gazeHMM: Parameter recovery simulation

Malte Lüken, Simon Kucharsky, & Ingmar Visser

07.05.2020

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. 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, we repeatedly generate data with the model under different conditions and different numbers of states. For every condition and number of states $k \in \{2, 3, 4\}$, the model will generate $D = 100$ data sets with parameters varied in a specified interval in equidistant steps.

The same model is then applied to estimate the parameters from the generated data. Then, the estimated parameters are compared with the "true" parameters that generated the data. We will assess parameter recovery descriptively by computing the mean squared error between the estimated and true parameter values separately for each parameter. Additionally, we will apply a bivariate linear regression for each parameter with the true parameter values as the independent and the estimated parameter values as the dependent variable. For a parameter to be recovered, we expect the regression intercepts to be close to zero and the regression slopes to be close to one. Next to parameter recovery, we will also examine the accuracy of the classficiation through the model. Therefore, we will compare the true states with the estimated states for each data set by using Cohen's Kappa. We will treat Kappa values above 0.8 as satisfactorily accurate.

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 $\beta = 1$. Mean parameters of the von-Mises distribution will always start at their true value.

In the first part of the simulation study, we will estimate parameter recovery under ideal conditions. These imply a medium sample size of $N = 2500$ samples per data set, no noise added to the data, and no missing samples. The parameters that we will use to generate the data are displayed in Table 1. We will vary one parameter at a time on the given interval while setting the other parameters to their fixed value. The initial state probabilities will always stay at their fixed value. We will only vary the transition probabilities for staying in the same state. The remaining probability mass will be evenly distributed across the probabilities for switching to a different state (see Table 1). For two states 10, for three states 15, and for four states 20 parameters will be varied one at a time. This simulation will result in $D_{total} = 100 \times (10 + 15 + 20) = 4500$ recoveries.

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 the fixed values from Table 1. For sample sizes of $N \in \{500, 2500, 10000\}$, we will generate data sets and add white noise. To velocity and acceleration, we will add Gaussian noise with $\mu = 0$ and $\sigma \in [1, 25]$. Noise from a von-Mises distribution with $\mu = 0$ and $\kappa \in 1/[0.1, 10]$ will be added to angle. This simulation will result in $D_{total} = 100 \times 3 \times 3 = 900$ recoveries.

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 from Table 1. For the shape, scale, and concentration parameters, we will increase the scale parameters of the gamma distributions to $\beta \in \{2, 4, 8\}$ all at the same time. The sample size will be $N = 2500$ and no noise will be added to the data. This simulation will result in $D_{total} = 100 \times 3 \times 3 = 900$ recoveries.

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 from Table 1. The size of the missing data interval will be $m \in [1, 100]$ and the number of intervals will be $n_{miss} \in \{1, 3, 5\}$. The sample size will be $N = 2500$ and no noise will be added to the data. This simulation will result in $D_{total} = 100 \times 3 \times 3 = 900$ recoveries.

Table 1

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

State	Parameter	Interval	Fixed	Description
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.