Bounded rationality in structured density estimation

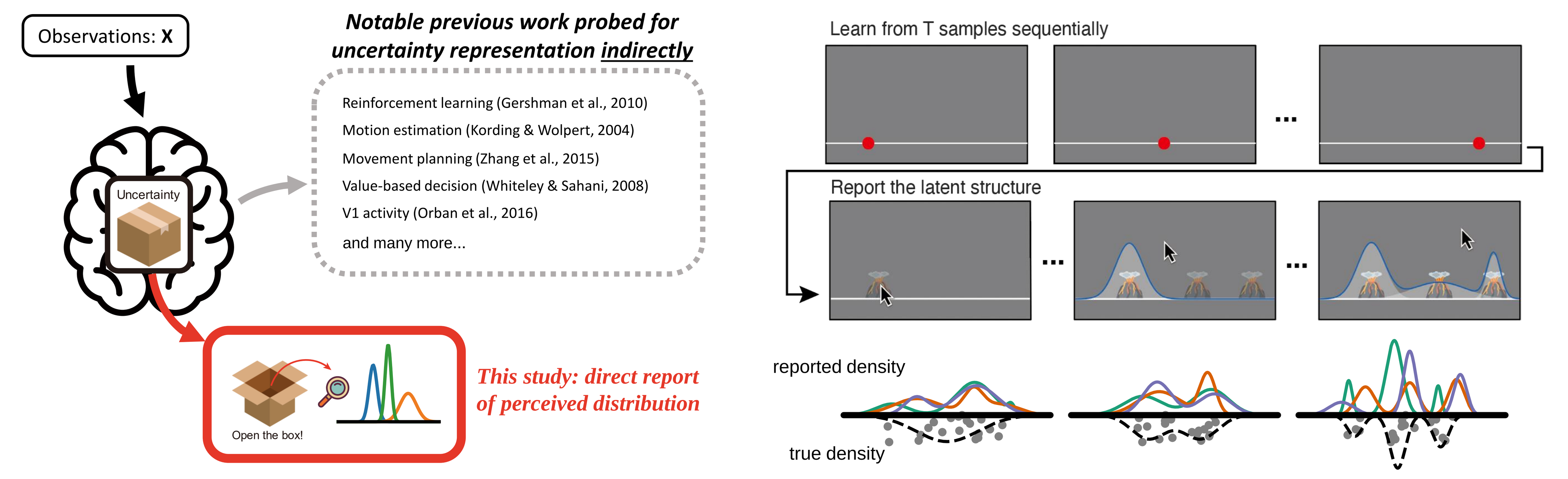
Tianyuan Teng[#,1], Li Kevin Wenliang[#,✉,2], Hang Zhang[✉,1,3,4]

# Equal contributions. ✉ Co-corresponding authors. 1. Peking-Tsinghua Center for Life Sciences, Peking University, Beijing. 2. Gatsby Computational Neuroscience Unit, University College London, London. 3. School of Psychological and Cognitive Sciences and Key Laboratory of Behavior and Mental Health, Peking University, Beijing. 4. PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing

Highlights

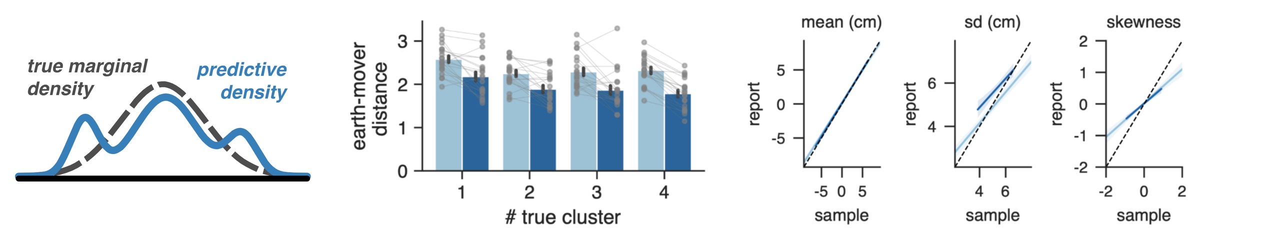
- Humans observe samples from an unknown complex distribution and report their estimated distributions through parameters of Gaussian mixture models (GMMs).
- Surprisingly, humans tend to **converge towards a wrong internal construct**, even with more data.
- We propose the Density Estimation Framework (DEF) to capture the complex behavior in this task.
- We can explain participants' representation of uncertainty by an **economical nonparametric GMM** under memory constraint and additional cognitive biases.

