apps and tzakiris 2013 model

niehaush

May 2020

Contents

1 Reverse engineering of depicted Data **2**
1.1 Figure 1 2
1.1.1 Figure 1 ('b'), bottom-left 3
1.1.2 Figure 1 ('c'), bottom-right 4
1.2 Figure 4, suppl. mat. 4
1.3 Introduction of Blocks 5
1.4 Introduction of new facial IDs over the experiment 5
1.5 Pseudo-code for data-generation 6
1.5.1 Introduction of new stimuli over the experiment 6
1.5.2 Distribution of the stimuli IDs 8

2 Component description **9**
2.1 Response model 9
2.2 Contextual familiarity 9
2.3 View (in)-dependent familiarity 9

3 Model description **10**
3.1 view-independent x context aka 'winning model' 10
3.1.1 view-Independence 10
3.1.2 context dependence 11
3.1.3 view-independent x context parameter summary 11
3.2 View-dependent 12
3.2.1 Parameters 12
3.3 View-dependent x context 12
3.3.1 Parameters 13
3.4 View-independent 13
3.4.1 Parameters 13
3.5 View-independent x view-dependent 13
3.5.1 Parameters 14
3.6 View-dependent x view-independent x context-dependent 14
3.6.1 Parameters 14

4 Computation **15**
4.1 Parameters 15
4.2 Optimization 15
4.3 Author implementation 15
4.4 Replication implementation 15
4.5 Model selection 15

5 Power analysis **16**
5.1 Computational Modeling 16
5.2 Neuroimaging 17

1 Reverse engineering of depicted Data

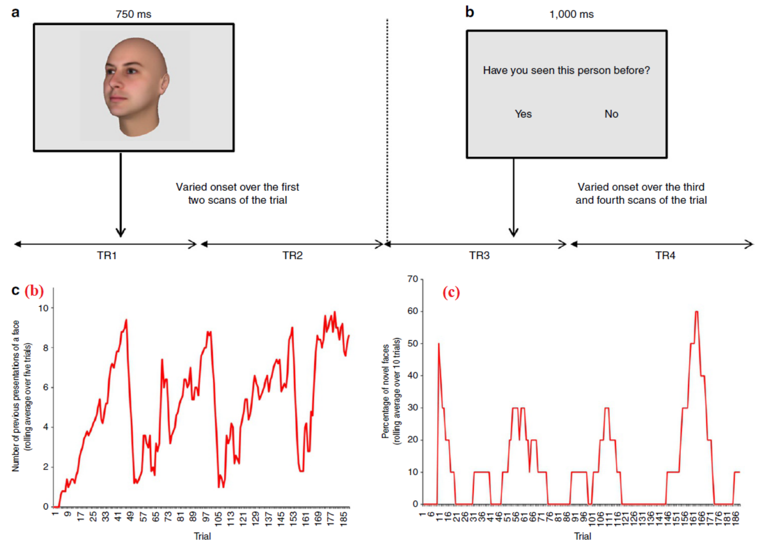
The authors did not respond to initiation of contact and subsequently did not provide information regarding:

- Within-Task/ Trial dynamics etc

To generate testable new presentation data we graphically extracted data from three figures found in the document.

1.1 Figure 1

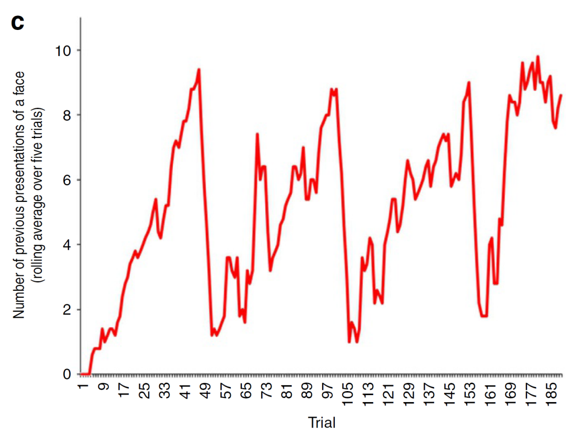
Note, that the description of the entire figure 1 is somewhat misleading: the letter 'b', indicating a sub-figure of figure 1, is positioned in the right upper corner, but the explanation in the description below the entire figure 1 clearly is focused on the sub-plot on the bottom left. Here it is assumed, that the letter 'b' was misplaced and actually belongs to the sub-plot on the bottom left (from here on referred to as plot 'b'). Additionally, subplot 'c' is placed next to the bottom-left subplot ('b') but the figure-description below clearly refers to the subplot on the bottom right. Therefore this bottom-right subplot will be referred from now on as subplot 'c' as it also matches the description. The adapted figure is given below:



