, which "emanates" and induces the observed states. Interestingly, these models can be regarded as an extension of MSMs, where a multistate model governs the transitions across *latent* rather than across *observed* states.

In the next subsections, we describe the most relevant features of multistate and latent Markov models, highlighting their usefulness for studying trajectories. Finally, in Subsection 3.3, we review possible limitations, some in common and some specific to each approach, and discuss some aspects that should be taken into account for their proper application.

3.1. Multistate models

Multistate models (see Meira-Machado, de Uña-Álvarez, Cadarso-Suárez, & Andersen, 2009, for an effective presentation) comprise a broad family of methods used to study sequences of categorical states by focusing on the time spent in each state and on the transitions out of a given state (Andersen & Keiding, 2002; Putter, Fiocco, & Geskus, 2007; Therneau & Grambsch, 2000). Such states can be transient and possibly recurrent, or absorbing, when transitions to other states are not possible. Fig. 2 reports a typical MSM, the illness-death model without remission, where boxes, circles, and arrows represent respectively transient states (health or illness), absorbing states (death) and possible transitions. The number of possible states and their types impact the possible transitions, and consequently, the complexity of the model. For example, Fig. 3 illustrates an MSM for the transition to adulthood, where states are recurrent and several transitions are possible.

Different assumptions can be made about the dependence of transitions on time. The *Markov* assumption states that the transition probabilities only depend on the history of the process through the current state. In some cases, it can be more realistic to rely on the *semi-*

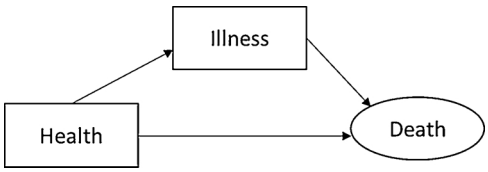

Fig. 2. The illness-death model. Boxes indicate transient states, the circle the absorbing state, and arrows possible transitions.

ARTICLE IN PRESS

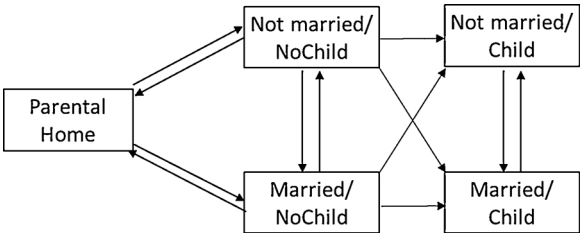

Fig. 3. An MSM to study the unfolding of family trajectories during the transition to adulthood. The model is simplified by ignoring rare transitions, such as going back to the parental home after parenthood.

Markov assumption, which relates the transition probabilities not only to the current state but also on its duration since the last entry. Despite some references in the literature concerning alternative non-Markov models (see e.g., Meira-Machado et al., 2009), the Markov or the semi-Markov properties are most commonly adopted in applied research.

For MSMs in *continuous time,* the data-generating stochastic process is fully characterized by the *instantaneous hazards of transitioning* to different states (given the current state and the past history). Covariates are often included in the model through the transition intensities to explain differences among individuals. A popular choice is the proportional hazards model, having a multiplicative structure with a baseline transition intensity assumed common for all individuals. It can be shown that under proper assumptions, maximizing the likelihood of the entire multistate process is equivalent to maximizing the probability of each transition separately, provided the model's coefficients are transition-specific (Putter et al., 2007). Therefore, in this case, the transitions instantaneous hazards can be estimated using standard EHA methods for competing risks. For instance, the model in Fig. 2 could be estimated by considering three transition intensities and transitions to illness or to death can be regarded as competing risks while being healthy.