1. Introduction and experiment design

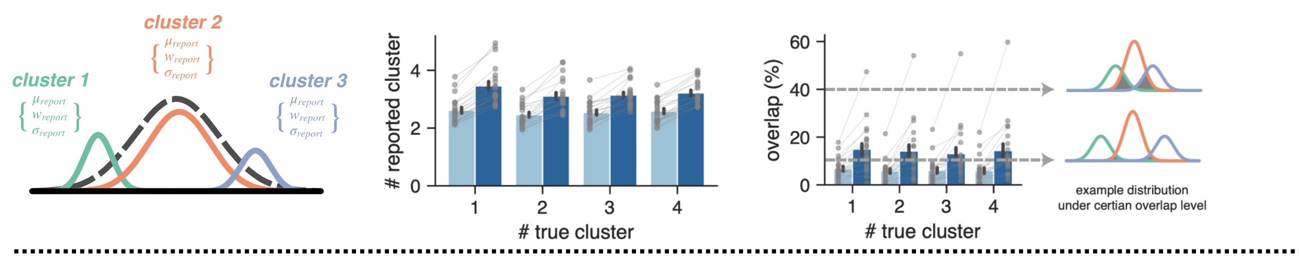


2. Main behavioral results: participants can't learn the true number of clusters

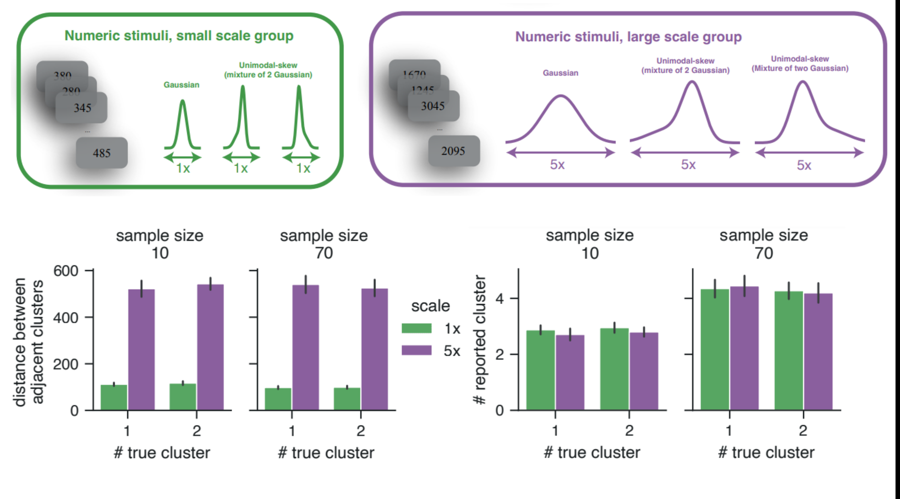
Improved representation of the **predictive density**