Note, that our inferred subplot order is depicted in red in the figure directly above.

1.1.1 Figure 1 ('b'), bottom-left

The original figure is reproduced below:



The description from the publication on pg. 3 regarding figure 1 b on the left side reads:

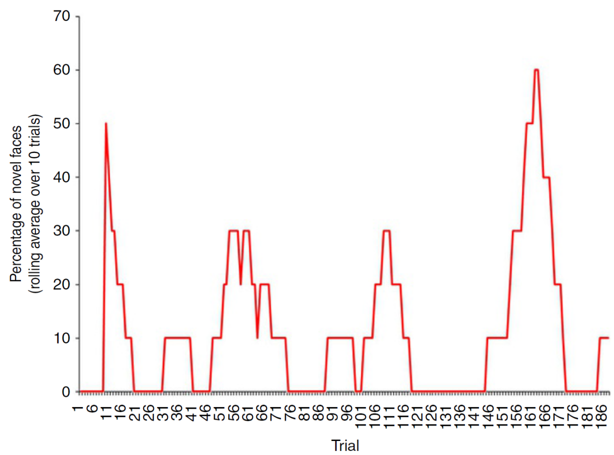
> *'Variability in the contextual familiarity was introduced by varying the average number of previous presentations of faces over a rolling average of five trials. When the rolling average of the number of*

times previous faces had been seen was high, the contextual familiarity could be considered as high and vice versa for low contextual familiarity.'

This information was manually extracted from the publication by HN.

1.1.2 Figure 1 ('c'), bottom-right

The original figure is reproduced below:



The description on pg. 3 regarding figure 1 'b' on the bottom-left side reads:

> *'Varying rate of novel faces over trials displayed as the rolling average of novel faces, regardless of viewpoint over trials. The first ten trials are displayed as zero, to reflect the absence of a rolling average over 10 trials for these first stimuli.'*

Again, this information was manually extracted from the publication by HN.

1.2 Figure 4, suppl. mat.

In the supplementary material, the authors talk about figure 4 depicting

> *'the latter five blocks [where] we controlled the average number of times that an identity had been presented during the experiment within each block (see supplementary fig.4).'*

on pg. 6 and

> *'[as] can be seen in supplementary figure 4 there is considerable variability in the average number of times a face had been seen previously across the experiment.'*

on pg.7. Note, that this figure was neither included in the supplementary material nor in the actual publication.

1.3 Introduction of Blocks

In their original publication, the authors summarize on pg.8

> *'Periodic changes in familiarity were introduced by breaking the trials up into five blocks and controlling the stimulus order in the latter four blocks.'*

but no further information regarding the actual stimulus order was given. In the supplementary material on pg. 6, they elaborate that

> *'Some important elements of the stimulus order were controlled [and that] one of the aims of the experiment was to look at the contextual effects of the recent history of the familiarity of faces, it was therefore important to vary the level of familiarity with faces throughout the experiment, such that there were blocks of trials with higher familiarity and blocks of trials of lower familiarity.'*

Finally they describe that

> *'Such periodic changes in familiarity were introduced by breaking the trials up into six blocks and controlling the stimulus order in the latter five blocks, whilst maintaining randomisation of the viewing angle and of the facial identities. Within each of the latter five blocks we controlled the average number of times that an identity had been presented during the experiment within each block (see supplementary fig.4)'*

Note that figure 4 is entirely missing from all available publicised material. They give an example of block 2, where

> *'we ensured that on average the facial identities presented in that block had been seen five times previously during the experiment'*

Note, that this constraint was added in the form of a prioritized distribution of stimuli that had the lowest deviation from 5 presentations between the 30th and 60th trial i.e. the second block.
They than add another constraint where

> *'Stimulus order was also controlled for by ensuring the rolling average of presentations of an identity across five stimuli did not diverge from the average that was set for that block by more than three.'*

Again, we have no further information due to the missing figure 4 and the specific stimulus order could not be replicated. In the next section, methods for estimating the by the figures implied stimulus order are discussed.

