How Occam's razor guides human inference

Eugenio Piasini*[1]; Shuze Liu*[2], Joshua Gold**[3,4], Vijay Balasubramanian**[3,4] (* indicates co-first authorship, ** indicates co-senior authorship)

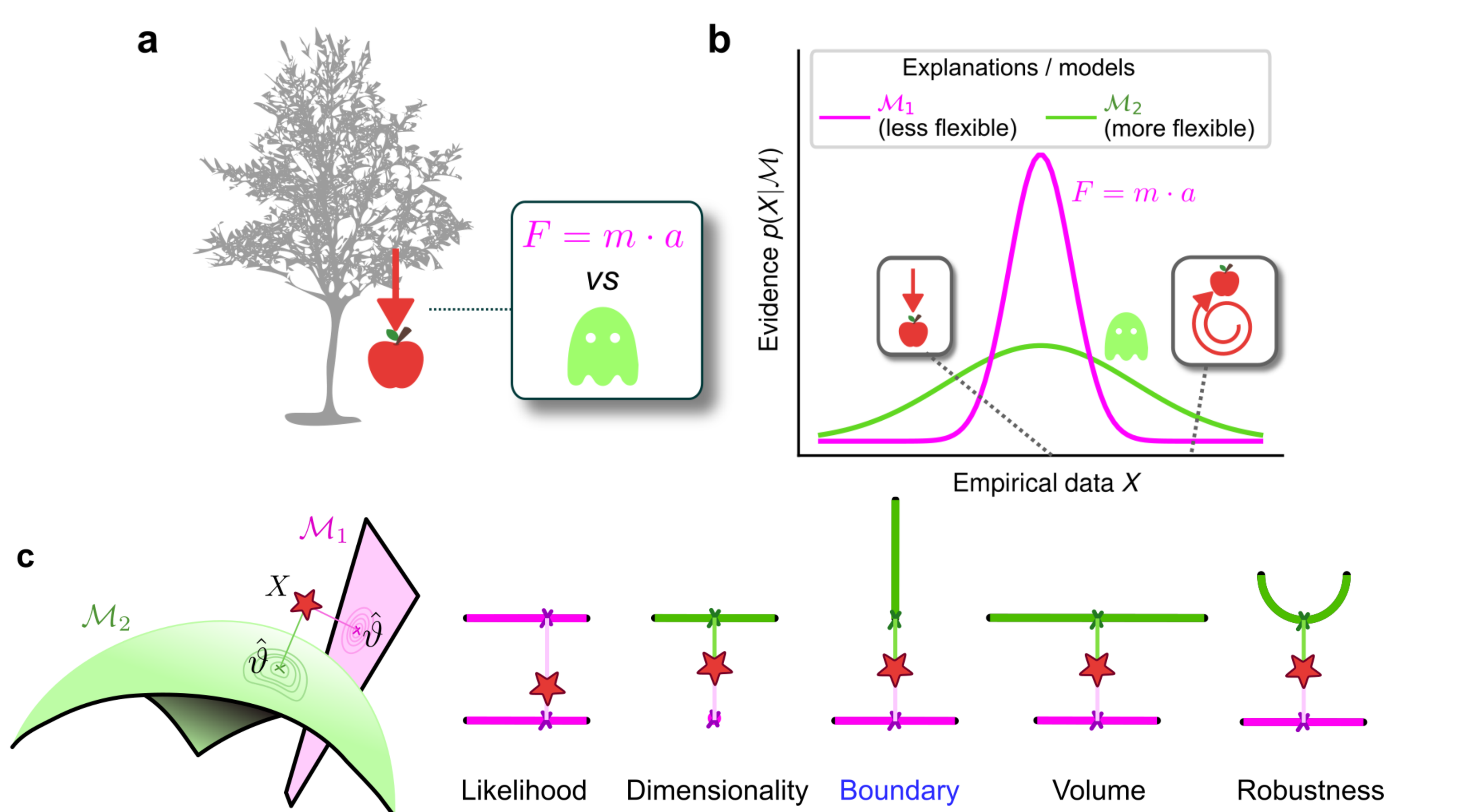
[1]International School for Advanced Studies (SISSA); [2]University of Pennsylvania College of Arts and Sciences (COL 2023); [3]University of Pennsylvania Neuroscience Graduate Group (NGG); [4]Computational Neuroscience Initiative (CNI)

Sponsored by Grants for Faculty Mentoring Undergraduate Research, UPenn Alumni Fellowship for Neuroscience Research

1. INTRODUCTION

- In **statistical model selection**, **Occam's razor** prefers the simplest model for a set of observations, when multiple models explain the observations equally well[1].
- Occam's razor plays a role in human perception and decision-making[2-6]. However, a **general, quantitative** account of the specific nature and impact of model complexity in human decisions is missing.
- When faced with uncertain evidence, naive human subjects bias their decisions in favor of simpler explanations, in a way that can be **quantified** by the framework of Bayesian model selection.

