gazeHMM: Parameter recovery simulation

Malte Lüken, Simon Kucharsky, & Ingmar Visser

08.05.2020

In this document, we are preregistering a simulation study to estimate the parameter recovery of a hidden Markov model (HMM) which is part of our recently developed algorithm gazeHMM. The purpose of this algorithm is to classify eye movements into distinct events. These events include fixations, saccades, and optionally postsaccadic oscillations (PSOs) and smooth pursuits. The full algorithm will be available on GitHub (www.github.com/maltelueken/gazeHMM) upon publication.

In general, gazeHMM consists of three steps: First, the raw gaze data is preprocessed. Missing samples are identified, the data are filtered, and for each sample the velocity, acceleration, and difference in angle to the next sample is calculated. Second, the HMM estimates the parameters of response, transition, and initial state models for the specified eye movement events. Moreover, the HMM computes the posterior probability of each sample for belonging to each event. The samples are labeled as the event with the highest posterior probability. Third, a postprocessing routine relabels samples that are violating theoretical and practical constraints. This simulation study will only consider the parameter recovery of the HMM but not the performance of the preprocessing and postprocessing routines in the algorithm.

Model Description

The HMM estimates non-observable states that have generated the data (CITE). In the context of eye-tracking, each state corresponds to an eye movement event. In gazeHMM, the model can have between two and four states. State one always corresponds to fixations, state two always to saccades, state three to PSOs, and state four to smooth pursuits. Moreover, the HMM is multivariate but conditionally independent on the states. Thus, it has three independent response models for every state describing velocity, acceleration, and angle. The velocity and acceleration response models for all four states use two-parameter gamma distributions (shape and scale parametrization). The angle response model for the fixation state uses a uniform distribution, while the other three states use a two-parameter von-Mises distribution (mean and concentration parametrization) to describe angle. Both the transition model and the initial state model use a multinomial distribution (with the logit and the identity link function, respectively). Note that no covariates are included in the model, only intercepts for the parameters are estimated.

The HMM is implemented in R (CITE) using the package depmixS4 (CITE). An expectation-maximization algorithm estimates the parameters given the hidden states, which are estimated through the Viterbi algorithm (CITE). The parameters of each response model (except for the uniform distribution) are estimated through maximum likelihood with a spectral projected gradient method (CITE) and Barzilai-Borwein step lengths (CITE) as implemented in the package BB (CITE).

Parameter Recovery

To estimate parameter recovery of the HMM, it will repeatedly generate data with a set of parameters (true parameter values). The same model will then be applied to estimate the parameters from the generated data (estimated parameter values). We will compare the true with the estimated parameter values to assess

whether a parameter was recovered by the model. Additionally, we will compare the true states of the HMM with the estimated states to judge how accurate the model recovers the states that generated the data.

Starting values

The HMM will always start with a uniform distribution to estimate the initial state and state transition probabilities. To generate random starting values for the estimation of shape, scale, and concentration parameters, we will use gamma distributions with the true parameter as the shape parameter and a scale parameter of $\beta = 1$. The gamma distributions ensure that the starting values will be positive. Mean parameters of the von-Mises distribution will always start at their true value.

Design

Parameter Variation

The simulation study will be divided into four parts. In the first part, we will vary the parameters of the HMM. For models with $k \in \{2, 3, 4\}$ states, $q \in \{10, 15, 20\}$ parameters will be varied respectively. For each parameter, the HMM will generate 100 data sets with $N = 2500$ samples and the parameter varied in a specified interval in equidistant steps. This will result in $100 \times (10 + 15 + 20) = 4500$ recoveries. Only one parameter will be varied at once, the other parameters will be set to their fixed values. We will not manipulate the initial state probabilities because these are usually irrelevant in the context of eye movement classification. For the transition probabilities, we will only vary the probabilities for staying in the same state (diagonals of the tranistion matrix) to reduce the complexity of the simulation. The left probability mass will be split evenly between the probabilites for switching to a different state (per row of the transition matrix). Moreover, we will not modify the mean parameters of the von-Mises distribution: As location parameters, they do not alter the shape of the distribution and they are necessary features for the HMM to distinguish between different states.

Where it was possible, we chose the intervals and fixed values for each parameter based on reported ranges for eye movement events in the literature. Otherwise, intervals and fixed values were chosen according to theoretical reasons and based on insights from data used to develop the model. Table 1 shows the theoretical ranges for each event and Table 2 shows the intervals and fixed values for each parameter in the simulation. Parameters were divided by 10 (compared to the reported ranges) to improve fitting of the gamma distributions. We will set the intervals for shape parameters of the gamma distribution for all events to [1,5] to examine how skewness influences the recovery (shape values above 5 approximate a symmetric distribution). The scale parameters will be set so that the respective distribution will match the ranges reported in the literature. Since the concentration parameters of the von-Mises distribution are the inverse of standard deviations, they will be varied on the inverse scale.