1.4 Introduction of new facial IDs over the experiment

Related to the above, the authors argued for controlling the order of presentations so

> *'that novel facial identities were introduced throughout the experiment, such that participants were not able to perform the task more accurately during the experiment, by increasing the number of "yes" responses in the latter stages of the experiment'*

so that

> *'all of the blocks in the experiment contained novel facial identities that had not been presented before during the experiment and also identities that had been presented only a small number of times during the experiment (i.e., only two or three previous repetitions)'*

1.5 Pseudo-code for data-generation

1.5.1 Introduction of new stimuli over the experiment

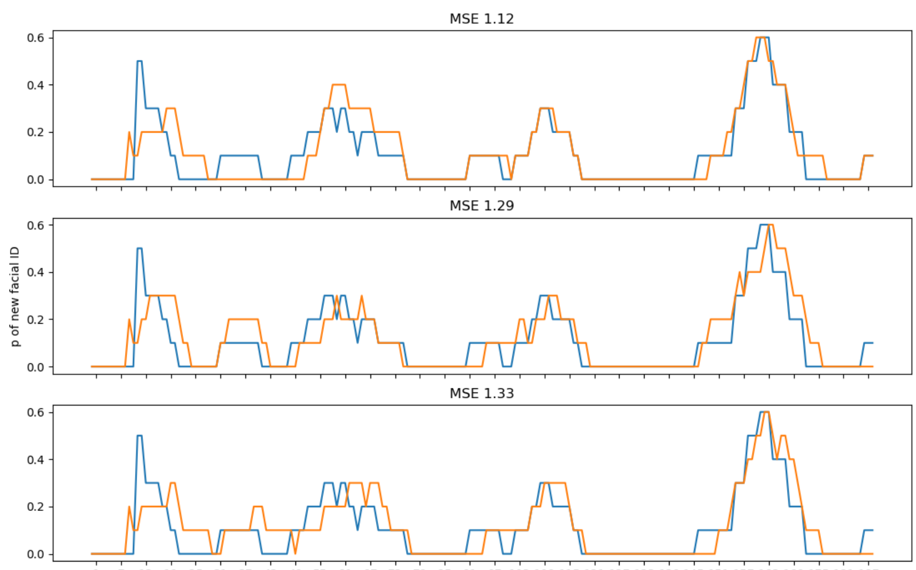
The data from figure 1 'c' was transcribed manually and inserted in the analysis program. The goal here was to iteratively find an array of the same length of the transcribed data that, when an rolling average was calculated, best resembled the original data. The following list summarizes steps, modifications and parameters:

- Transcribed data x_{fig} has the length of 189 for 189 trials.
- Values at index $i = 1, 2$ where set to 0.999, as new faces necessarily need to be presented at the beginning.
- 1000000 iterations were used.
- The estimated array y_{est} hat a length equal to x_{fig}.
- The rolling average window was $w = 10$.
- The sum of y_{est} had to be 24, to insure same amount of stimuli as used by authors.

The following pseudo code summarizes the procedure:

Data: x_{fig} i.e. modified transcribed data
Result: best fitting introduction of new facial IDs given the transcribed data
initialization;
for *n iterations* **do**
  draw random binomial vector y_{est} with p= x_{fig} ;
  if *sum of random binomial vector = 24* **then**
    compute rolling average with window $w = 10$;
    compute MSE between x_{fig} and y_{est} ;
    if *MSE at current iteration < MSE at last iteration* **then**
      store y_{est}
    else
      discard y_{est}
    end
  else
    start next iteration;
  end
end

The following plot illustrates the best-fitting array y_{est} found after 1000000 iterations:



Blue line represents the transcribed data and and yellow the best fitting line from our method. With a mean-squared error of 1.12, the best fitting estimation is shown at the top. Note that this combination was not used in the actual experiment.