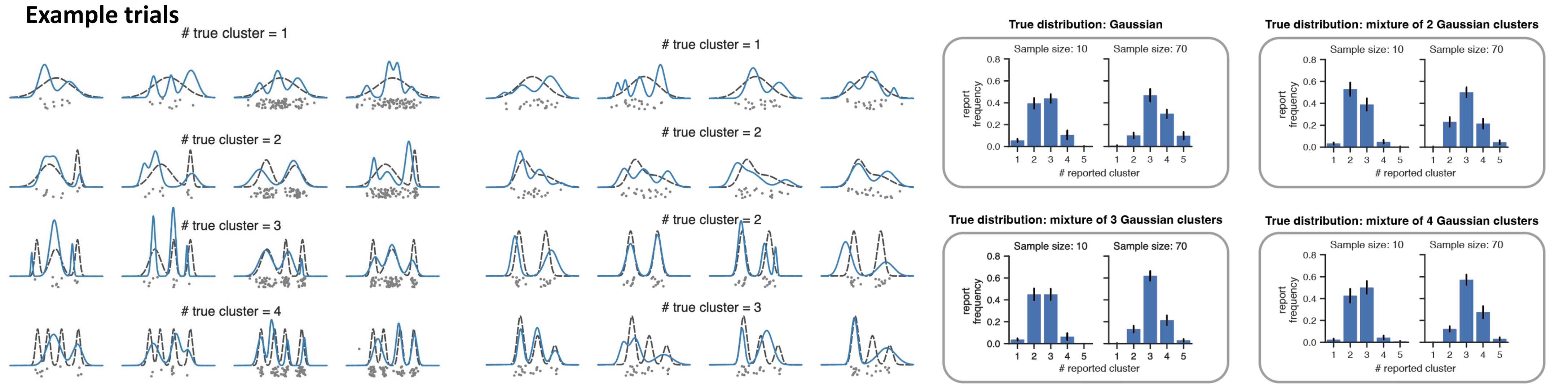
Inconsistency of the inferred structure



Scale- and modality-invariance



Example trials



3. Modeling: Density Estimation Framework (DEF)

Problem formulation

- Interpretable model p
- Likelihood fitting and critique

$$\mathcal{L}(\theta) := \mathbb{E}_{\text{data}}[\log p_\theta(\varphi^r \mid x_{1:T})]$$

$$\theta^* = \max_\theta \mathcal{L}(\theta)$$

$$\text{AIC} = -2\mathcal{L}(\theta^*) + \frac{2}{m}\log|\theta^*|$$

Key notations
$x_t \in \mathbb{R}$: sample at time t
$z_t \in \mathbb{N}_+$: the cluster identity for sample
$n_k = \sum_\tau \mathbb{I}[z_\tau = k]$: # samples in cluster k
K: reported number of clusters
φ: sufficient statistics of all clusters
φ^r, φ^i: reported, inferred

Rational model: non-param GMM [Anderson, 1991, Sanborn et al., 2010]

- Cluster means and variances specified by sufficient statistics φ
- A **rational** prior for generating **internal constructs:**
  cluster prior $p(z_{t+1}|\varphi_t)$ a likelihood $p(x_{t+1}|\varphi_t, z_{t+1})$
- Particle $z_{t+1} \sim p(z_{t+1}|x_{t+1}, \varphi_t) \propto p(z_{t+1}|\varphi_t)p(x_{t+1}|z_{t+1}, \varphi_t)$
- Key parameters: prior sd σ_0 and mean μ_0 with "confidence" a_0, λ_0
- But, it cannot capture reported K
- It has *infinite* capacity, while human memory is *finite*
- Performs *suboptimal* inference on an *optimal* model

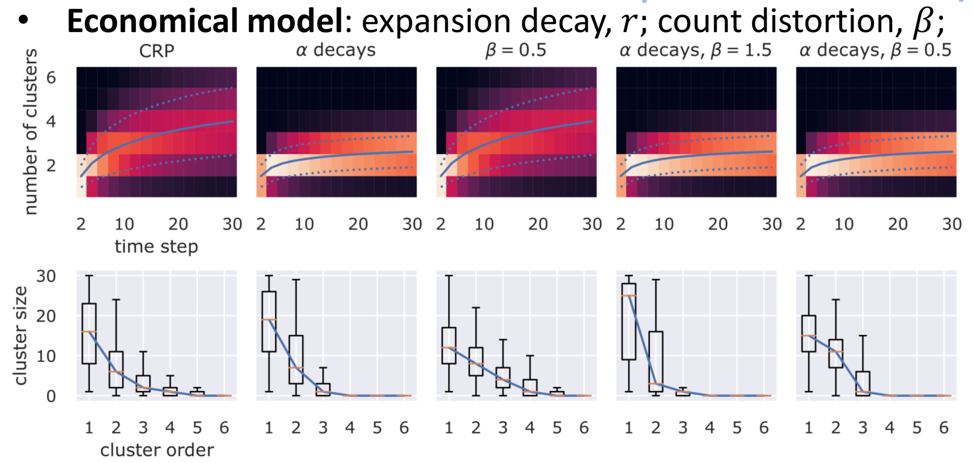
Modeling contributions:

