

Neuroscience Area - PhD course in
Cognitive Neuroscience

A Bayesian Need for Simplicity

A study of how the brain selects and implements
internal models, using a normative approach
and Bayesian data analysis methods

Candidate:
Francesco G.
Rinaldi

Advisor:
Prof. Eugenio Piasini

Academic Year 2024-25



Acknowledgements

I would like to express my deepest gratitude to the many people who have guided, supported, and inspired me throughout the course of these four years and during the writing of this thesis. My first and most earnest thanks go to my supervisor, Eugenio, for the incredible dedication to your role and the guidance you have given me throughout these years. Given the insightfulness, balance, and patience you demonstrated from day one, it is genuinely hard to believe that I was your first PhD student. To be honest, you might have spoiled me for my future bosses.

Crucial contributions to this work were made by my collaborators Mattia and Davide, as well as Evan and Marc: thank you for making me part of your projects.

I extend my sincere appreciation to my colleagues of the Neural Computation Lab for the exchange of ideas and the (many) presentations and posters they helped improve with their comments and suggestions. Likewise, I am grateful to the School for the friendly and collaborative environment it fosters, and for financially supporting my studies.

A special thanks goes to my parents for the never-ending support throughout my whole journey, and for moving mountains whenever I'm in need. And to Bi, with whom I shared every important moment of these years, and who was always there in the hardest times. You're the closest thing I have to a brother.

I don't think I'll be able to thank Kendall enough. Before moving to Trieste, I couldn't even imagine that something as good as you would ever happen to me. I know I can be a difficult person to share time with (particularly when writing a thesis). Yet whenever I turn, you're always there at my side.

Finally, I was fortunate to spend these years surrounded by SISSA Very Important People. Nobody fits that description better than Claudia, who was with me from the very beginning of this PhD. I don't know how I could have survived that first year without you. Or the second. Or the third. Or the fourth. And the same goes for Giulio, Ben and Shre, together in SISSA up to the very last moment, Alfredo, the most genuine man I know, Clem, who boosted my self confidence to levels before unknown, the unmatched generosity of Paso, Marina's loads of music suggestion, Lodo and his impressive ability to cut vegetables, Fra's imitations (the

ones you were allowed to do, and the ones you weren't), the incredible thoughtfulness of Anant, and all the others who made these four years truly special. Thanks to all of you.

Abstract

A prominent hypothesis in cognitive neuroscience is that the brain operates as an inference engine, relying on internal models to interpret sensory input and guide decision-making. While a great variety of behaviors and biases are explained through this approach, it is less clear how the brain builds and, most importantly, chooses among different models to interpret its environment. A general bias toward simpler interpretations is well-established, but critical gaps remain regarding the computational mechanisms driving this preference. Specifically, it is unclear whether the brain assesses the simplicity (or complexity) of an interpretation through optimal, theory-aligned computations or if it relies on heuristic approximation that can be skewed by perceptual biases. Furthermore, we don't know if the measure of complexity is fixed or adapted to the detailed features of the models at hand.

The first part of this work examines the criteria humans use to select between alternative interpretations of noisy data. First, investigating the effect of a variable amount of data, we show that human interpretations, while qualitatively following the principle of parsimony, diverge from optimal behavior. Our results suggest that intuitive model selection is based on a perceptual evaluation of the sample size, rather than a principled computation of the model fit. Second, we address which properties of the model affect the perception of its complexity. We show that the brain does not rely on a generic complexity metric, as some model selection criteria do. Instead, characteristics unique to the models considered have a direct effect on the choice between interpretations.

Broadening the scope from intuitive model selection to biological implementation, the latter half of this work investigates how internal models shape sensory perception and how they are instantiated in neural populations, leveraging the same probabilistic frameworks used in the first half. In the domain of sensory processing, we apply Bayesian inference to explain perceptual biases in rats. We show that the influence of task-irrelevant sounds on visual discrimination is best explained by an internal model affected by a compressive warping of visual representations caused by auditory inputs, providing behavioral evidence for direct sensory interaction. Finally, we develop a novel Bayesian method for analyzing neural population activity. Applied to data from mice performing a value-based decision-making

task, this technique allows for the decoding of latent variables required to construct and update an internal model of reward contingency.