1.5.2 Distribution of the stimuli IDs

After finding a suitable array y_{est} sufficiently similar to the transcribed data, the facial IDs needed to be distributed over the course of the trials. Again, the transcribed data of figure 1 'b' was used to calculate a MSE metric to ensure sufficient similarity to the transcribed data. The following summarizes the procedure:

- The stimulus-pool contained a total of 24 IDs.
- 15 IDs were presented 12 times each over the course of the experiment(15x12 stimuli).
- 9 IDs were presented once each over the course of the experiment (9x1 stimuli).
- To ensure sufficient stimuli in a stimulus pool, 9x1 stimuli were more likely to appear at the end of the experiment.
- As in figure 1 'b', the rolling average had a window of $w = 10$.
- In the second Block (i.e. trial 30 to 60) the IDs were prioritized that have been seen 5 times previously during the experiment.

Note that in this approach a purely numerical/brute-force method was used requiring substantial amount of iteration. The following code illustrates the procedure:

Data: generated data from function above, modified transcribed data
Result: best distribution of stimuli over the course of the experiment given generated list
initialization;
for *n iterations* **do**
  initialize dictionary with constraints mentioned above;
  try *iterate through list***:**
    compute rolling average with window $w = 10$;
    compute MSE between x_{fig} and y_{est} ;
    if *MSE at current iteration ¡ MSE at last iteration* **then**
      store y_{est}
    else
      discard y_{est}
    end
  start next iteration;
  catch *didnt make it to the end***:**
    break
  end
end

2 Component description

2.1 Response model

The predictions by each model were converted to probabilities by the 'Luce Choice Rule' (aka. softmax, see [1]) by the formula

$$p(yes|t) = \frac{1}{1 + (e^{-\beta * F_t})} \tag{1}$$

i.e. the probability of answering yes is the softmax function of the total familiarity at a given time t multiplied by the individually estimated parameter β which reflects the

> *'...stochasticity of the participants choices on the task, and therefore their sensitivity to the value of F(t).'*

2.2 Contextual familiarity

Apps and Tzakiris (2013) pg.8

> *'...importance of top-down context-dependent expectations about likely sensory input as modulators of behavioural responses. We [...] assumed that facial recognition will be dependent on the moderating effects of contextual information.'*

> *'...important elements of the stimulus order were controlled in order to examine, in line with one of the aims of the experiment, the contextual effect of the familiarity of previously perceived faces.'*

2.3 View (in)-dependent familiarity

> *'We therefore make the assumption that the subject is calculating the difference between the maximum familiarity of a face and its actual familiarity, before the maximum level of familiarity with that face is known ... [and] we therefore make the assumption that the subject is calculating the difference between the maximum familiarity of a face and its actual familiarity, before the maximum level of familiarity with that face is known.'*

We therefore make the assumption that the subject is calculating the difference between the maximum familiarity of a face and its actual familiarity, before the maximum level of familiarity with that face is known.

suggest that this maximum familiarity is developed ontogenetically and reflects a tuning property of neurons involved in the perceptual learning of facial identities

[1] 'Luce's choice axiom - Scholarpedia'

3 Model description

The authors described:

a total of 24 stimuli so $i \in 1, .., 24$. From those 24, *15 where presented multiple times* and *9 where presented only once*

3.1 view-independent x context aka 'winning model'

Here the authors proposed that the **overall familiarity at a given trial** F_t is a function of the view-independent familiarity of stimulus i presented at time t i.e. V_i and the current contextual familiarity.

$$F_t = V_i \cdot C_t \tag{2}$$

3.1.1 view-Independence

The view-independent familiarity is updated on each trial where the specific facial identity is presented via

$$V_{i,n+1} = V_{i,n} + \alpha\delta \tag{3}$$

i.e. the view-independent familiarity with a given stimulus after having observed the stimulus. Here, δ is defined as the difference between maximum possible familiarity of any facial identity (defined as λ, model-parameter) and the view-independent familiarity before a given stimulus was observed thus giving

$$\delta = \lambda - V_{in} \tag{4}$$

or as the authors phrase it on pg. 8

> *'...discrepancy is calculated between the probability of the occurrence of a sensory event and the actual sensory event'.*