Bayesian Model Selection Framework[7]



d

$$-\log p(X|\mathcal{M}) = -\log p(X|\hat{\vartheta},\mathcal{M}) + \frac{d}{2}\log\frac{N}{2\pi} + \left(\frac{1}{2}\log\frac{N}{2\pi} + \log\left[2\pi\|\hat{l}\|_{\hat{h}^{-1}}\right]\right) + \log\int d^d\vartheta\sqrt{\det g(\vartheta)} + \frac{1}{2}\log\left[\frac{\det h(X;\hat{\vartheta})}{\det g(\hat{\vartheta})}\right] + \dots$$

Given a set X of N noisy observations, and possible statistical models $\{M_1, M_2, \dots\}$, we seek the model M that is the best explanation for X.

- Occam's razor penalizes more flexible/complex models (typically to avoid overfitting noise in the observations), while also considering the likelihood of observing X under each model.
- The optimal **trade-off between likelihood and model complexity** is given by Bayesian model selection. **The evidence for a model** requires integrating over its entire parameter (θ) space with respect to some prior $w(\theta)$ over parameters:

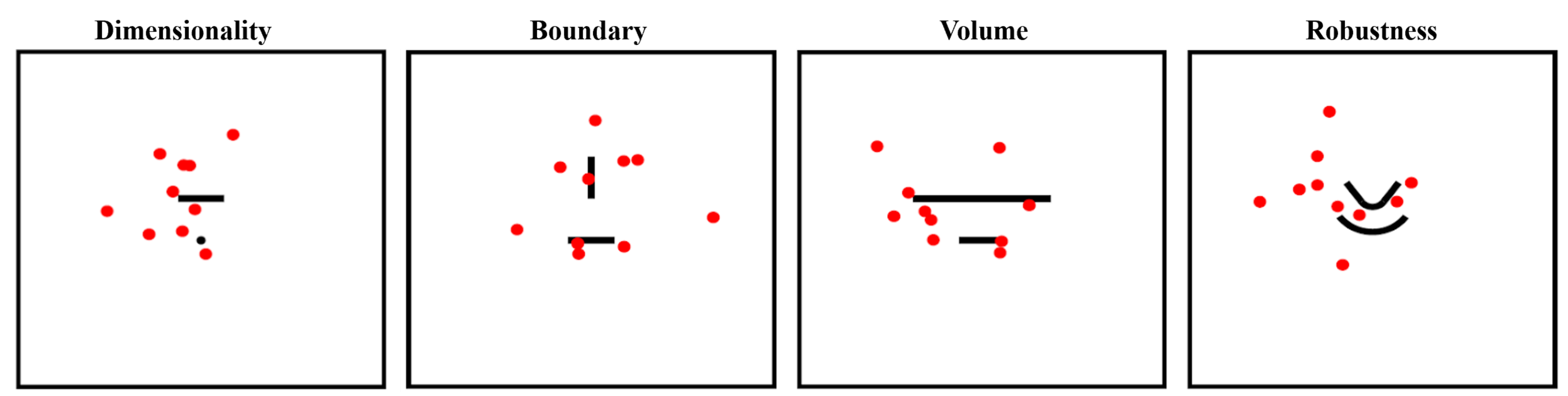
$$p(X|M) = \int d\theta\ w(\theta)\, p(X|M(\theta))$$

- **Fisher information approximation (FIA) of the log evidence ratio** identifies **four measures of model complexity** that the optimal decision depends upon:
  1. ***Dimensionality:*** *the number of parameters of the model (akin to BIC);*
  2. ***Boundary:*** *introduced when the maximum likelihood (ML) estimate along the model $M(\hat{\theta})$ for X is at the boundary of the model;*
  3. ***Volume:*** *range that the model can cover by varying θ in data space;*
  4. ***Robustness:*** *related to the curvature of the model towards/away from X.*