Sample Size and Noise Variation

In the second part, we will vary the sample size of the generated data and the amount of noise added to it. The model parameters will be set to their fixed values. For models with $k \in \{2, 3, 4\}$ states and sample sizes of $N \in \{500, 2500, 10000\}$, we will generate 100 data sets ($100 \times 3 \times 3 = 900$ recoveries). We chose samples sizes roughly corresponding to small, medium, and large eye-tracking data sets of a single participant and trial. To simulate measurement error, we will add Gaussian white noise with $\mu = 0$ and $\sigma \in [1, 5]$ varying between data sets to velocity and acceleration. Values that decrease to zero or below will be set to 0.001. We will add white noise from a von-Mises distribution with $\mu = 0$ and $\kappa \in 1/[0.1, 10]$ varying between data sets to angle.

Variation of Starting Values

In the third part, we will increase the variation in the starting values used for parameter estimation. The model parameters will be set to the fixed values. For the shape, scale, and concentration parameters, we will simultaneously increase the scale parameters of the starting value gamma distributions: For $k \in \{2, 3, 4\}$ states and $\beta \in \{2, 4, 8\}$, 100 data sets with $N = 2500$ samples will be generated each ($100 \times 3 \times 3 = 900$ recoveries).

Missing data

In the last part, we will set intervals of the generated data to be missing. The model parameters will be set to the fixed values. For $k \in \{2, 3, 4\}$ states and $n_{miss} \in \{1, 3, 5\}$ intervals, 100 data sets with $N = 2500$ samples will be generated ($100 \times 3 \times 3 = 900$ recoveries). The size of the missing data interval $m \in [1, 250]$ samples will vary between the data sets.

Recovery Measures

For each parameter, we will calculate the mean squared error (MSE) between the true and estimated parameter values. Additonally, we will apply a bivariate linear regression with the estimated parameter values as the dependent and the true parameter values as the independent variable to each parameter that has been varied on an interval. The MSE and the regression coefficients will serve as descriptive measures for parameter recovery. Lower MSE values, regression intercepts closer to zero, and regression slopes closer to one will be interpreted as better recovery.

To assess state recovery, we will compute the proportion of states (for all states taken together, not for each state seperately) that have been classified correctly by the model for each generated data set. Higher proportions will be interpreted as better model accuracy. Overall, we require a median accuracy of 0.9 for the model to be satisfactorily accurate.

Table 1

Theoretical ranges of response variables used to generate parameter values

Event	Resp. variable	Range	Reference
Fixation	Velocity	<30	Larsson et al. (2013)
Fixation	Acceleration	<15	-
Fixation	Angle	uniform	-
Saccade	Velocity	30-500	Larsson et al. (2013)
Saccade	Acceleration	10-250	-
Saccade	Angle	~0	Pekkanen & Lappi (2017)
PSO	Velocity	20-100	Larsson et al. (2013)
PSO	Acceleration	10-90	-
PSO	Angle	$\sim\pi$	Pekkanen & Lappi (2017)
Smooth pursuit	Velocity	<100	Larsson et al. (2013)
Smooth pursuit	Acceleration	<15	-
Smooth pursuit	Angle	~0	Larsson et al. (2013)

Note. Units are °/s (velocity), °/s^2 (acceleration), and radians (angle). ~ indicates that the distribution has a peak at this value.

Table 2

HMM parameter values for generating the data

State	Parameter	Interval	Fixed	Description
1-4	p_i^+	-	$1/k$	Initial state probability for starting in state i
1-4	$a_{i=j}$	[.01,.99]	0.9	Transition probability for staying in the previous state i
1-4	$a_{i\neq j}$	$(1-a_{i=j})/(k-1)$	$0.1/(k-1)$	Transition probability for switching to from state i to a different state j
1	α_{vel}	[1,5]	3	Shape parameter of the velocity gamma distribution
1	β_{vel}	[0.1,0.6]	0.35	Scale parameter of the velocity gamma distribution
1	α_{acc}	[1,5]	3	Shape parameter of the acceleration gamma distribution
1	β_{acc}	[0.05,0.25]	0.15	Scale parameter of the acceleration gamma distribution
1	a^+	-	0	Minimum of the uniform distribution
1	b^+	-	2π	Maximum of the uniform distribution
2	α_{vel}	[1,5]	3	
2	β_{vel}	[5,25]	15	
2	α_{acc}	[1,5]	3	
2	β_{acc}	[1,5]	3	
2	μ^+	-	0	Mean parameter of the von-Mises distribution
2	κ	$1/[0.1, 10]$	1	Concentration parameter of the von Mises distribution
3	α_{vel}	[1,5]	3	
3	β_{vel}	[1,5]	3	
3	α_{acc}	[1,3]	2	
3	β_{acc}	[1,3]	2	
3	μ^+	-	π	
3	κ	$1/[0.1, 10]$	1	
4	α_{vel}	[1,5]	3	
4	β_{vel}	[1,2]	1.5	
4	α_{acc}	[1,5]	3	
4	β_{acc}	[0.05,0.25]	0.15	
4	μ^+	-	0	
4	κ	$1/[0.1, 10]$	1	

Note. Parameters marked with $^+$ will not be varied but always set to their fixed value. k is the number of states in the model.