Probabilistic modeling of mouse-tracking data

A state-space approach

[1]Antonio Calcagnì, [2]Luigi Lombardi, [2]Marco D'Alessandro

[1] Dep. of Developmental and Social Psychology, University of Padova, IT
[2] Dep. of Psychology and Cognitive Science, University of Trento, IT

EMPG Meeting 2018
Genova, July 31 - August 2

UNIVERSITÀ DEGLI STUDI DI PADOVA

Mouse-tracking methodology

Over last decades, mouse-tracking is becoming a popular approach to collect real-time cognitive measures.

Mouse-tracking allows time-based recording of x-y computer mouse positions during experimental tasks (e.g., lexical decisions, categorization)

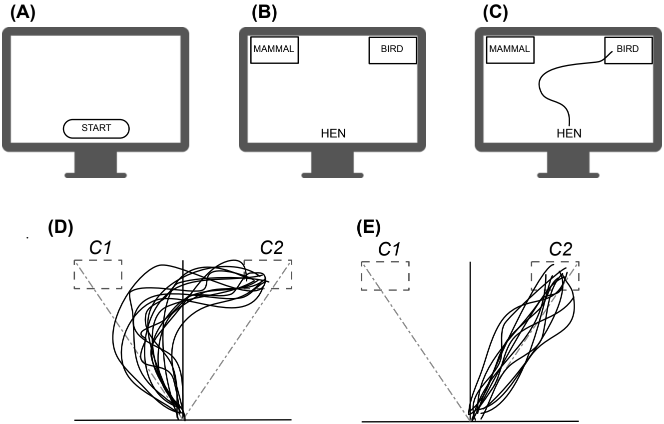
Main idea: x-y trajectories can unfold the decision process underlying hand movement

Example: in a two-choice categorization task (*hen: mammal vs. bird*), stimuli with higher ambiguity (*hen*) push the hand movement toward the incorrect response (*mammal*)

Mouse-tracking methodology

Introduction

Università degli Studi di Padova



Several methods are available to analyse mouse-tracking trajectories:

- **model-free methods:** spatial measures (e.g., MD, AUC, x/y flips. Hehman et al., 2014), raw temporal measures (e.g., initiation-time, velocity/acceleration profiles. Kieslich & Henninger, 2017)
- **model-based methods:** decision landscapes (Zgonnikov et al., 2017), latent gaussian processes (Cox et al., 2012), entropy-based decompositions (Calcagnì et al., 2017)

Mouse-tracking data analysis usually proceeds with a **two-stage process**:

- first computes summary measures (e.g., AUC, MD, entropies)
- then applies statistical model on the summary measures

Mouse-tracking data analysis

Standard approaches

Università degli Studi di Padova

Mouse-tracking data analysis usually proceeds with a **two-stage process**:

- first computes summary measures (e.g., AUC, MD, entropies)
- then applies statistical model on the summary measures

Both the model-free and model-based approaches lack a unified way to **simultaneously model** and **analyse** mouse-tracking data.

Mouse-tracking data analysis

Why a unified approach?

A **unified approach** is needed:

- Because summary measures can neglect **movement variability** of the x-y trajectories (e.g., dissimilar trajectories treated as similar)
- To separate **experimental variability** (task manipulation) from **individual variability** (hand motor programs)
- Because post-hoc analyses ignore the **data generation process** of the observed x-y trajectories

Our proposal

A dynamic probabilistic model

We use a **state-space modeling** (SSMs) approach to represent the transformed observed trajectories data $\boldsymbol{y}$ (in terms of angles in $[0, \pi/2]$) as a function of:

- latent individual movement profiles ($\boldsymbol{z}$)
- experimental manipulations (β)

Our proposal

A dynamic probabilistic model

Data

$\mathbf{Y}_{I \times J \times N}$ obs. movement angles

$i = 1, \ldots, I$ individuals
$j = 1, \ldots, J$ stimuli/trials
$n = 1, \ldots, N$ time-step of recording
y_{ijn}: movement angle