Even if transitions occur continuously, in some applications the available data are interval-censored (because individuals are observed at equally spaced time points, for example monthly or yearly). In this situation, a possible approach is to use *discrete-time* MSMs, describing individuals' movements between states in discrete time. In these models, attention is focused on the transition probabilities, and the covariates are typically related to such probabilities through generalized multinomial regression (Agresti, 2002). Discrete-time MSMs are particularly interesting when dealing with sequences, which are typically built registering the state experienced at regular equally spaced time points. An interesting discrete-time MSM, explicitly built to model life courses, is the *State Change Model* (SCM) introduced by Bonetti et al. (2013). The SCM takes into account only transitions to *different* states. The probability of observing a specific sequence is expressed as the combination of discrete time to event distributions and transition probabilities. Specifically, the duration of the k -th visited state, and therefore, the (discrete) time to the next generic transition, is assumed to follow a geometric distribution with a parameter possibly depending both on the covariates available when the k -th state was entered and on the last state visited before entering the k -th one. The probability of transitioning from one state to another is modeled using multinomial logistic regressions, which are allowed to include covariates and the duration of the state visited prior to transition.

Importantly, MSMs allow easily dealing with right-censored data, provided the censoring mechanism is independent from the process. In this case, information on censored individuals can be "completed" based on that on cases without censoring (see Meira-Machado et al., 2009, and their references for in-depth discussion).

Despite their undeniable appeal, the application of MSMs to the study of life courses presents some limitations. The complexity of the model is one of the key issues of MSM, because interpretation can become cumbersome very soon. In its simplest form, the MSM implies modeling one set of coefficients for each transition. This is perfectly manageable for a simple model, such as the illness-death model, which involves only three transitions. However, considering all the possible transitions between a set of s states results in $s(s-1)$ transitions and in as many sets of coefficients. To simplify the MSM, the set of possible transitions is often reduced by ignoring some (rare) transitions, or by considering a reduced set of states. In these cases, the impact of this simplification should be thoroughly justified and discussed. Other aspects to carefully take into account will be discussed in Section 3.2, after describing the latent Markov models.

3.2. Mixed latent Markov models for sequences

In this section, we illustrate latent class models and various specifications of hidden Markov models, which are based on the assumption that a *latent* structure exists underlying the observed sequences. Such latent structure should identify the most relevant features of the trajectories, by filtering out negligible individual differences, which are attributed to "sampling variation" (i.e., due to the probabilistic relation between latent and observed states). From the substantive perspective, the distinction between latent and observed states has an appealing interpretation with reference to life course data. Indeed, life courses can be regarded as the outcomes of life planning, and the latency can reflect plans and/or decisions taken at different stages of life, resulting in the experience of specific observed states (Billari & Piccarreta, 2005).

To describe the most relevant characteristics of such approaches, we refer to the convenient overarching framework offered by the so-called mixed latent Markov model (Vermunt et al., 2008). Such a model can include both a time-constant latent *class,* ω, accounting for the possible partition of cases into (latent) groups, and a sequence of time-varying discrete latent *states*, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_T)$, describing the stochastic process generating the observed states. It is assumed that ω is independent of each latent state, and that $\boldsymbol{\sigma}$ is described by a Markov *chain*, so that σ_t only depends on σ_{t-1} (first-order Markov assumption).

The model is fully characterized by four types of probabilities: the probability of belonging to a certain latent group, the probability of having a particular *initial* latent state, the probability of *transitioning* to a specific latent state at each time-point, and the *emission* probability of observing each actual state, which depends only on the *concomitant* latent state. The case when more domains are considered can be easily accommodated assuming that the states experienced in different domains are all independent conditioned to the latent states. In other words, each latent state emanates the concomitant states observed across all the domains simultaneously.

As underlined by Vermunt et al. (2008), such articulation allows accounting for three relevant aspects in longitudinal data analysis: *autocorrelation*, through the relationship between the time-varying latent states; *misspecification or measurement errors* through the imperfect relationship between latent and observed states; and *unobserved heterogeneity*, through the possible partition of the trajectories induced by the latent class.

