gazeHMM: Parameter recovery simulation

Malte Lüken, Simon Kucharsky, & Ingmar Visser

5 5 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 (e.g., fixations). The full algorithm will be available on Github 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 the events. Each sample is labeled as the event with the highest posterior probability. Third, a postprocessing routine relabels samples that are violating theoretical and practical constraints. Since the HMM might not model the data perfectly, this step improves classification performance. This simulation study will only consider the parameter recovery of the HMM but not the performance of the preprocessing and postprocessing routines of the algorithm.

Model description

The HMM estimates unobservable 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 postsaccadic oscillations (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 model 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.

Depending on the number of states, the HMM has a different number of parameters. The response models of the HMM contain six parameters per state. The transition models have k^k parameters in total with k being the number of states. The initial state models have one paramter per state. In sum, HMMs with two, three, and four states have X, X, and X paramters. The denotation of each parameter is shown in Table 1.

Optimization

Table 1

Parameter	Description
π	Initial probabilities for state i
a_{ij}	Transition probability from state i to j
$	

Model parameter recovery

To estimate parameter recovery of the HMM, we repeatedly generate data with the model under different conditions and a different number of states. For every condition and number of states $k \in \{2, 3, 4\}$, the model will generate $D = 100$ data sets with parameters evenly distributed on a given interval. The same model is then applied to estimate the parameters from the generated data. Afterwards, the estimated parameters are compared with the "true" parameters that generated the data. We treat a parameter as successfully recovered when the 95% confidence bands of the estimated parameters include the true parameters. We further require a 90% median rate of correct classifications across the data sets for the model to be satisfactorily accurate.

The HMM will always start with a uniform distribution to estimate the initial state and state transition probabilites. 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 condititions. These imply an average 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 2. We will vary one parameter at a time on the given interval and set the other parameters to their fixed value. This simulation will result in $D_{total} = 100 \times 3 \times 20 = 6000$ recoveries.

In the second part, we will vary the sample size of the generated data and the amount of noise added to it. 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. 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 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 2

State	Parameter	Interval	Fixed
1-4	p*	-	$1/k$
1-4	$a_{i=j}$	[.01,.99]	0.9
1-4	$a_{i \neq j}$	$(1 - a_{i=j})/(k-1)$	$0.1/(k-1)$
1	α_{vel}	[1,5]	3
1	β_{vel}	[0.5,1]	0.75
1	α_{acc}	[1,5]	3
1	β_{acc}	[0.05,0.5]	0.15
1	a*	-	-
1	b*	-	-
2	α_{vel}	[1,5]	3
2	β_{vel}	[5,15]	10
2	α_{acc}	[1,5]	3
2	β_{acc}	[1,5]	3
2	μ*	-	0
2	κ	$1/[0.1, 10]$	1
3	α_{vel}	[1,5]	3

State	Parameter	Interval	Fixed
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 values. Parameters a and b (the minimum and maximum of the uniform distribution) are always obtained from the data.