Our proposal

A dynamic probabilistic model

UNIVERSITÀ DEGLI STUDI DI PADOVA

Model representation

$$y_{ijn} \sim \text{mVM}(\mu_D, \mu_T, \kappa_D, \kappa_T, \pi_{ijn})$$

Von-Mises Observation equation:
$\{\mu_D, \mu_T\}$: locations of DISTRACTOR/TARGET
$\{\kappa_D, \kappa_T\}$: variability of DISTRACTOR/TARGET
π_{ijn}: probability to select DISTRACTOR

$$\pi_{ijn} = \text{logit}^{-1}(z_{in}, \beta_j)$$

Attraction probability equation:
z_{in}: latent individual movement dynamics
β_j: information of experimental trials

$$z_{in} = \mathcal{N}(z_{i,n-1}, \sigma_i^2)$$

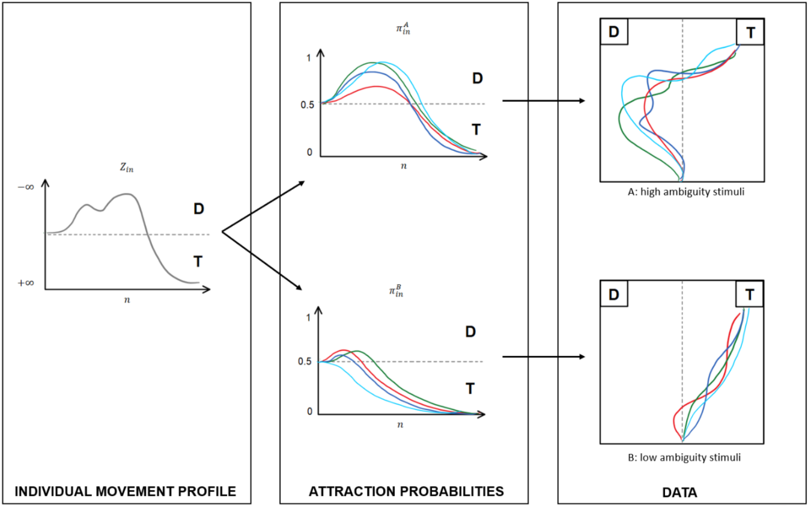
Individual movement equation:
$z_{i,n-1}$: lag-1 AR process
σ_i^2: motor variability

$$\beta_j = \sum_k d_{jk}\gamma_k + x_j(\eta + \sum_k d_{jk}\delta_k)$$

Stimuli equation:
d_{jk}: design matrix
x_j: stimuli covariate
$\{\gamma_k, \delta_k, \eta\}$: design parameters

Our proposal

A dynamic probabilistic model



Model identification

Model identification requires estimating:

- latent states $\mathbf{Z} = (\mathbf{z}_{1,1:N}, \ldots, \mathbf{z}_{I,1:N})$
- stimuli parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$

via the following decomposition:

$$\underbrace{\log f(\mathbf{Z}, \Theta|\mathbf{Y})}_{\text{joint posterior density}} = \underbrace{\log f(\Theta|\mathbf{Y})}_{\text{likelihood}} + \underbrace{\log f(\mathbf{Z}|\mathbf{Y})}_{\text{filtered density}} + \underbrace{\log f(\Theta)}_{\text{prior density}}$$

Our proposal

A dynamic probabilistic model

Model identification

$$\underbrace{\log f(\mathbf{Z}, \Theta|\mathbf{Y})}_{\text{joint posterior density}} = \underbrace{\log f(\Theta|\mathbf{Y})}_{\text{likelihood}} + \underbrace{\log f(\mathbf{Z}|\mathbf{Y})}_{\text{filtered density}} + \underbrace{\log f(\Theta)}_{\text{prior density}}$$

Samples from $f(\mathbf{Z}, \Theta|\mathbf{Y})$ are obtained with a marginal-Metropolis Hastings which alternates between updating $f(\mathbf{Z}|\mathbf{Y})$ and evaluating $f(\Theta|\mathbf{Y})$.

Our proposal

A dynamic probabilistic model

Università degli Studi di Padova

Model identification

$$\underbrace{\log f(\mathbf{Z}, \Theta|\mathbf{Y})}_{\text{joint posterior density}} = \underbrace{\log f(\Theta|\mathbf{Y})}_{\text{likelihood}} + \boxed{\underbrace{\log f(\mathbf{Z}|\mathbf{Y})}_{\text{filtered density}}} + \underbrace{\log f(\Theta)}_{\text{prior density}}$$

$f(\mathbf{Z}|\mathbf{Y})$ is recursively computed over $n = 0, \ldots, N$ via **gaussian approximation filter** on:

$$\log f(\mathbf{Z}|\mathbf{Y}) \propto \log \underbrace{f(y_{ijn}|z_{in}, \theta)}_{\text{observation density}} + \log \underbrace{\int_{\mathbb{R}} f(z_{i,n}|z_{i,n-1}, \theta) f(z_{i,n-1}|y_{ij,0:n-1}, \theta)\, dz_{i,n-1}}_{\text{Chapman-Kolmogorov eq.}}$$

for each $i = 1, \ldots, N$.

