# DYNAMIC SEQUENTIAL ANALYSIS OF CAREERS

Fulvia Pennoni[1] and Raffaella Piccarreta[2]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: fulvia.pennoni@unimib.it)

[2] Department of Decision Sciences, Bocconi University Milano, (e-mail: raffaella.piccarreta@unibocconi.it)

**ABSTRACT**: Individuals' life courses may be represented as ordered collections of the activities experienced during a given period. We consider data on the school-work transition for a set of young individuals from Northern Ireland, and show how some interesting features of their careers can be described by a dynamic latent variable model accounting for some unobserved individuals' characteristics, such as attitudes.

**KEYWORDS**: Expectation-Maximization algorithm, typologies, latent Markov model.

## 1 Introduction

In the recent years an increasing attention has been devoted to the study of life courses described as "careers", i.e. as meaningful units. Life course theory suggests (see e.g. Aisenbrey & Fasang, 2010) that, conceptually, trajectory is superior to discrete transitions, because it emphasizes that single events should be understood in their continuity.

Sequence Analysis (SA) is an umbrella term for tools defined to describe careers adopting such holistic perspective, thus studying life courses represented as sequences (ordered collection) of states describing the activities experienced by individuals during a given period. Specifically, SA focuses on the description and identification of the sequences' most salient distinctive features, and, at least in its original formulation, it is based upon criteria to properly evaluate the dissimilarity between sequences. Indeed, the development of SA started with the seminal papers of Abbott and his co-authors (Abbott, 1995, Abbott & Hrycak, 1990), who extended to sociology the optimal matching algorithm, originally introduced in molecular biology to study protein or DNA sequences (Sankoff & Kruskal, 1983).

Whilst from a descriptive point of view the analysis of trajectories is largely adopted, a "holistic" inferential approach has not been individuated yet, and most empirical research in quantitative life course sociology is based on the

analysis of the timing and comparison of transitions. Indeed, the study of the whole careers is complicated by their categorical nature combined with an extremely high number of realizations. Following some recent contributions (for example, Han *et al.* , 2016) we refer to the latent Markov (LM) model proposed by Bartolucci *et al.* , 2013 to study timing, sequencing, and quantum of the events experienced by each individual. Focusing on an historical dataset, we discuss the advantages of this approach, allowing to account for the unobserved individuals' heterogeneity, and emphasize some aspects that practitioners should carefully consider for its proper application.

## 2 Data

We use the data analysed by McVicar & Anyadike-Danes, 2002, collected (in two waves, 1995 and 1999) on a cohort of 712 young people living in Northern Ireland, who were interviewed about the labour market activities experienced from the month in which they were first eligible to leave compulsory education, at the age of 16 (in July 1993), to the age of 22. Individuals' sequences were built describing their activities in the considered 72 months: Unemployment (U), still being at School (S), attending a Further Education college (FE) or Higher Education (HE), being in Training (T), or Employed (E).

The goal of the analysis is to relate such sequences to a set of covariates measured at the age of 16: region of residence, gender, religion (catholic or not), type of secondary school (grammar or not), grades at the end of compulsory education (high or not), father's occupation (manager or not), father's unemployment, cohabitation with both parents. McVicar & Anyadike-Danes, 2002 adopt a two step method at this aim: optimal matching is first used to obtain clusters of careers, and a multinomial logistic regression is then used to relate the probability of experiencing trajectories in each cluster to the explanatory variables.

In the next sections, we describe instead a unique model-based and data driven procedure, which also allows understanding the pattern of change in the data.

## 3 Latent Markov model

Let $Y_{it}$ denote a categorical variable taking $J$ levels, representing the activity experienced by the $i$-th individual at the $t$-th time point, with $i = 1, \ldots, n$, $t = 1, \ldots, T$. Also, let $\boldsymbol{X}_i$ indicate a vector of time-fixed covariates (which can be also be time-varying) for the $i$-th individual.

The LM models are promising candidates for the analysis of life courses, because the observed activities are regarded as the emanation of a latent process representing, for instance, the individual's attitudes which supposedly change over time. Specifically, let $\boldsymbol{V} = (V_1, \ldots, V_T)$ be a latent process following a first-order Markov chain with a discrete number ($k$) of states. We assume that the observed activities are independent one another conditionally to the latent process and that the monthly activities are independent conditionally to the $V_t$'s.