Further, α is defined as an 'idiosyncratic' parameter between 0 and 1 which

> *'...scaled the rate at which familiarity was acquired with the facial identities [by each subject]'.*

Finally, the authors proposed setting

> *'...the initial value of V for all of the identities [...] idiosyncratically for each participant [...] as the percentage of false-positive responses on the 24 stimuli that were the first presentations of each face, multiplied by λ, the maximum familiarity parameter'.*

which gives

$$V_{i,0} = rate_{FP} \cdot \lambda \tag{5}$$

3.1.2 context dependence

The original publication proposes that the contextual familiarity is updated at every trial, regardless of facial identity through

$$C_{t+1} = C_t + \sigma\varepsilon \tag{6}$$

where C_t is the contextual familiarity before observing a stimulus and ε is defined as a 'prediction error'

$$\varepsilon = C_t - V_{i,t} \tag{7}$$

i.e. the difference between current contextual familiarity and the unupdated view-independent familiarity from the stimulus shown at time t. The parameter σ functions as a idiosyncratic learning-rate that reflects

> *'...how sensitive the participant is to trial-by-trial changes in the familiarity of facial stimuli.'.*

Note, that when implementing the original model, inclusion of the term described in equation 6 returned *severe errors*. More specifically, when the weighted contextual-prediction error was *added* (cf. equation 6) to the previous contextual familiarity, it rapidly produced large negative values causing the model-estimation to crash. Interestingly, exchanging addition with *subtraction*, the model produced plausible results and remained numerically stable. From here on, the correct formula for updating contextual familiarity is assumed to be:

$$C_{t+1} = C_t - \sigma\varepsilon \tag{8}$$

where, as defined above, C_t is the contextual familiarity before observing a stimulus and ε is defined as a contextual 'prediction error'

$$\varepsilon = C_t - V_{i,t} \tag{9}$$

3.1.3 view-independent x context parameter summary

In this model ('view-dependence x context') the following parameters had to be estimated per-person. The the parameter space was defined by the authors as categorical. Note that the formulation of the authors (pg. 9) 'between value x and y' i.e.

> *'...varied the view-independent learning rate parameter (a) between 0 and 1 in steps of 0.01, the context-dependent learning rate (d) between 0 and 1 in steps of 0.01, the maximum familiarity (l) between 0 and 2 in steps of 0.01 and we varied the stochasticity parameter (b) between 0.1 and 20 in steps of 0.1.'.*

is understood as an open interval *excluding the endpoints*. Thus it follows that

- 'stochasticity parameter' $\beta \in (0.1, 20) = \{x \in \mathbb{R} | a < x < b | \Delta\beta = \pm 0.1\}$
- 'maximum familiarity' $\lambda \in (0, 2) = \{x \in \mathbb{R} | a < x < b | \Delta\lambda = \pm 0.01\}$

- ’view-independent learning rate’ $\alpha \in (0,1) = \{x \in \mathbb{R} | a < x < b | \Delta\alpha = \pm 0.01\}$
- ’context learning rate’ $\sigma \in (0,1) = \{x \in \mathbb{R} | a < x < b | \Delta\sigma = \pm 0.01\}$

Note that in the parameter definition on page 9, the authors used the Greek letter delta δ but on pg.8 they used the letter σ. As this is likely an *error*, from here on further the ’context learning rate’ is denoted as σ.

3.2 View-dependent

Control model to identify if familiarity acquisition is view-dependent. Total familiarity is given by

$$F_t = Vd_i \tag{10}$$

and is updated for a specific facial identity upon its presentations by

$$Vd_{i,n+1} = Vd_{i,n+1} + \gamma * \tau \tag{11}$$

where γ is the view-dependent learning rate (model-parameter) and

$$\tau = \lambda - Vd_{i,n} \tag{12}$$

is defined as the difference between the maximum view-dependent familiarity λ (model-parameter) and the un-updated view-dependent familiarity $Vd_{i,n}$ of the current stimulus i with n prior presentations.