Model identification

$$\underbrace{\log f(\mathbf{Z}, \Theta|\mathbf{Y})}_{\text{joint posterior density}} = \boxed{\underbrace{\log f(\Theta|\mathbf{Y})}_{\text{likelihood}}} + \underbrace{\log f(\mathbf{Z}|\mathbf{Y})}_{\text{filtered density}} + \underbrace{\log f(\Theta)}_{\text{prior density}}$$

$f(\Theta|\mathbf{Y})$ is recursively computed via integration over $n = 0, \ldots, N$:

$$\sum_i \sum_j \left(\sum_n \log \int_{\mathbb{R}} f(y_{ijn}|z_{in}, \theta) f(z_{in}|\mathbf{y}_{ij,0:n-1}, \theta) dz_{in} \right)$$

for each $i = 1, \ldots, N$.

Our proposal

A dynamic probabilistic model

Data analysis

Data analysis is performed using **marginal posterior distributions** of:

- filtered states $f(\mathbf{Z}|\mathbf{Y})$
- stimuli parameters $f(\beta|\mathbf{Y})$

Exemplary application

A lexical decision task (Barca et al., 2012)

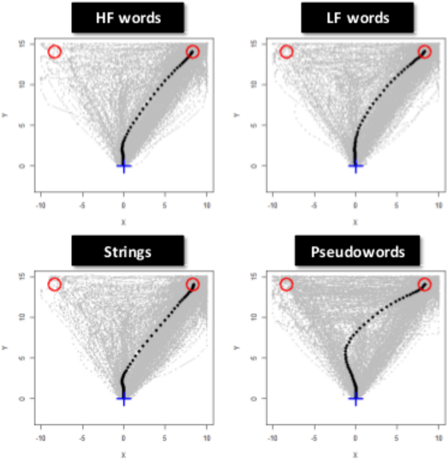
Lexical decision task (in Italian) where *stimuli type* was varied with four levels:

- **HF - High frequency words** (e.g., "acqua", water)
- **LF - Low frequency words** (e.g., "cervo", deer)
- **PW - Pseudowords** (e.g., "dorto")
- **ST - Strings** (e.g., "btfpr")

Participants ($n = 22$, age 20-35, right-handed) saw a total of 96 stimuli to be categorized as **word** or **non-word**.

Exemplary application

A lexical decision task (Barca et al., 2012)



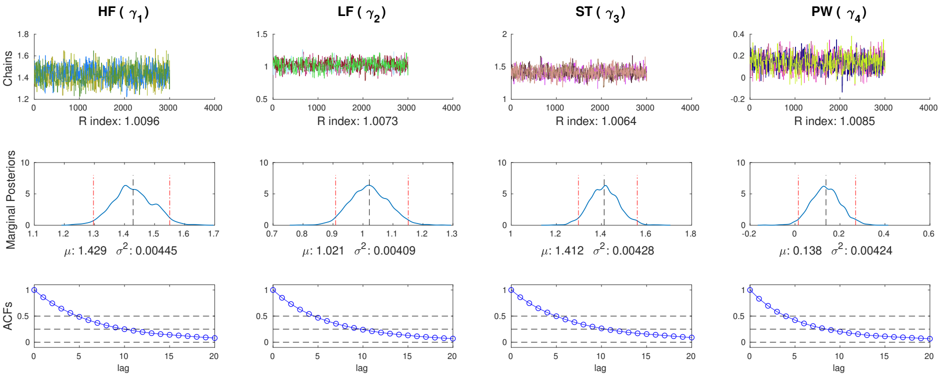
Goals of the SSM based analysis:

- estimate and evaluate how experimental manipulations affect individual responses:

  $\boldsymbol{\beta} = \text{HF}\ \gamma_1 + \text{LF}\ \gamma_2 + \text{ST}\ \gamma_3 + \text{PW}\ \gamma_4$ (constrast equation)

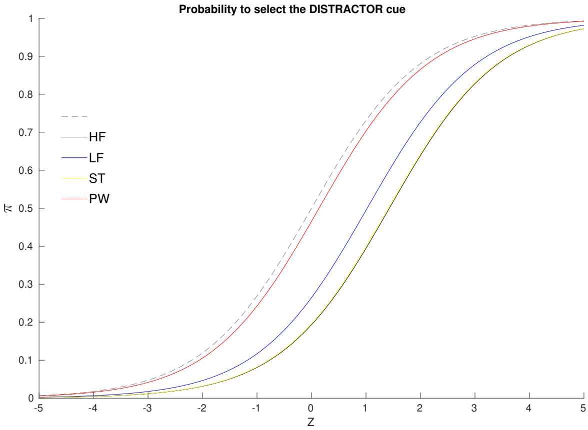
- estimate individual movement profiles **Z** and $\boldsymbol{\pi}$