The initial probabilities of the chain and the transition probabilities are assumed to depend on the covariates through multinomial logit models:

$$\log \frac{p(V_1 = v|\boldsymbol{X} = \boldsymbol{x})}{p(V_1 = 1|\boldsymbol{X} = \boldsymbol{x})} = \log \frac{\pi_{v|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \beta_{0v} + \boldsymbol{x}^\top \boldsymbol{\beta}_{1v},\ v \geq 2, \quad (1)$$

$$\log \frac{p(V_t = v|V_{(t-1)} = \bar{v}, \boldsymbol{X} = \boldsymbol{x})}{p(V_t = 1|V_{(t-1)} = \bar{v}, \boldsymbol{X} = \boldsymbol{x})} = \log \frac{\pi_{v|\bar{v}\boldsymbol{x}}}{\pi_{\bar{v}|\bar{v}\boldsymbol{x}}} = \gamma_{0\bar{v}v} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{1\bar{v}v},\ t, v \geq 2. \quad (2)$$

To obtain the maximum likelihood estimates it is first necessary to compute the manifest distribution of the responses conditional to the covariates, through a forward recursion. Subsequently, the model log-likelihood,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i|\boldsymbol{x}_i),$$

(where $\boldsymbol{\theta}$ denotes all the model parameters, including also those related to the manifest distribution) is maximized using the Expectation-Maximization (EM) algorithm, which requires the computation of the quantities involved in the complete data log-likelihood using forward-backward recursions, (for a more detailed description see Bartolucci *et al.* , 2014).

A suitable function of the R package `LMest` (Bartolucci *et al.* , 2017) allows us to check the multimodality of the likelihood function and to choose properly the number of latent states. The function estimates the model many times for a specified range of number latent states and it returns the best model according to the chosen optimality criterion, for example the Bayesian Information Criterion (BIC) (Schwarz, 1978). The latter is based upon the maximum value of the log-likelihood penalized for both the number of cases and the number of free parameters in the model. The standard errors may be obtained by considering the observed information matrix. The `LMest` package also permits to find the patterns of activities with the highest a posteriori probability on the basis of the observed activities and of the estimated model, for each individual in the sample.

## 4 Highlight on some results

An exploratory analysis of the data described in Sec. 2 evidenced that 40 people were always employed, and only 164 people showed identical careers. By applying the LM model with covariates, the BIC criterion lead us to choose a model with $k = 5$ latent states (with maximum log-likelihood value equal to $\hat{\ell} = -16049.34$, and 313 free parameters).

On the basis of the probability of the manifest model, i.e. the conditional probability of the response given the latent variable, the latent states turn out to be $V_S$, $V_E$, $V_{FE}$, $V_{U\&T}$, $V_{HE}$ (the letters indicate the activities described in Sec. 2). Interestingly, $V_{U\&T}$ refers to an alternation between unemployment and firm training, consistently with the observed data structure. The estimates of the $\boldsymbol{\beta}$ parameters in (1) – not reported for space limitations – allow evaluating the impact of covariates on the initial probabilities, that is the probability of starting with a state different from the reference latent state, here $V_S$.

| $\hat{\boldsymbol{\gamma}}_{1v_{U\&T}v}$ | $V_E$ | $V_{FE}$ | $V_{U\&T}$ | $V_{HE}$ |
|---|---|---|---|---|
| Intercept | $-9.09^{**}$ | $-1.08$ | $-3.91^{**}$ | $13.74^{**}$ |
| Male | $-0.37$ | $0.08$ | $-0.23$ | $-0.14$ |
| Catholic | $0.39$ | $-0.36^{**}$ | $-0.55^{\dagger}$ | $-0.47$ |
| Belfast | $0.29$ | $-0.30^{\dagger}$ | $-0.42$ | $-8.58$ |
| South | $-1.43^{*}$ | $-0.23$ | $0.23$ | $-1.00$ |
| SouthEast | $-0.03$ | $-0.24$ | $-0.17$ | $-8.06$ |
| West | $0.18$ | $-0.27$ | $0.13$ | $0.71$ |
| Grammar school | $1.09^{**}$ | $0.09$ | $0.16$ | $1.51^{\dagger}$ |
| Father unemployed | $0.14$ | $-0.38^{\dagger}$ | $-0.07$ | $-7.75$ |
| High grades at school | $1.09^{**}$ | $0.11$ | $0.88^{**}$ | $1.93^{*}$ |
| Father manager | $0.69^{\dagger}$ | $-0.08$ | $0.20$ | $0.85$ |
| Cohabiting with both parents | $0.60$ | $-0.06$ | $-0.20$ | $-0.25$ |

**Table 1.** *Estimates of the logit regression parameters affecting the transition from $V_{U\&T}$ to the other latent states under the chosen LM model. (Significant at 10%($^{\dagger}$), 5%($^{*}$), 1%($^{**}$)).*

For the sake of convenience, for the estimates of the $\boldsymbol{\gamma}$'s coefficients in (2), in Table 1 we limit attention to transitions from $V_{U\&T}$ to the other latent states. It is interesting to observe that the odds ratio related to move towards most 'favourable' states, for example $V_E$, is higher for people who attended grammar school and achieved high grades.