Taken together, these findings offer a multi-level perspective on internal models, from cognition to perception to neural implementation, shedding light on the algorithms the brain uses to manage uncertainty and providing methodological tools to investigate them further.

Table of contents

List of figures	xi
List of tables	xxv
1 Introduction	1
1.1 Internal models of the brain	1
1.2 Bayesian models in cognition	3
1.3 Cognitive application of Bayesian methods	5
1.4 Model Selection	7
1.5 Thesis Outline	12
2 Sample size dependence of Occam's razor in human decision-making	15
2.1 Introduction	15
2.2 Results	19
2.3 Discussion	25
2.4 Methods	28
2.4.1 Behavioral experiment	28
2.4.2 Experimental data analysis	30
2.4.3 Bayesian regression	38
2.4.4 Model comparison	43
3 Characterization of Complexity in Intuitive Model Selection	47
3.1 Introduction	47
3.2 Sublinear dimensionality	49
3.3 Laplace approximation of Bayesian Model Selection	53
3.4 Corrections to BIC	55
3.5 Remainder term	59
3.6 Discussion	63
3.7 Supplementary Material	65

3.7.1	Experimental Data Analysis	65
3.7.2	Multiplicity threshold of the Symmetry term	67
4	Sounds compress rat visual perceptual space	69
4.1	Introduction	69
4.2	Results	71
4.2.1	Task-irrelevant sounds affect rat perception of drifting gratings in a way that is independent of the temporal congruency between the auditory and visual stimuli	73
4.2.2	A computational model accounts for the behavioral impact of task-irrelevant sounds as the result of visual perceptual space compression	78
4.3	Discussion	84
4.4	Materials and methods	88
4.4.1	Subjects	88
4.4.2	Behavioral apparatus	89
4.4.3	Visual and auditory stimuli	90
4.4.4	Estimation of sound intensity	91
4.4.5	Behavioral task	94
4.4.6	Experimental design	94
4.4.7	Statistical analyses	97
4.4.8	Ideal observer model	98
4.4.9	Inferring the model parameters from experimental data	105
4.4.10	Model comparison	107
5	A hierarchical Gaussian process model to study the activity of neuronal populations	111
5.1	Introduction	111
5.2	Experiment	114
5.3	Bayesian Hierarchical Model	115
5.4	Results	120
5.5	Discussion	127
6	Conclusion	133
	References	145
	Appendix A Supplementary material	161
A.0.1	Supplementary Tables	161

A.0.2	Marginal Likelihood of Gaussian Mixture Models	164
A.0.3	Higher-Order Correction to the Laplace Approximation	165
A.0.4	Supplementary Tables	168
A.0.5	Supplementary Figures	171
A.0.6	Supplementary Tables	178
A.1	Chapter 5	179
A.1.1	Hilbert Space Gaussian Process	179
A.1.2	Supplementary Tables	182

List of figures

- 2.1 **Balancing of goodness-of-fit and complexity in Bayesian Model Selection.**
- a.** Bayesian Model Selection (BMS) between a model with one parameter (Model 1, linear model with fixed slope = 0, blue) and a model with two parameters (Model 2, linear model, orange). Given the same prior for both models and uniform prior across their parameters, the best model will be the one with the higher likelihood averaged across all the possible choices of parameters (dashed box, left). This quantity, the marginal likelihood, corresponds to the integral of the likelihood times the prior (dashed box, right). By using Laplace approximation, this integral can be decomposed in the product of a “height” term (proportional to the maximum likelihood of the model) and “width” terms (proportional to the inverse square root of the number of data-points N). In the logarithmic space, these terms correspond to the log-likelihood L_M and the dimensionality D_M of the Bayesian Information Criterion (BIC, solid-lined box). **b.** Effect of increasing the sample size. The height of the likelihood landscape is exponential in the number of points N (since it is the product of the single-point likelihood of N data-points), while the widths of the peaks decrease as $N^{-1/2}$. In the logarithmic space, this translates to a log-likelihood term that grows linearly in N and a dimensionality that grows logarithmically. Since the goodness-of-fit grows faster than the complexity penalty, the simplicity bias gets relatively weaker for higher sample sizes, allowing for the selection of more complex models (solid-lined box).