This generic model offers a unified framework to address some of the issues described in the previous sections. Covariates (possibly time varying) can be included in the model and, at least theoretically, they can influence each of the probabilities mentioned before (typically using multinomial regression models). Thus, it is possible to relate clusters (latent class) to covariates; however, additional and more articulated relations can be explored too. Consider for example a study on work careers. One could be interested in evaluating whether women present a higher probability of belonging to specific groups, for example, a group characterized by long permanence in the inactive status. Alternatively, one might evaluate whether women experience the transition from a latent state describing participation in the labor market to one describing exclusion from it more frequently. Finally, it is possible to assess whether conditioned to participation in the labor market (latent state), women exhibit a higher propensity to work part-

ARTICLE IN PRESS

time (observed state). Ideally, the relation between sex and work trajectories could be investigated accounting for all these aspects. However, allowing the same variable to influence different probabilities could hinder a proper interpretation and evaluation of its effect. Therefore, it is common practice to relate each covariate only to one type of probability (see Han, Liefbroer, & Elzinga, 2016).

In addition, the model can handle missing values in the trajectories (if covariates are not missing). Cases with partially observed trajectories can contribute to the likelihood when they provide information at a given occasion. This is viable only when data are missing completely at random. Nonetheless, in the case when it is possible to identify the missing data mechanism (that is the underlying cause of missing data) the model can be extended to incorporate it (see Vermunt et al., 2008).

Most of the models used in the life-course literature to study sequence data can be regarded as special cases of a mixed latent Markov model.

Latent class analysis (see e.g. Barban & Billari, 2012) assumes that cases are partitioned into latent classes, with class membership possibly depending on baseline covariates. A class-specific distribution describes the simultaneous occurrence of states across domains, and each trajectory is regarded as the realization of a sequence of independent draws conditional on its class distribution. Thus, within each class, the realization of a state in one domain is independent of all the previously experienced states. This strong assumption implies disregarding possible associations among subsequent events, which is one of the main weaknesses of this approach.[2]

Hidden Markov models (HMMs) that are increasingly employed in life-course research (see, e.g., Bolano, Berchtold, & Ritschard, 2016; Han et al., 2016) constitute an interesting improvement. Fig. 4 provides a graphical representation of a standard HMM, and illustrates the relationships between covariates, latent states, and actual states in the simplified case when only one domain is taken into account. It is worth noting that for the model to be identifiable, it is commonly assumed that the transition probabilities are time-homogeneous (i.e., constant over time).

Observe that HMMs do not include latent classes, and therefore assume that the unobserved heterogeneity can be ignored (Vermunt et al., 2008). Alternatively, Helske, Helske, and Eerola (2016) propose to model sequences using the so-called *Mixture* HMM. Such model assumes that cases are partitioned into groups, and that the sequences within each class are the emanation of class-specific HMMs. In Helske et al. (2016) covariates are allowed to affect only the latent class membership. Therefore, time-varying covariates cannot be included in the model. This surely constitutes a major limitation of the model. In addition, in many applications, one may assume that covariates play different roles, influencing different features of the trajectories (class membership, or latent states, or observed states). Unfortunately, to the best of our knowledge, no statistical software is yet available that allows estimating a "full" HMM conditioned to each class.

While seeming very promising for sequence-data analysis, latent Markov models present some drawbacks. Indeed, the introduction of additional layers of complexity (i.e., latent variables) leads to a large number of model parameters, making the interpretation of the results and of the effect of the covariates difficult. For the same reason, the parameters' estimation might be cumbersome when the data are too sparse. Moreover, the estimates might depend on the initialization of the EM-algorithm used for estimation. Such algorithm might fail to find a global optimum when the likelihood has multiple local maxima. This can be partially solved by running the algorithm with different initial parameters settings, and by selecting a final model based on goodness of fit criteria (such as the AIC or the BIC). In addition, some authors (see, e.g., Helske et al., 2016) propose the preliminary use of SA to individuate reasonable algorithm starting points.

Under a substantive perspective, the estimation of the number of latent states and/or classes can be a relevant issue too. Typically, models with different levels of complexity are contrasted, and the final model specification is chosen based on goodness of fit criteria. In many situations, this leads to select complex models, characterized by a too large number of latent states, for instance. This poses serious problems from the interpretative perspective when one assumes, for example, that the latent states describe the "stages in which the subjects will take demographic decisions" (Han et al., 2016, p. 159).

Other issues, also common to multistate models, are discussed in the next subsection.

3.3. Discussion on model-based approaches: what limitations?

It is important to consider that model-based approaches rely on simplifying assumptions that do not necessarily hold, and which may lead to unreliable results when violated.