4 chains x 4000 iterations = 12000 samples



Exemplary application

Effect of stimuli type



p(distractor|HF) = 0.186

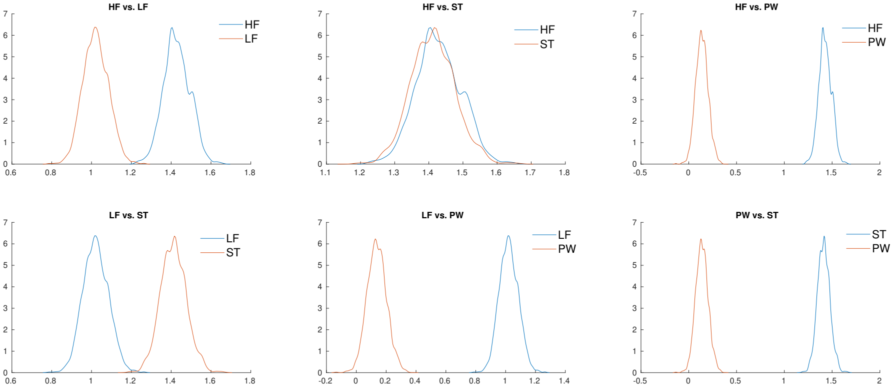
p(distractor|LF) = 0.258

p(distractor|ST) = 0.189

p(distractor|PW) = 0.458

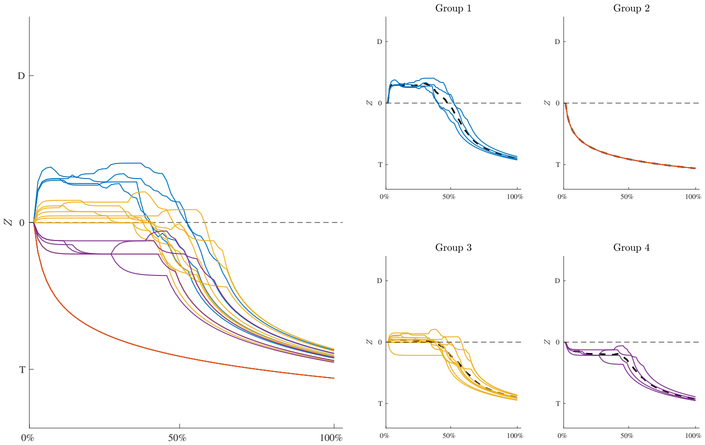
Exemplary application

Posteriors analysis



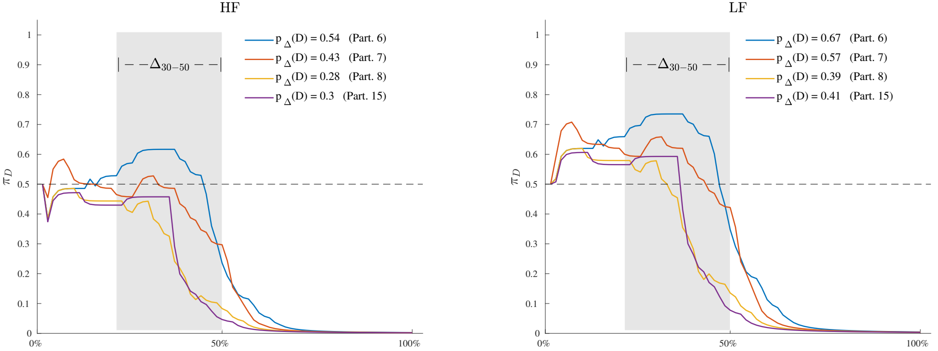
Exemplary application

Individual profiles **Z**



Exemplary application

Attraction probability profiles π (group 1)



Final remarks

A dynamic probabilistic model

- Our proposal offers a unified framework for modeling and analysing mouse-tracking trajectories
- Individual's movement heterogeneity is included as AR(1) stochastic process
- Experimental manipulations are included as linear combination of categorical and continuous variables
- Group-level and individual-level analyses can be performed by assessing marginal posterior distributions of parameters
- Movement profiles can be further analysed in terms of proximities (e.g., clustering)