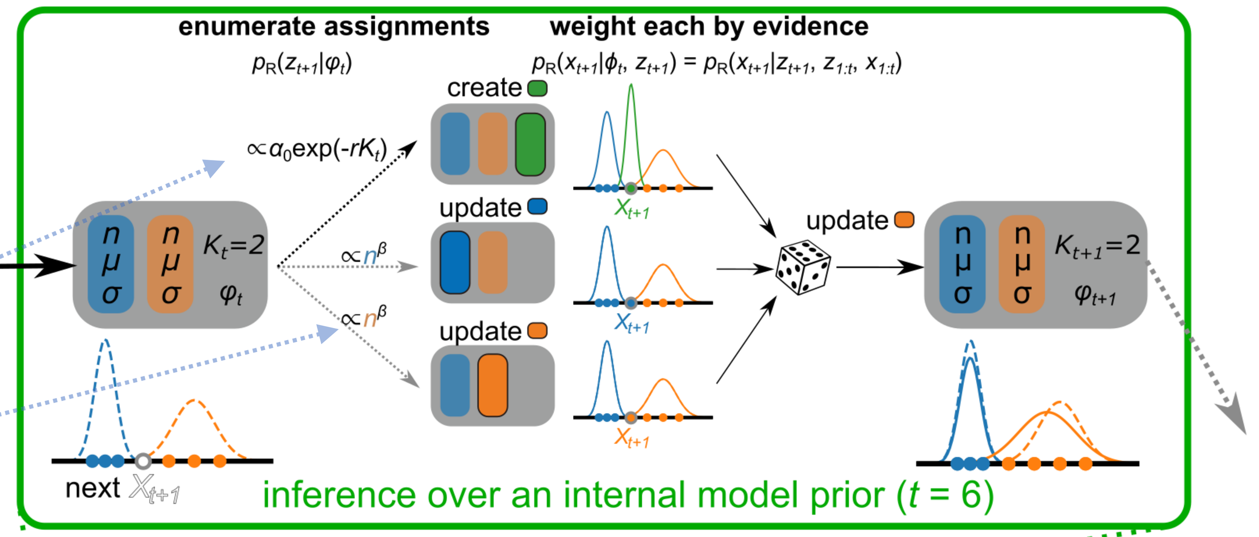
1. **The density estimation framework (DEF)**
2. **An economical internal construct prior**