2. METHODS

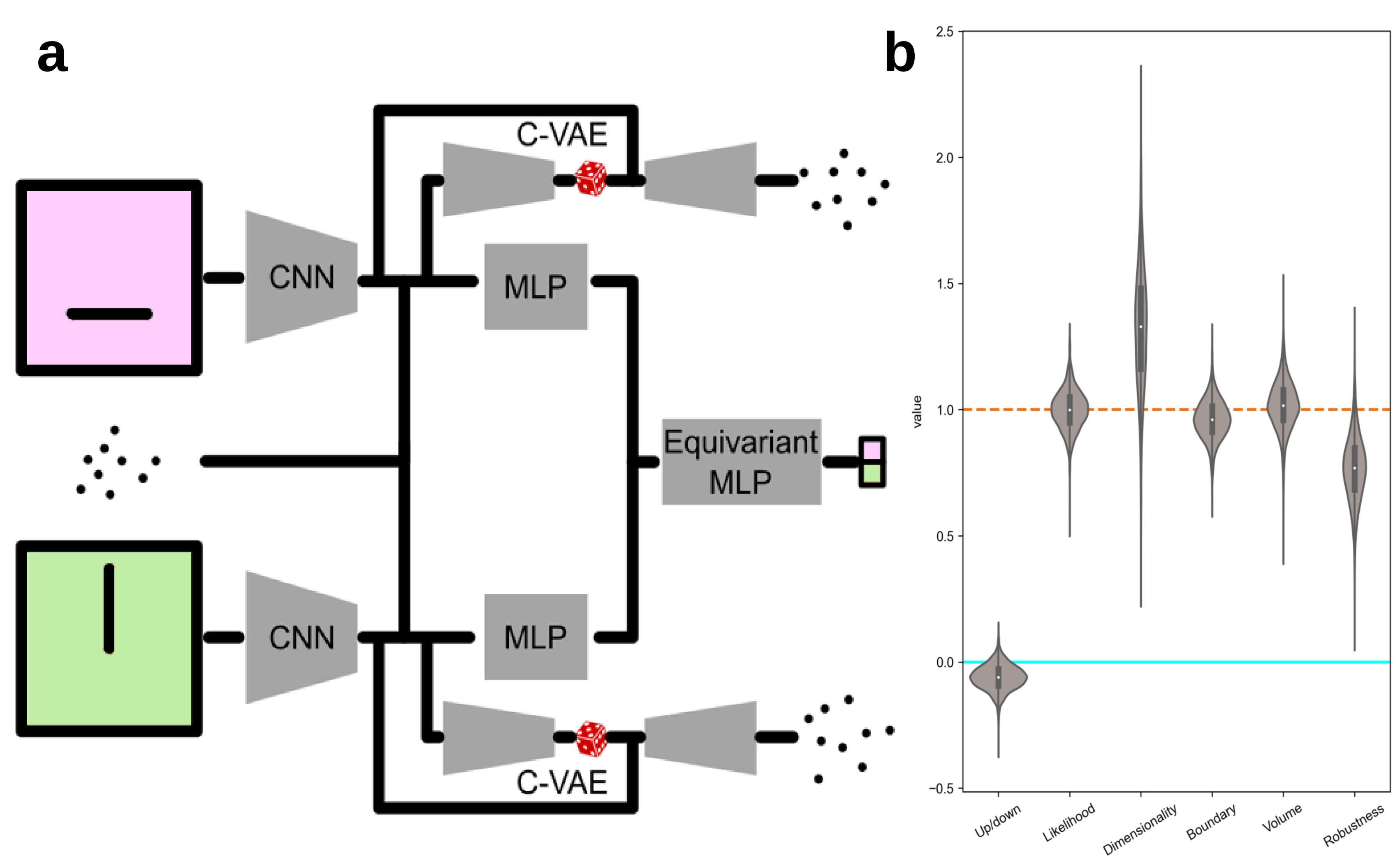
Perceptual Model Selection Task

- In each trial, given two possible shapes (black lines):
- Chose one shape to generate observations randomly with equal probability.
- Sample the location of a Gaussian center uniformly within the chosen shape.
- Sample $N = 10$ observations (red dots) from a 2D Gaussian with the center above.
- Given only the two shapes and 10 observations, the subjects decides which shape (model) is more likely to contain the center of the Gaussian.
- Four task types correspond to the four Bayesian model complexity measures.