The (first-order) Markov assumption implies that the "shadow of the past" is efficiently summarized by the last visited state (or by the last entered state in SCM). Nonetheless, in many applications, the "memory" of the process might be longer, and the combination of experienced states, their ordering, and/or their duration might be relevant. As suggested by Meira-Machado, de Uña-Álvarez, Cadarso-Suárez, and Andersen (2009), for MSM, a way to assess whether such assumption is tenable is to include among covariates variables depending on the past history, such as the last state entered before the current one, or the duration of the current state, and to test their significance. Some proposals deal with higher-order multistate Markov models (see e.g., Berchtold & Raftery, 2002); however, at least to our knowledge, they are not commonly adopted in social sciences, mostly because of the difficulties arising from the high number of parameters and the consequent estimation and interpretation problems. Similar considerations hold for latent Markov models, with the additional problem of a very large number of parameters implied by this solution. While these aspects prevent the wider adoption of higher-order Markov models in life-course research, we are convinced that developments are worthy, both from the theoretical and from the computational point of view.

In addition, the assumption of time-homogeneity can be unrealistic in social sciences, where transitions are typically highly time-dependent. Allowing time-varying covariates (e.g., the age at transition) to influence transitions can at least partially mitigate the impact of such assumption (when violated). Nonetheless, this would imply assuming a linear or quadratic impact of time on transitions, which might not be an adequate specification in some cases.

These aspects are particularly relevant in the assessment of the quality of models, and in the selection of their features (e.g., selection of covariates, and choice of the number of latent states or of the latent classes), which are often based on criteria (e.g., AIC or BIC), which intrinsically relate to assumptions concerning the data-generating process. It is worthwhile to remind that the model is assumed to hold, and that its parameters are estimated at maximizing the coherency between the model and the data. While this is not a serious issue when the assumptions—even if not necessarily met—allow a proper description of the data structure, it is important to understand whether the simplified (and more readable and interpretable) structure arising from a model satisfactorily describes the data. Titman and Sharples (2008, 2009) offer an interesting review of some goodness of fit measures that are adequate both for Markov and for hidden Markov models. Without entering into details that are beyond the scope of this paper, the available proposals aim at evaluating the models' performance with respect to the *transitions*. While this is surely a relevant preliminary investigation when these models are applied to sequence data, the ability of a model to adequately explain/reproduce trajectories unfolding should also be considered.

[2] Han et al. (2016) discuss a formal heuristic to relate the stochastically defined latent classes to the distance-based clusters found with SA.

ARTICLE IN PRESS

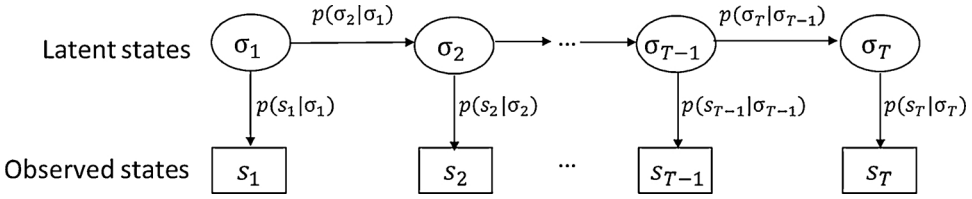

Fig. 4. Graphical representation of an HMM. σ_t and s_t denote the latent and the observed states respectively, $p(\sigma_t\sigma_{t-1})$ denotes the probability of transitioning from one latent state to another at time t (i.e., the Markov chain), and $p(s_t|\sigma_t)$ denotes the probability of observing an actual state at time t given the concomitant latent state (emission probability). All the probabilities can depend on the covariates.