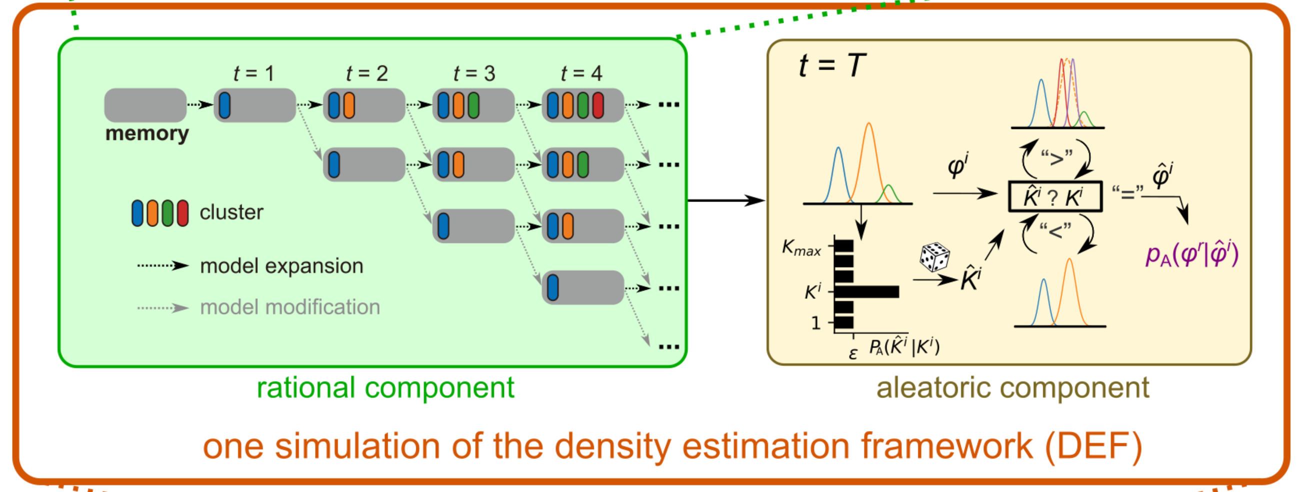
The DEF factorizes the likelihood as

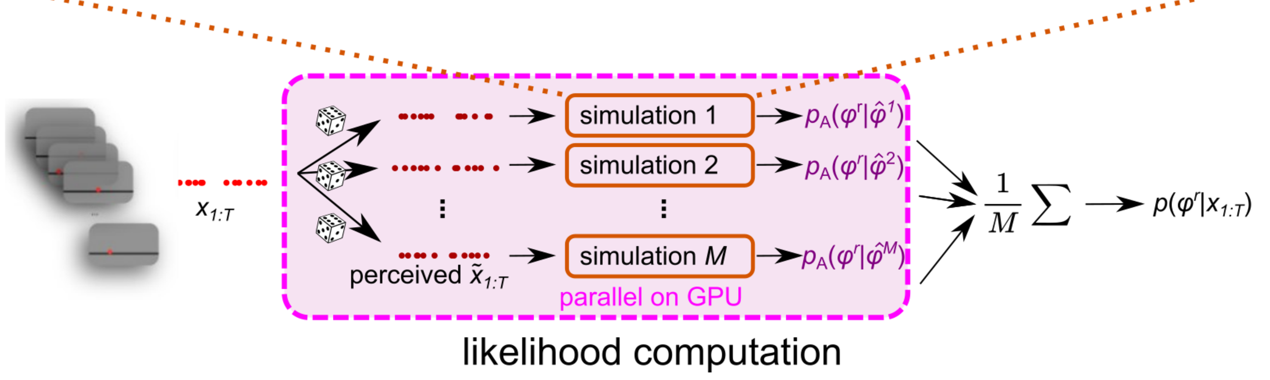
$$p_\theta(\varphi^r|x_{1:T}) = \int p_A(\varphi^r|\varphi^i)\, \mathrm{d}p_R(\varphi^i|x_{1:T})$$

- **Rational component** $p_R(\varphi^i|x_{1:T})$, e.g.:
  - Posterior of the rational model
  - **Economical model**: expansion decay, r; count distortion, β;









- **Aleatoric component** $p_A(\varphi^r|\varphi^i)$: **structured noise model**
  - Leads to overall likelihood
    $$p(\varphi^r \mid x_{1:T}) = \int p_A(\varphi^r|\varphi^i)\mathrm{d}p_R(\varphi^i|x_{1:T})$$
  - Noise structure: gives well defined likelihood
    $$p_A(\varphi^r|\varphi^i) = \int p_A(\varphi^r|\hat{\varphi}^i)\mathrm{d}p_A(\hat{\varphi}^i|\varphi^i)$$
  - $p_A(\hat{\varphi}^i|\varphi^i)$: **slacks** on choosing K, then modifies clusters (Fig)
  - when inferred # clusters is as reported, $K^i = K^r$
    $$p_A(\varphi^r|\hat{\phi}^i) = \frac{1}{|\mathcal{S}(K^r)|}\sum_{\pi\in\mathcal{S}(K^r)} \mathrm{Dir}\left(w; \pi(\hat{n}^i)\right)\mathcal{N}(\mu^r; \pi(\hat{\mu}^i), s_\mu^2\mathbf{I})\mathrm{Log}\mathcal{N}(\sigma; \pi(\hat{\sigma}_k^i), s_\sigma^2\mathbf{I})$$
    averaging over all permutations $\mathcal{S}(K^r)$
  - when $K^i \neq K^r$: $p_A(\varphi^r|\hat{\phi}^i) = 0$, penalising slacks