Novel Deep Neural Network Architecture

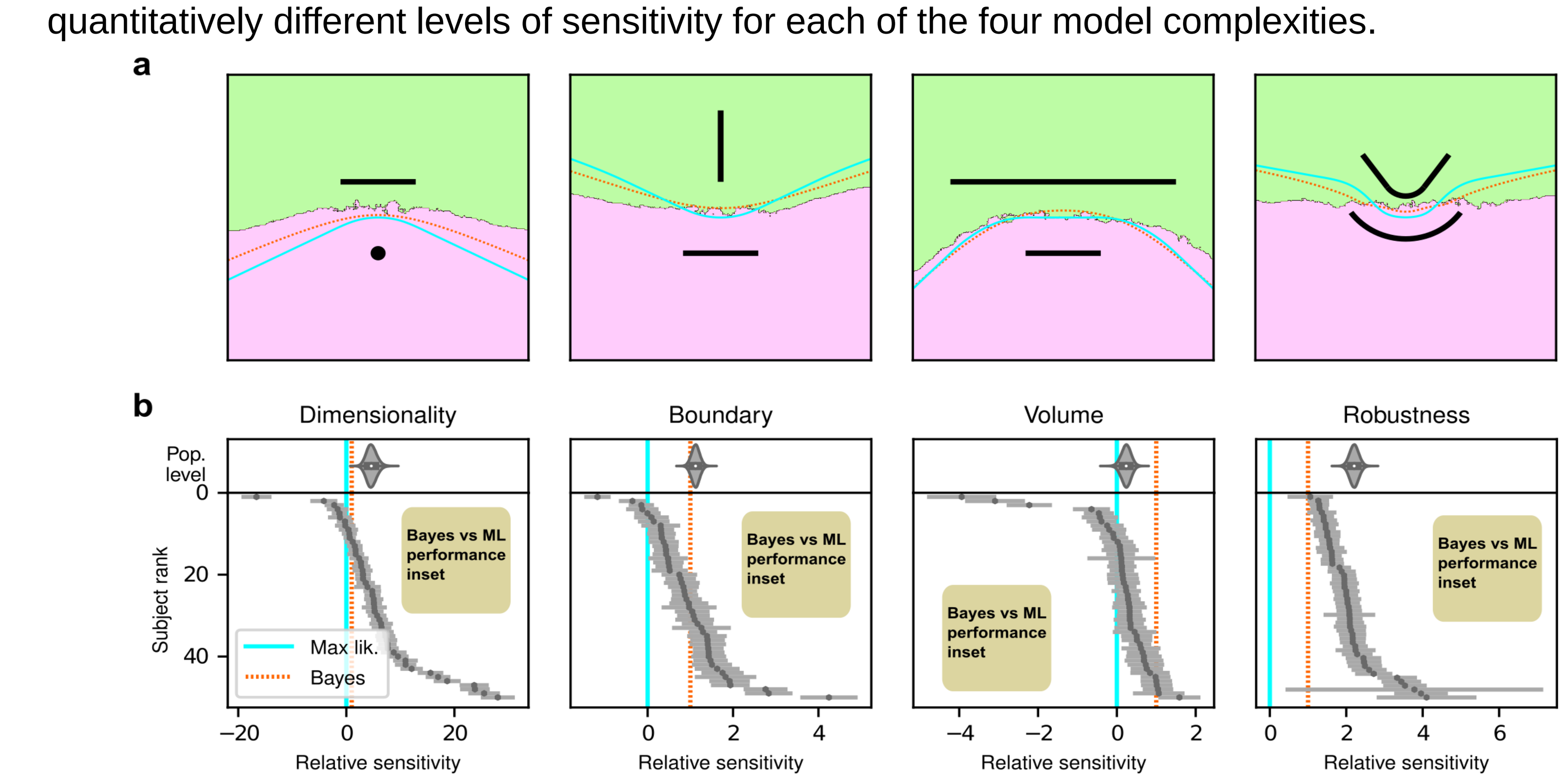
- We designed a novel deep neural network architecture and trained it on the task, to show that the complexity terms are **learnable** quantities that impact decision accuracy,



*(**a**) Novel deep network architecture. It intakes the observations as a data vector and the two models as images. It outputs a decision between the two models (encoded as a softmax vector). The pre-trained convolutional neural network (CNN) VGG16 is used for model image processing, and conditional variational autoencoders (C-VAE) assist learning of the data generation process. (**b**) Population level estimates of the sensitivity to each model complexity fitted for 10 trained networks, which differ only in random initialization. The logistic regression intakes a complexity-independent up/down bias term, the data likelihood (ml), and differences in the four Bayesian complexity terms computed for the two models, to predict the network's decisions. Dotted red line: a Bayesian observer would have sensitivity=1 to all complexities. Solid blue line: an observer without simplicity bias would have sensitivity=0 to all complexities. The networks' fitted coefficients normalized with respect to likelihood (reflecting relative sensitivities to likelihood and each complexity term) are close to the Bayesian optimal value of 1.*

3. RESULTS

- **Model selection:** data is better described by considering model complexities, instead of considering only the minimum distance between model and data and reporting the model with maximum likelihood.
- **Hierarchical Bayesian model:** Humans intuitively bias away from more complex models, but with quantitatively different levels of sensitivity for each of the four model complexities.