The impact of the covariates on the initial and transition probabilities can

also be explored analyzing the average initial and transition probabilities for people differing with respect to one (or more) features of interest. For the sake of illustration, Table 2 and 3 report results relative to the father type of employment. At the end of compulsory school, individuals whose father was a manager have a lower probability to be employed (i.e., to start with $V_E$) and a higher probability of continuing the studies (i.e., to start with $V_S$). After the first period of time, these individuals have a higher probability of remaining in $V_{HE}$ (higher education) compared to those whose father is not a manager.

| Father manager: $\hat{\pi}_1$ | | | | | Father not manager: $\hat{\pi}_1$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $V_S$ | $V_E$ | $V_{FE}$ | $V_{U\&T}$ | $V_{HE}$ | $V_S$ | $V_E$ | $V_{FE}$ | $V_{U\&T}$ | $V_{HE}$ |
| 0.21 | 0.23 | 0.14 | 0.42 | 0.00 | 0.18 | 0.25 | 0.14 | 0.43 | 0.00 |

**Table 2.** *Estimates of the initial probabilities broken down by father occupation (manager vs not manager) under the chosen LM model.*

| | Father manager: $\hat{\pi}_{v\|\bar{v}}$ | | | | | Father not manager: $\hat{\pi}_{v\|\bar{v}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $V_S$ | $V_E$ | $V_{FE}$ | $V_{U\&T}$ | $V_{HE}$ | $V_S$ | $V_E$ | $V_{FE}$ | $V_{U\&T}$ | $V_{HE}$ |
| $V_S$ | 0.95 | 0.01 | 0.01 | 0.01 | 0.02 | 0.94 | 0.02 | 0.01 | 0.01 | 0.01 |
| $V_E$ | 0.00 | 0.97 | 0.01 | 0.01 | 0.01 | 0.00 | 0.98 | 0.01 | 0.01 | 0.00 |
| $V_{FE}$ | 0.00 | 0.02 | 0.96 | 0.01 | 0.01 | 0.00 | 0.03 | 0.95 | 0.02 | 0.00 |
| $V_{U\&T}$ | 0.01 | 0.05 | 0.03 | 0.91 | 0.00 | 0.00 | 0.04 | 0.02 | 0.94 | 0.00 |
| $V_{HE}$ | 0.00 | 0.02 | 0.07 | 0.00 | 0.98 | 0.00 | 0.01 | 0.08 | 0.00 | 0.89 |

**Table 3.** *Estimates of the transition probabilities from $\bar{v}$ (row) to $v$ (column), broken down by father occupation (manager vs not manager) under the chosen LM model.*

As a further piece of information, an adaptation of the Viterbi (Viterbi, 1967) algorithm is considered to get the optimal decoding which is based on the results of the EM algorithm. For the sake of illustration, we report below the sequence of observed activities $\tilde{\boldsymbol{y}}$ and the corresponding predicted profile $\hat{\boldsymbol{v}}^*(\tilde{\boldsymbol{y}})$ for a catholic male, living in Belfast who attended a non grammar secondary school, achieving low grades at the end of compulsory education, cohabiting with both parents at the age of 16, and whose father was employed as a manager.

$$\begin{aligned} \tilde{\boldsymbol{y}} &= (U(2), T(20), E(2), FE(12), E(36)) \\ \hat{\boldsymbol{v}}^*(\tilde{\boldsymbol{y}}) &= (V_{U\&T}(22), V_E(2), V_{FE}(12), V_E(36)). \end{aligned}$$

Note that the predicted sequence may be also very different from the observed one.

## References


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

ABBOTT, A., & HRYCAK, A. 1990. Measuring resemblance in social sequences. *American Journal of Sociology*, **96**, 144–185.

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**, 420–462.

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2013. *Latent Markov models for longitudinal data*. Boca Raton: Chapman and Hall/CRC press.

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2014. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, **23**, 433–465.

BARTOLUCCI, F., PANDOLFI, S., & PENNONI, F. 2017. *LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data*. (forthcoming on Journal of Statistical Software).

HAN, Y., LIEFBROER, A.C., & ELZINGA, C.H. 2016. Understanding social-class differences in the transition to adulthood using Markov chain models. *In:* RITSCHARD, G., & STUDER, M. (eds), *Proceedings of the International Conference on Sequence Analysis and Related Methods*.

MCVICAR, D., & ANYADIKE-DANES, M. 2002. Predicting successful and unsuccessful transitions from school to work by using sequence methods. *JRSSA*, **165**, 317–334.

SANKOFF, D., & KRUSKAL, J.B. (EDS.). 1983. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–463.

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