Clustering multidimensional life sequences with hidden Markov models

Satu Helske[1], **Jouni Helske**[2], **Mervi Eerola**[3], **and Ioan Tabus**[4]

[1] University of Jyväskylä, Finland
[2] University of Jyväskylä, Finland
[3] University of Turku, Finland
[4] Tampere University of Technology, Finland

In social sciences, sequence analysis is being more and more widely used for the analysis of longitudinal data such as life courses (Abbott and Tsay, 2000; Aisenbrey and Fasang, 2010). Life courses are described as sequences, categorical time series, which constitute of one or sometimes multiple parallel life domains (Gauthier et al., 2010). Sequence analysis is a model-free data-driven method that is used for computing the (dis)similarities of sequences. Often the goal is to find typical and atypical patterns in histories using cluster analysis, usually with Ward's hierarchical method which minimizes the total within-cluster variance (Ward, 1963).

To a great degree, choosing the best clustering result and the number of clusters is a subjective choice. Also, describing, visualizing, and comparing large sequence data with multiple life domains is complex. We suggest methods for using hidden Markov models (Rabiner, 1989) for solving both problems.

The methods are tested and illustrated using data from the German National Educational Panel Survey (NEPS; Blossfeld et al., 2011). We focus on life courses of an age cohort of 1731 individuals born in 1955-1959. Each individual is represented by three sequences corresponding to different life domains: studies and work, partnerships, and parenthood. Sequences contain 434 monthly statuses between the ages 15-50. Altogether 306 individuals have some missing information in one or at most two life domains.

Hidden Markov models are constructed using a single transition matrix (transitions between underlying life stages) but three emission matrices (actual observed states emitted by the life stages). This way the number of observed states is much lower than if the life domains are combined (e.g. 8+5+2=15 states instead of 8*5*2=80 states). Also, only partly observed states that contain missing information in some life domains can be more efficiently accounted for. This allows a more parsimonious representation of the model, as well as a more accurate inference of general life stages.

At first, Ward's agglomerative clustering algorithm based on generalized Hamming distances (Hamming, 1950) is used to give initial clustering solutions to start with. We test two different options.

1. The initial clustering solutions are assumed fixed, and for each cluster an independent hidden Markov model is estimated. The goal is to find

the correct number of clusters and hidden states and to find the hidden state paths for individuals. This is a simpler approach in terms of computations and interpretability.

2. The initial clustering solution is used to calculate the membership probabilities of each cluster and other starting values of block-HMM which is fitted to the whole data. In the block-HMM the transition and emission matrices are independent block matrices and the initial hidden states of each sequence define the cluster memberships. Although computationally heavier than the first option, this is an intuitively simpler approach which takes account of clustering uncertainty in a probabilistic way – the membership is not fixed but a random variable based on the initial states.

After finding the final HMMs, we want to find the most probable paths of hidden states for each individual, as well as the optimal hidden state path given all the sequences in the cluster, i.e. the representative sequence. The first goal is achieved by the standard Viterbi algorithm (Viterbi, 1967). For the latter goal we present a multivariate extension of the Viterbi algorithm.

Keywords: Clustering, Hidden Markov models, Social sequence analysis.

References:

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

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.

Blossfeld, H.-P., Robach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift fr Erziehungswissenschaft, 14.* Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

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

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal 29*, 147–160.

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

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

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 236–244.