3.2.1 Parameters

Note, that the authors did not provide information if the sample space of the two model-parameters λ, γ was different, therefore is was assumed by us that (similar to the winning model)

- ’stochasticity parameter’ $\beta \in (0.1, 20) = \{x \in \mathbb{R} | a < x < b | \Delta\beta = \pm 0.1\}$
- view-dependent learning rate $\gamma \in (0,1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma = \pm 0.01\}$
- maximum view-dependent familiarity $\lambda \in (0,2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda = \pm 0.01\}$

3.3 View-dependent x context

View-dependent familiarity aquisition modulated by contextual expectations therefore

$$F_t = Vd_i \times C_t \tag{13}$$

Note that C_t is updated the same way as in formulas 6 and 7 and view-dependent familiarity was updated by formulas 9 and 10.

3.3.1 Parameters

Again, as no additional info was provided by the authors regarding the sample-space the following was assumed:

- 'stochasticity parameter' $\beta \in (0.1, 20) = \{x \in \mathbb{R} | a < x < b | \Delta\beta = \pm 0.1\}$
- view-dependent learning rate $\gamma \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma = \pm 0.01\}$
- maximum view-dependent familiarity $\lambda \in (0, 2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda = \pm 0.01\}$
- 'context learning rate' $\sigma \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\sigma = \pm 0.01\}$

3.4 View-independent

Here, it is assumed that total familiarity at a given trial is a function only of view-independent familiarity and context does not play a role. Thus

$$F_t = V_i \tag{14}$$

and V_i is updated through equations (3) and (4). Initialization values are assumed the same as in equation (5).

3.4.1 Parameters

Again, as no additional info was provided by the authors regarding the sample-space the following was assumed:

- 'stochasticity parameter' $\beta \in (0.1, 20) = \{x \in \mathbb{R} | a < x < b | \Delta\beta = \pm 0.1\}$
- view-dependent learning rate $\gamma \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma = \pm 0.01\}$
- maximum view-dependent familiarity $\lambda \in (0, 2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda = \pm 0.01\}$

3.5 View-independent x view-dependent

Here, it is assumed that total familiarity at time t is an additive of the view-independent and view-dependent familiarity therefore

$$F_t = V_i + Vd_i \tag{15}$$

where V_i is updated through equations (3) and (4) and Vd_i is updated by (9) and (10). Again, initialization values are assumed to be (5).

3.5.1 Parameters

Again, as no additional info was provided by the authors regarding the sample-space the following was assumed:

- 'stochasticity parameter' $\beta \in (0.1, 20) = \{x \in \mathbb{R} | a < x < b | \Delta\beta = \pm 0.1\}$
- view-dependent learning rate $\gamma_d \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma_d = \pm 0.01\}$
- maximum view-dependent familiarity $\lambda_d \in (0, 2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda_d = \pm 0.01\}$
- view-independent learning rate $\gamma \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma = \pm 0.01\}$
- maximum view-independent familiarity $\lambda \in (0, 2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda = \pm 0.01\}$

3.6 View-dependent x view-independent x context-dependent

In this control model, total familiarity at time t is a function of view-dependent, view-independent and contextual factors. Note that

$$F_t = (V_i + Vd_i) \times C_t \tag{16}$$

and V_i is updated through equations (3) and (4), Vd_i is updated by (9) and (10) and C_t is updated by equations (6) and (7). Note that equation (7) slightly modified takes the form

$$\epsilon = C_t - (V_i + Vd_i) \tag{17}$$

3.6.1 Parameters

Again, as no additional info was provided by the authors regarding the sample-space the following was assumed:

- 'stochasticity parameter' $\beta \in (0.1, 20) = \{x \in \mathbb{R} | a < x < b | \Delta\beta = \pm 0.1\}$
- view-dependent learning rate $\gamma_d \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma_d = \pm 0.01\}$
- maximum view-dependent familiarity $\lambda_d \in (0, 2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda_d = \pm 0.01\}$
- view-independent learning rate $\gamma \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\gamma = \pm 0.01\}$
- maximum view-independent familiarity $\lambda \in (0, 2) = \{x \in \mathbb{R} | 0 < x < 2 | \Delta\lambda = \pm 0.01\}$
- 'context learning rate' $\sigma \in (0, 1) = \{x \in \mathbb{R} | a < x < b | \Delta\sigma = \pm 0.01\}$

4 Computation

4.1 Parameters

Original: constrained and categorical space (space definition unclear) Replication: constrained and categorical space, unconstrained?