Indeed, this relates to a fundamental aspect that is often overlooked. Models focused on transitions might prevent keeping a holistic perspective on the life course. In a sense, such models are well suited for sequences characterized by "natural ordering" or "natural histories" in the words of Abbott (1992). For example, when family formation patterns are considered, individuals tend to move from being single, to being in a union and to have children, transitions to previous states are relatively rare, and more importantly, the transitions across different states provide a quite exhaustive description of the life course. Nonetheless, this does not hold in all situations. For example, when analyzing processes including recurrent states and transitions, such as work careers, back and forth movements are more frequent and there is not necessarily a predetermined order in transitions. A woman could enter the labor market with a part-time job, or experience a long period of unemployment before getting a job, but she could also alternate between full and part-time jobs or unemployment after the birth of her children. In this case, the adequacy of a model should be assessed also with respect to its ability to reproduce the sequencing of states observed in data. This is even more essential when more trajectories are studied jointly, because in such cases it is also crucial to evaluate whether the model allows for a satisfactory explanation of all the domains or only of a subset of them.

The complexity of the presented model-based approaches can rapidly increase when a high number of states or transitions are considered, and/or when latent class or latent states are added to the model. However, one should bear in mind that—as for any statistical procedure—this additional complexity does not guarantee results that are more meaningful from a sociological viewpoint. Indeed, in some situations, an overly articulated analysis might further complicate the unveiling of patterns in data, and might prevent a meaningful interpretation of the results. These considerations relate to the more general problem of comparing alternative models with respect to their ability of recovering trajectories. This can be difficult because alternative models are typically non-nested, and possibly rely on different assumptions.

Piccarreta, Bonetti, and Lombardi (2018) discuss a possible approach in this direction. They suggest evaluating the goodness of fit of alternative models by simulating, for each observed case, a set of sequences, possibly conditional on covariate values, and measuring the similarity between observed and model-based generated sequences. As in SA, this procedure depends on the criterion chosen to measure the dissimilarity between sequences. On the one hand, such dependence can be considered a limitation; however, sensitivity analyses can be run to evaluate if (and to what extent) the obtained results depend on the chosen dissimilarity measure. On the other hand, the choice of a dissimilarity criterion allows one to emphasize those features relevant to sequence comparisons, according to the researcher. Such dissimilarity-based evaluations can be considered as a way to reconcile the main ideas underlying the standard SA-dissimilarity-based descriptive framework with the increased attention of the scientific community towards (holistic) model-based approaches.

4. Conclusions

The goal of this study was to account for the advantages and limitations of different approaches in studying life courses from a holistic perspective. We distinguished among methods based on SA and its extensions, and model-based procedures, which are increasingly used in the SA community. While the strength of SA is its ability to provide a holistic descriptive view of trajectories, it does suffer from some limitations, mostly related to its use in inferential analysis. Procedures developed within the original SA framework rely, at least to a certain extent, on decisions that need to be taken by the researcher. The most relevant issues include the choice of a dissimilarity criterion, of a clustering algorithm, of the number of clusters, and of the actual way in which results are used (e.g., as input for a multinomial regression, or to draw conclusions about relationships among more domains). This can raise concerns when these decisions are not thoroughly discussed and justified, when instead they need to be well grounded, both theoretically and for the data at hand.

From this perspective, model-based approaches might seem less "subjective," as model specifications are often chosen based on statistical indicators. However, it is important to fully discuss and justify the choice of a specific model, particularly with reference to the adequacy or suitability of the underlying assumptions and to the possible estimation problems that might arise when fitting complex models. It is true that models based on specific assumptions—even if not necessarily met—can lead to good results for a given dataset; nonetheless, this should not be taken for granted. Similarly, one should consider that a model that is theoretically well suited for a specific problem might nonetheless perform very poorly. This suggests the necessity of carefully revising results and their implications, particularly with respect to the original goal of the analysis. For example, when a model for transitions is employed with the ultimate aim of explaining the unfolding of entire trajectories, it is crucial to assess its performance with respect to the original object of interest.

Even if the methods presented in this work can all be employed to study trajectories in their entirety, they focus on different aspects. MSMs and HMM describe the generative process by focusing on transitions and (possibly) on time spent in each states. In some cases, due to the complexity of the model or of the data, this decomposition might result in losing sight of the whole trajectory. Instead, SA and LCA point to the entire trajectories, but do not provide insights on the data generating process. Combinations of SA and EHA hold an intermediate position, focusing on sub-sequences, and therefore, weakening the holistic perspective of SA. Whatever the chosen approach, the methodological focus should fit the substantive research question; this is a crucial aspect in making an informed methodological choice. In turn, choosing a specific approach might itself require further specification or articulation of the goal of the analysis.