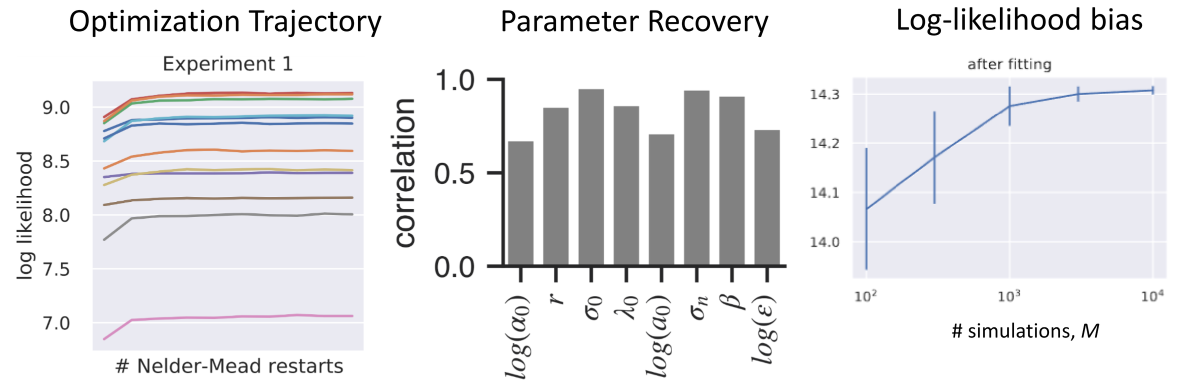
- **GPU-friendly array implementation**

Example implementation of sampling z_t given $z_{1:t-1}$ for $t = 6$ and $K_{\max} = 4$

$$\underset{\text{particles, } i}{\overset{\text{Counts, } n_k^i}{\begin{bmatrix} 2 & 3 & 0 & 0 \\ 5 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}}} \xrightarrow{\text{fancy index}} \overset{\text{Logits}}{\begin{bmatrix} 2 & 3 & \alpha_t & 0 \\ 5 & \alpha_t & 0 & 0 \\ 3 & 1 & 1 & \alpha_t \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}} \xrightarrow{\text{normalize, sample}} \overset{\text{OneHot}(z^i)}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}}$$

