Sequence analysis with Variable Length Markov Chains

Extended Abstract
Lausanne Conference On Sequence Analysis (LaCOSA)
University of Lausanne, June 6-8, 2012

Alexis Gabadinho
Institute for Demographic and Life Course Studies (IDEMO)
University Geneva

May 9, 2012

Sequence analysis in the social sciences mostly relies on the computation of pairwise edit distances (e.g. Optimal Matching) or distances based on common attributes. Once a pairwise distance matrix is obtained the usual next step is to submit it to a clustering procedure or to the more recently introduced procedures for discrepancy analysis [13]. An important part of the recent methodologically oriented litterature on sequence analysis is devoted to Optimal Matching parameters setting or development of new distance measures [4, 5, 6].

This orientation can be traced back to Abbott's work that introduced sequence analysis methods in the social sciences. He distinghishes between approaches that treat sequences step by step and an approach that consider sequences as whole units [1]. The central issue in whole sequence analysis is to find (typical) patterns in the sequences. To achieve this goal the solution proposed by Abbott, which he calls the metric approach, consists in using a measure of resemblance (Optimal Matching distance) between sequence giving access to standard classification method.

Markov chain models are one of methodological tools classified by Abbott in the "step by step" methods. Such models have been used for longitudinal data analysis for a long time, especially for modelling individual or collective behavior like geographical or professional mobility [11, 12, 2, 7]. In this framework categorical sequences encoding life histories can be seen as the result of a stochastic process in which the probability of occurrence of one state depends on a given amount of the past states, called the order of the model. This theoretical framework is intuitive for and fits most of the social processes considered in the social sciences.

Standard, fixed length Markov chain models suffer however from severe drawbacks. One of them is the exponentially increasing complexity of the models when their order, that is the amount of considered past states, increase. Another drawback is the stationarity assumption that is usually made to gain analytic simplicity. This assumption restricts the conditional probabilities to be fixed in time, which usually represent an important sacrifice in realism and flexibility [10].

We propose to explore a class of models called Variable Length Markov

Chains (VLMC) [8, 9, 3] that overcome the complexity problem of fixed length Markov Chains by allowing the amount of considered past states (the memory) to vary according to a particular *context*. The gain in flexibility together with the ease of estimation of these models allow to study high order dependencies in the sequences that may appear in many life course domains.

Although the original framework for such models assums stationarity, it is possible to extend these models to the non stationary case, where the conditional probability of occurence of one state varies according to time, that is the position in the sequence, allowing a much more realistic modelling in many cases.

Variable length Markov Chains are stored in a construction called Probabilistic Suffix Tree (PST). Graphical representations of a PST are a usefull description of the process supposed to have generated a set of sequences, and thus of the underlying social process producing life courses. The main outcome of the model is to assign a probability for each observed state of an individual sequence, the whole sequence probability being simply the product of the state probabilties it is composed of.

Once the probibility of each sequence has been predicted, one can easily extract typical (most likely) patterns as well as outliers sequences. Morevover, separate or segmented models can be fitted for subgroups of the population or even single sequences, allowing to efficiently compare individuals or subpopulations defined by covariates (age, sex, socioeconomic level). PST based sequence clustering can be developed, either by computing pairwise distance matrices or by using model-based clustering. Other possible usages include imputation of missing states.

Variable Length Markov Chains models therefore escape Abbott's dichotomic classification between step by step and whole sequence methods. They are a way of considering sequence resemblance that can be an alternative to the dominant 'metric approach' introduced by Abbott, which has suffered and is still suffering many criticism.

References

[1] Andrew Abbott. Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21:93–113, 1995.

[2] Jr Brent, Edward E. and Richard E. Sykes. A mathematical model of symbolic interaction between police and suspects. *Behavioral Science*, 24(6), Nov 1979.

[3] Peter Bühlmann and Abraham J. Wyner. Variable length markov chains. *The Annals of Statistics*, 27(2):pp. 480–513, 1999.

[4] Jacques-Antoine Gauthier, Eric D. Widmer, Philipp Bucher, and Cédric Notredame. How much does it cost? Optimization of costs in sequence analysis of social science data. *Sociological Methods and Research*, 38:197–231, 2009.

[5] Matissa Hollister. Is Optimal Matching Suboptimal? *Sociological Methods Research*, 38(2):235–264, 2009.

[6] Laurent Lesnard. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods Research*, 38(3):389–419, 2010.

[7] Ronald W. Manderscheid, Donald S. Rae, Anne K. McCarrick, and Sam Silbergeld. A stochastic model of relational control in dyadic interaction. *American Sociological Review*, 47(1):pp. 62–75, 1982.

[8] J. Rissanen. A universal data compression system. *Information Theory, IEEE Transactions on*, 29(5):656 – 664, sep 1983.

[9] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.

[10] Burton Singer. Estimation of nonstationary markov chains from panel data. *Sociological Methodology*, 12:319–337, 1981.

[11] Burton Singer and Seymour Spilerman. Social mobility models for heterogeneous populations. *Sociological Methodology*, 5:pp. 356–401, 1973.

[12] Burton Singer and Seymour Spilerman. The representation of social processes by markov models. *American Journal of Sociology*, 82(1):pp. 1–54, 1976.

[13] Matthias Studer, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40(3):471–510, August 2011.