The differences in the analytical focuses of the different methods could be exploited in a "sequence of analyses" combining the techniques revised in this paper. A preliminary application of SA would provide insights into the most relevant patterns in the data, and possibly unveil different mechanisms governing the unfolding of the sequences. For instance, a data-driven type of sequence might suggest that being a graduate always precedes reaching a managerial position. Clearly, to generalize this observation (e.g., concluding that graduation is a mandatory step), it is necessary to test it. Under this perspective, model-based approaches offer the opportunity to verify and test considerations based on the evidence provided by SA. The SA results could also provide indications useful for models' specification (e.g., which transitions can be neglected, or how the states can be simplified, or to

ARTICLE IN PRESS

what extent transitions reflects unfolding of trajectories), thus, allowing to properly define the model and avoid over-specification. In addition, SA could prove useful in the identification of covariates whose effects on trajectories are worth consideration.

Ultimately, the choice of a specific model or approach is driven by the researcher's opinions regarding the adequacy of the assumptions for the data at hand, the goals of the analysis, or the possible ease of estimation. Nonetheless, it is important to consider that alternative approaches (e.g., MSMs instead of HMM) could be considered, provided they satisfactorily fit the data. From this perspective, the development of criteria to compare competing models is of key importance and an area that requires further investigation.

Acknowledgment

M. Studer gratefully acknowledges that he benefited from the support of the Swiss National Centre of Competence in Research LIVES – Overcoming vulnerability: Life course perspectives (NCCR LIVES), which is financed by the Swiss National Science Foundation (grant number: 51NF40-160590). =

References

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

Abbott, A. (1992). From causes to events: Notes on narrative positivism. *Sociological Method and Research, 20*, 428.

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

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

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musician's careers. *The American Journal of Sociology, 96*(1), 144–185.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology, review and prospect. *Sociological Methods & Research, 29*, 3–33.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley & Sons.

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*, 430–462.

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research, 11*, 91–115.

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, 61*(5), 765–784.

Berchtold, A., & Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science, 17*(3), 328–356.

Bernardi, L., Huinink, J., & Settersten, S. (2018). The life course cube: A tool for studying lives. *ALCR*.

Billari, F. C., & Piccarreta, R. (2005). Analysing demographic life courses through sequence Analysis. *Mathematical Population Studies, 12*, 81–106.

Bolano, D., Berchtold, A., & Ritschard, G. (2016). A discussion on hidden Markov models for life course data. G. Ritschard, & M. Studer (Eds.). *Proceedings of the international conference on sequence analysis and related methods,* 241–260.

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.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Bürgin, R., & Ritschard, G. (2014). A decorated parallel coordinate plot for categorical longitudinal data. *The American Statistician, 68*(2), 98–103.

Elder, G. H., Kirkpatrick Johnson, M., & Crosnoe, R. (2003). The emergence and development of life course theory. In J. Mortimer, & M. Shanahan (Eds.). *Handbook of the life course, handbooks of sociology and social research*Springer US pp. 3–19.

Elzinga, C., & Studer, M. (2016). Normalization of distance and similarity in sequence analysis. *Sociological Methods & Research* Accepted for publication.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). New York: John Wiley & Sons.

Fasang, A. E., & Liao, T. F. (2014). Visualizing sequences in the social sciences. *Sociological Methods & Research, 43*, 643–676.

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.

Halpin, B. (2016a). Multiple imputation for categorical time series. *The Stata Journal, 16*(3), 590–612.

Halpin, B. (2016b). Missingness and truncation in sequence data: A non-self-identical missing state. G. Ritschard, & M. Studer (Eds.). *Proceedings of the international conference on sequence analysis and related methods,* 443–444.

