Title of the envisaged research: The multilevel explicit-duration hidden Markov model for real time behavioural data
Applicant(s) (co-promotor(es): Dr. Emmeke Aarts
Promotor(es): Prof. Irene Klugkist
Department: Methodology and Statistics

Description of the research idea: Due to technological advances, it becomes increasingly easy within the social sciences to collect data on behavior as it unfolds in real time, measured for a prolonged period of time. Take for instance the interaction between a therapist and a patient: we can automatically annotate different types of nonverbal communication based on a video recording, for every second for a period of 15 minutes. Other methods that can be used for collecting real time behavioral data are, for example, experience sampling, GPS tracking, and accelerometer data. These new data enable a novel perspective on investigating behavior: studying the dynamics of behavior over time. This in contrast to the static summaries of behavior that are currently typically obtained.

To extract the dynamics of behavior over time, the statistical model of choice is the hidden Markov model (HMM; Rabiner, 1989; Zucchini, MacDonald & Langrock, 2016). HMMs are a machine learning method that have been used for several decades in many different scientific fields, such as speech recognition and DNA segmentation. Within the social sciences, the HMM is still a rarely used statistical method. When applied to real time behavioral data, it enables one to extract latent (i.e., hidden) behavioral states over time - based on one or several dependent variables - and model the dynamics of behavior over time. This model shows great potential for application to data collected within the social sciences, and answering new research questions.

To make the HMM the perfect match for real time behavioral data, the conventional HMM must be extended in two ways. First, the HMM is extended to the multilevel framework such that we can model the observed sequences of multiple subjects simultaneously, and are able to investigate how the dynamics in behavior are influenced by covariates (see for example de Haan-Rietdijk et al., 2017). That is, the conventional HMM is typically used to analyze only one long sequence of data, such as one string of DNA or one speech sequence. Second, the durations of the latent behavioral states need to be explicitly modeled and allowed to deviate from a geometric distribution by using an explicit duration HMM (ED-HMM; Guédon, 2003). In the conventional HMM, it is implicitly assumed that a shorter duration of a (behavioral) state is always more probable than a longer duration, which is not a very good match with behavioral data. The ED-HMM within the multilevel framework (in which all model parameters are allowed to be random) is a novel method and not yet described in literature, and is a viable method as shown by extensive preliminary results (Aarts, 2016).

Research problem: First, the multilevel ED-HMM needs to be further developed and implemented. That is, the multilevel ED-HMM still requires an optimal estimation algorithm. The algorithm typically used for explicitly modeling the state durations is computationally very intensive. Especially when used within a multilevel and iterative Bayesian estimation framework, this results in computation times that are not user friendly. To improve the computational speed, several promising options are available. Second, an (open source) and user friendly software package to apply the multilevel ED-HMM should be developed. Third, guidelines for applied researchers for both the multilevel HMM and the multilevel ED-HMM on proper use of the models (e.g., required sample sizes) are still missing, and should thus be developed. This information is vital to obtain unbiased and reliable estimates of the model parameters, and a reasonable level of statistical power to detect the influence of a covariate on model parameters.

Goals:

1. Improve the algorithm for estimating the multilevel ED-HMM to reduce computational intensity while maintaining robust and unbiased estimation performance.

2. Develop a user friendly and open source software package such that applied researchers can use the developed statistical method.
3. Investigate on how many subjects observational sequences should be collected and how long these observational sequences should be when applying the multilevel ED-HMM.
4. Investigate how the required sample size of the ED-HMM depends on the complexity (e.g., number of hidden states, number of dependent variables) of the data.

Methods: The multilevel ED-HMM is implemented within the statistical package R (and partly in C++). Regarding the specific research questions:

1: A literature study will be conducted to narrow down optimal possibilities to improve the algorithm for estimating the multilevel ED-HMM. A small number of possibilities will be implemented and tested using simulation studies.

2: An official R package will be developed, including an extensive tutorial and workshop.

3 & 4: Simulation studies will be conducted.

Rationale and approach: The project consists of methodological research, in which new methods are developed, implemented (i.e., programmed in R), and applied. Collaboration with applied researchers will be maintained and actively sought during the entire period. This to ensure that developed methodologies meet the needs of the field, the operationalization of the research questions correspond to realistic settings, and the developed method is applied to real data.

Institutional environment: The PhD student will be employed at the Department of Methodology and Statistics of the Faculty of Social and Behavioural Sciences at Utrecht University. The research of the Department of Methodology and Statistics focuses on a broad array of methods and techniques for the social and behavioural sciences and comprises topics like: longitudinal studies, Mplus and multilevel analyses, collection and analysis of intensive big data, survey research, research synthesis techniques, best practices when doing research, and qualitative research and mixed methods research. In addition, the Department of Methodology & Statistics provides teaching in methods and statistics within all bachelor's and master's degree programmes of the Faculty of Social and Behavioural Sciences and University College Utrecht. The department also advises staff and students with respect to their research activities. The PhD student will find a dynamic and pleasant working environment, in a group that is actively involved in scientific research at the highest international level. In addition, the PhD students will become a member of the Interuniversity Graduate School of Psychometrics and Sociometrics (www.iops.nl).

Relevance: To unravel the complex interactions between psychological characteristics and/or external influences and the behavior of an individual, real time behavioral information provides an unprecedented wealth of information. Due to technological advances, it becomes increasingly easy to collect this type of data. However, to optimally exploit the information present in real time behavioral data, statistical approaches that go beyond the standard statistical test are required. The multilevel ED-HMM provides the perfect match to summarize such data and extract novel information: it allows one to model the dynamics of behavior over time, and quantify and predict heterogeneity in behavioral dynamics between subjects.

References:

Aarts, E. (2016). *Beyond the average: Choosing and improving statistical methods to optimize inference from complex neuroscience data*. VU University, Amsterdam.

de Haan-Rietdijk, S., Kuppens, P., Bergeman, C. S., Sheeber, L. B., Allen, N. B., & Hamaker, E. L. (2017). On the use of mixed Markov models for intensive longitudinal data. *Multivariate behavioral research*, *52*(6), 747-767.

Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, *12*(3), 604-639.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.