4.2 Optimization

4.3 Author implementation

No info:

- Optimization
- Implementation

Maximum likelihood-estimation is usually only available via numerical iterative procedures and closed form equations are rarely used. The authors only report using a maximum likelihood estimation without any information about the implementation of the code (e.g. Python, R, Matlab) nor the specific optimization routine (e.g. gradient-descent, random-sampling etc.). Here, θ_{ML} is a n-dimensional vector sampled from the n-dimensional sample-space defined per model. Finally the maximum-likelihood parameter estimate θ_{ML} is chosen by

$$\theta_{ML} = arg\,max\,\theta_{ML} \tag{18}$$

and the maximum-likelihood estimate is given by

$$\theta_{ML} \;= \sum_{t}^{T} log(\,p(answer|t)) \tag{19}$$

4.4 Replication implementation

Random Sampling, Bayesian Optimization, (Gradient-descent)
Original: Random Sampling (no specific input)
Replication:

4.5 Model selection

Original: ML, delta neg. log-likelihood with t-test comparison vs. 0
Replication: delta neg. log-likelihood with t-test comparison, Information Criterion, Bayes Factor?, LOOCV?

5 Power analysis

5.1 Computational Modeling

Here, the view-independent x context is considered the 'winning'-model H_1 and the set of the control-models $H0 = H_{01}, ..., H_{0i}$ is considered the null-hypothesis. Note that model evidence is the sum of log-likelihood of the models. Derivation of BF_{10} is partially taken from [2] We start with Bayes-rule and for a simple model selection between two models H_1, H_0 where $i = 1, 0$ we get

$$p(model_i|data) = \frac{p(data|model_i) * p(model_i)}{p(data)} \tag{20}$$

and

$$p(data) = p(data|model_{H1}) * p(model_{H1}) + p(data|model_{H1}) * p(model_{H1}) \tag{21}$$

It is also defined that

$$p(model_{H0}) = 1 - p(model_{H1}) \tag{22}$$

an therefore the posteriors

$$p(model_{H0}|data), p(model_{H1}|data) \tag{23}$$

also follow

$$p(model_{H0}|data) = 1 - p(model_{H1}|data) \tag{24}$$

so we can restate the rule as

$$\frac{p(model_{H1}|data)}{p(model_{H0}|data)} = \frac{p(data|model_{H1}) * p(model_{H1})}{p(data|model_{H0}) * p(model_{H0})} \tag{25}$$

and since we have a uniform/uninformative prior over all models where $p(model) = \frac{1}{n_{models}}$ which cancels out so we get

$$BF_{10} = \frac{p(data|model_{H1})}{p(data|model_{H0})}. \tag{26}$$

With more than one control model we get

$$\begin{aligned} p(model_{H1}|data) &= \frac{p(data|model_{H1})}{\sum_{i=1}^{i=I} p(data|model_{H0i})} & (27) \\ \Leftrightarrow \log\left(p(model_{H1}|data)\right) &= \log\left(\frac{p(data|model_{H1})}{\sum_{i=1}^{i=I} p(data|model_{H0i})}\right) \\ \Leftrightarrow \log\left(p(model_{H1}|data)\right) &= \log\left(p(data|model_{H1})\right) - \left(\sum_{i=1}^{i=I} p(data|model_{H0i})\right) & (28) \end{aligned}$$

[2] http://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf

which is transformed to [HOW? WHY?]

$$\log(p(model_{H1}|data)) = \log(p(data|model_{H1})) - log(\sum_{i=1}^{I} exp(p(data|H_{0i}))) \tag{29}$$

5.2 Neuroimaging

Andreas and Nestor Power analysis??