Han, Y., Liefbroer, A. C., & Elzinga, C. H. (2016). Understanding social-class differences in the transition to adulthood using Markov chain models. G. Ritschard, & M. Studer (Eds.). *Proceedings of the international conference on sequence analysis and related methods,* 155–177.

Helske, S., Helske, J., & Eerola, M. (2016). Analysing complex life sequence data with hidden Markov modeling. G. Ritschard, & M. Studer (Eds.). *Proceedings of the international conference on sequence analysis and related methods,* 209–240.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis, 52*, 258–271.

Hennig, C. (2008). Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis, 99*, 1154–1176.

Hoem, J. M., & Kreyenfeld, M. (2006). Anticipatory analysis and its alternatives in life-course research. Part 1: The role of education in the study of first childbearing. *Demographic Research, 15*, 461–484.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis.* New York: John Wiley & Sons.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR, 163*(4), 845–848 in Russian.

Levine, J. (2000). But what have you done for us lately. *Sociological Methods & Research, 29*(1), 35–40.

Levy, R., Gauthier, J.-A., & Widmer, E. (2006). Entre contraintes institutionnelle et domestique: les parcours de vie masculins et féminins en suisse. *Cahiers canadiens de sociologie, 31*(4), 461–489.

McVicar, V., & 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, 165*(2), 317–334.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., & Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research, 18*(2), 195–222.

Piccarreta, R. (2017). Joint sequence analysis: Association and clustering. *Sociological Methods & Research, 46*(2), 252–287.

Piccarreta, R., & Billari, F. C. (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society Series A, 170*(4), 1061–1078.

Piccarreta, R., Bonetti, M., & Lombardi, S. (2018). Comparing models for sequence data: Prediction and dissimilarities. *Dondena Centre for Research on Social Dynamics and Public Policy working paper,* 113.

Piccarreta, R., & Elzinga, C. H. (2013). Mining for association between life course domains. In J. J. McArdle, & G. Ritschard (Eds.). *Contemporary issues in exploratory data mining in the behavioral sciences. Quantitative methodology*New York: Routledge pp. 190–220.

Piccarreta, R., & Lior, O. (2010). Exploring sequences: A graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society Series A, 173*(1), 165–184.

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society A, 170*(1), 167–183.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multistate models. *Statistics in Medicine, 26*, 2389–2430.

Rossignon, F., Studer, M., Gauthier, J.-A., & Le Goff, J.-M. (2018). Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. In G. Ritschard, & M. Studer (Eds.). *Sequence analysis and related approaches: Innovative methods and applications* (pp. 83–100). Cham: Springer.

Sampson, R. J., & Laub, J. H. (2005). A life-course view of the development of crime. *The Annals of the American Academy of Political and Social Science, 602*, 12–45.

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

Shanahan, M. J. (2000). Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. *Annual Review of Sociology, 26*, 667–692.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy Analysis of State Sequences. *Sociological Methods & Research, 40*(3), 471–510.

Studer, M. (2013). *Weighted cluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES working papers.* 24.

Studer, M. (2015). Comment: On the use of globally interdependent multiple sequence analysis. *Sociological Methodology, 45*(1), 81–88.

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In G. Ritschard, & M. Studer (Eds.). *Sequence analysis and related approaches: Innovative methods and applications* (pp. 223–239). Cham: Springer.

Studer, M., Liefbroer, A. C., & Mooyaart, J. E. (2018). Understanding trends in the transition to adulthood: An application of competing trajectories analysis (CTA). *Advances in Life Course Research, 36*, 1–12.

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, 179*, 481–511.

Studer, M., Struffolino, E., & Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. *Sociological Methodology, 48*(1), 103–135.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model.* New York: Springer.

Titman, A. C., & Sharples, L. D. (2008). A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine, 27*, 2177–2195.

Titman, A. C., & Sharples, L. D. (2009). Model diagnostics for multi-state models. *Statistical Methods in Medical Research, 19*(6), 621–651.

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 pp. 373–385.

Warren, J. R., Luo, L., Halpern-Manners, A., Raymo, J. M., & Palloni, A. (2015). Do different methods for modeling age-graded trajectories yield consistent and valid results? *The American Journal of Sociology, 120*(6), 1809–1856.