***(a)** K-nearest-neighbor interpolation of subject choices (green=up, pink=down), as a function of the centroid of the observations, pooled over $M = 202$ subjects. K is chosen by 10-fold cross-validation. **(b)** Population and subject level estimates of the sensitivity to each model complexity. Dotted red line: a Bayesian observer would have sensitivity=1 to all complexities. Solid blue line: an observer without simplicity bias would have sensitivity=0 to all complexities. Dark gray dots: posterior mean for complexity terms and bias/likelihood for individual subjects. Light gray bars: standard deviation of the posterior distribution for the same parameters. Subjects ranked based on the mean posterior (dark gray).*

Explain Different Sensitivities to Four Complexities

Control experiment: Point-Point and Line-Line model pairs.

- No difference in model complexity within each pair;

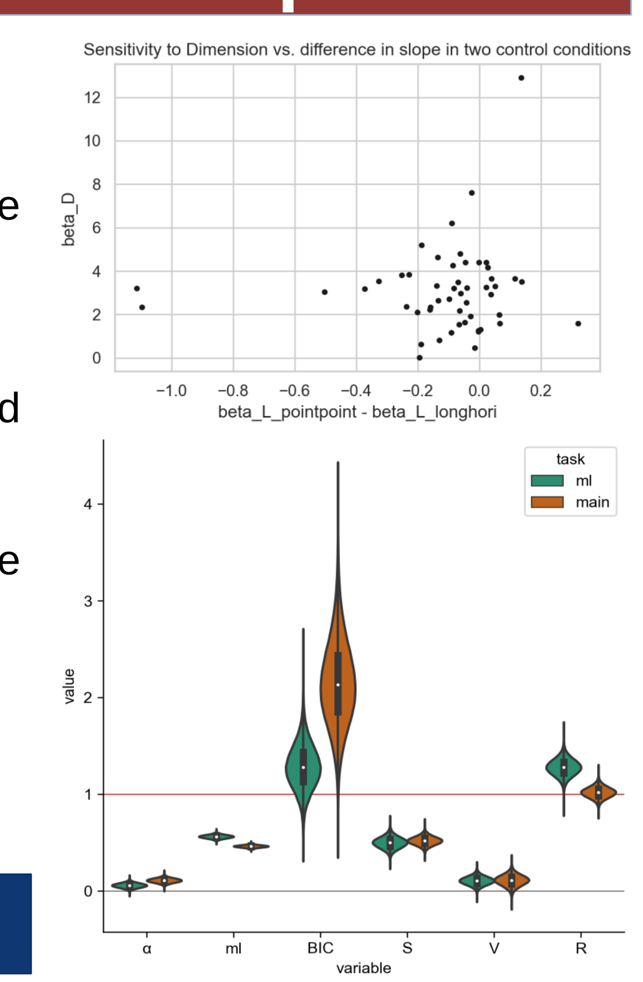
Higher sensitivity to Dimensionality is not related to difference in noise levels for the point and line individual models.

Possible Heuristic: counting the point model as the 11th data point.

- Then, regressing using the correct model complexity terms would overestimate the sensitivity to Dimensionality.

Maximum Likelihood task: ask human subjects to only consider the closest point along each model to the data points.

- Discourages consideration of model shape and thus complexity;
- Strong human tendency to remain sensitive to model complexity.



4. FUTURE DIRECTIONS

- Further investigate sensory noise in estimating the location of data points, as a potential source of sensitivity differences to different model complexities.
- Further validate the normative framework and human biases on other model selection tasks.

5. REFERENCES

1. Baker, A. Simplicity. in The Stanford Encyclopedia of Philosophy. (2022)
2. Pothos, E. M. & Chater, N. A simplicity principle in unsupervised human categorization. (2002).
3. Genewein, T. & Braun, D. A. Occam's Razor in sensorimotor learning. (2014).
4. Gershman, S. & Niv, Y. Perceptual estimation obeys Occam's razor. (2013).
5. Little, D. R. B. & Shiffrin, R. Simplicity Bias in the Estimation of Causal Functions. (2009).
6. Johnson, S., Jin, A. & Keil, F. Simplicity and Goodness-of-Fit in Explanation: The Case of Intuitive Curve-Fitting. (2014).
7. Balasubramanian, V. Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions. (1997).