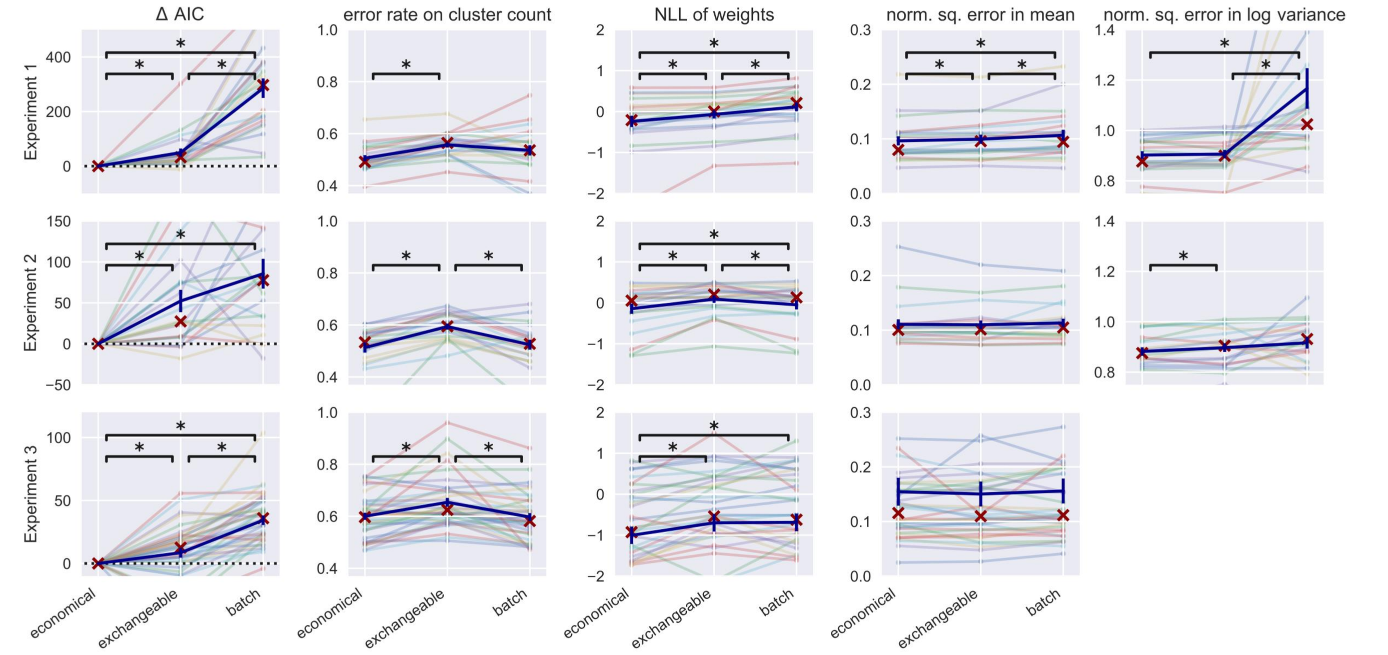
clusters, k

Large array operations can be accelerated using PyTorch with CUDA enabled

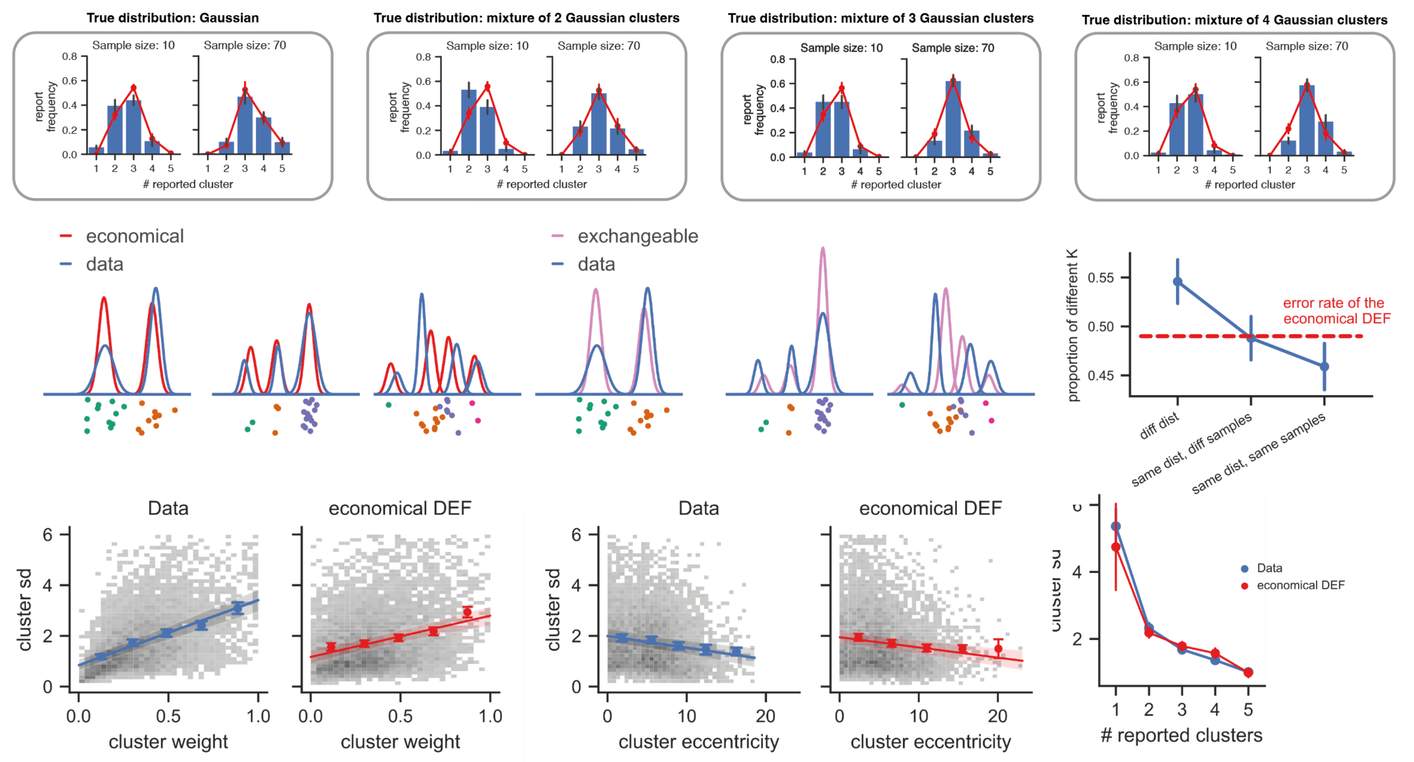


4. Modeling Results

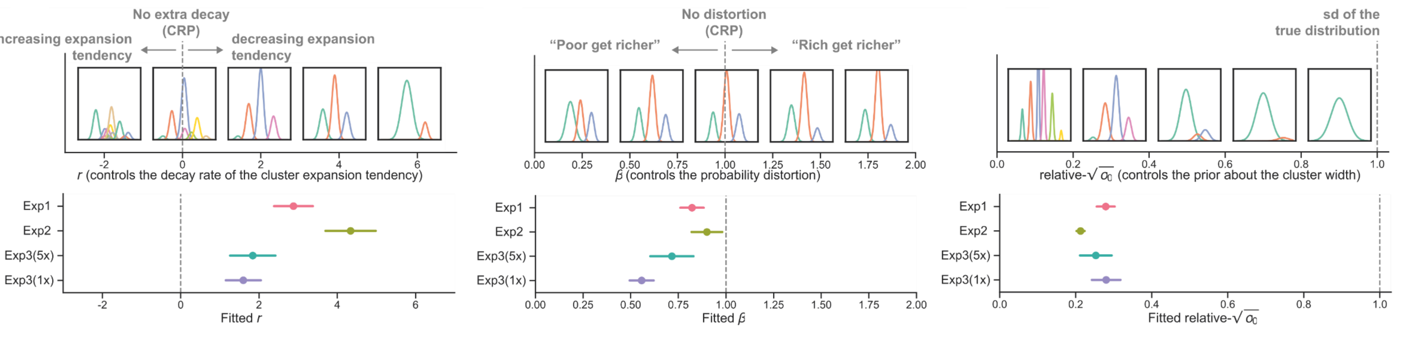
1. The economical prior fits best in all 3 Experiments



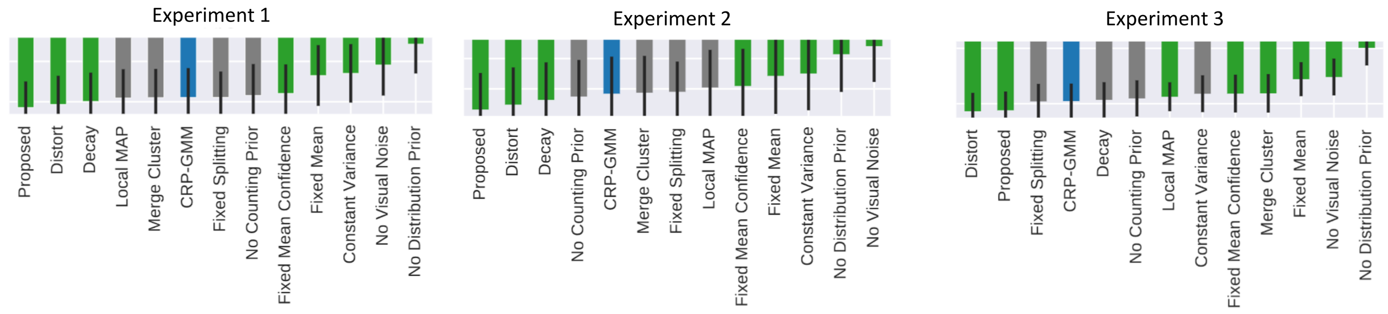
2. Captures human behavior



3. Reflects the construction bias



4. Survives extensive model ablation tests



References
Anderson, J. R. (1991). *Psych Rev*
Gershman, S. J., Blei, D. M., & Niv, Y. (2010). *Psych Rev*
Orban, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). *Neuron*
Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). *Psych Rev*
Whiteley, L. & Sahani, M. (2008). *Journal of Vision*
Zhang, H., Daw, N. D., & Maloney, L. T. (2015). *Nature Neuroscience*
Körding, K. P., & Wolpert, D. M. (2004) . *Nature*

tengtianyuan@pku.edu.cn kevinli@gatsby.ucl.ac.uk hang.zhang@pku.edu.cn T.T. is now at Max Planck Institute for Biological Cybernetics, Germany. L.K.W. is now at Google DeepMind.