- 2.2 **Task, models and participants' answers.** **a.** The five models that participants had to choose among (bottom) and an example of a trial from the task (top). The trial shown was generated by a model with three components; two had close means and generated the cluster on the left, one had an isolated mean and generated the cluster on the right. **b.** Average confusion matrix across participants and trials: while for models with low dimensionality ($K_M = 1, 2$) the answers mostly correspond to the true model, when the number of parameters increases, the answers are skewed towards values lower than the true ones, showing the effect of the simplicity bias. **c.** Average answer across all participants and all trials with N data-points. The values consistently fall below the true mean $\langle K_M \rangle = 3$, once again indicating the presence of a simplicity bias. However, the upward trend suggests a gradual weakening of the bias as N increases, as quantified by the Weighted Least Squares regression (dashed line; slope: 0.0012, one-sided p-value 0.04), supporting the hypothesis of a scaling in N 20
- 2.3 **Model comparison between candidate descriptions.** **a.** suPareto-Smoothed-Important-Sampling Leave-One-Out Cross-Validation (PSIS LOO CV) comparison between the two Goodness-of-fit descriptions (W-I and Likelihood-only) and two descriptions based on AIC and BIC. The x-axis shows the Expected Log Predictive Density (ELPD) of each description. The errorbars represent the standard error of the ELPD differences with the best model, $\text{ELPD}_{\text{best}} - \text{ELPD}_M$. The inset shows the PSIS LOO comparison between the Likelihood-only and the W-I descriptions, showing that the Likelihood-only description has a significantly higher ELPD than the W-I one. **b.** ELPD comparison between the “computation”, “independent” and “heuristic scaling” description. The “computation” ones perform significantly worse than the others. The inset highlights that the difference between the AIC “heuristic” and AIC “independent” descriptions is within three standard deviations. **c.** ELPD comparison between the AIC “heuristic”, AIC “independent” and “free scaling” descriptions. The “independent” description is the only one to perform significantly worse than the others. 21

- 2.4 **Graphical representation of the three descriptions we used to model the participants' strategies and dependence of their terms on the sample size N .** Only the BIC versions are shown. The “computation” description (top left) assumes that participants' choices depend on a computation like Bayesian integration: as shown in Fig 2.1a, the integration sums the likelihoods for each possible parameter value, here represented as different values of the components' means. The resulting goodness-of-fit scales linearly in N , while the resulting complexity penalty scales logarithmically (bottom left). The “independent” description (top center) assumes that participants only rely on the density of the data (gray line) and compare it with the best-fitting shape of the models. In the figure, the red areas represent the difference between the empirical data distribution and the model distributions. Both the goodness-of-fit and the dimensionality are constant in N (bottom center). The “heuristic scaling” description (top right) assumes that participants rely on a process similar to that in the “independent” description, but try to recover the optimal scaling by weighting the result by the perceived number of data-points $\tilde{N}$. The resulting goodness-of-fit scales as the perceived numerosity of the data, which we assume to be $\tilde{N} = \log(N)$ (Fechner, 1860), while the penalty scales as $\log(\tilde{N})$ 23
- 2.5 **Scaling of the perceived sample size. a.** The inferred functions $\tilde{N} = GP(N)$, computed for both the AIC (red) and BIC (blue) “free scaling” descriptions, compared with the linear scaling of the “computation” descriptions (gray), the flat scaling of the “independent” description (green) and the logarithmic scaling of the “heuristic scaling” descriptions (pink). Both the AIC and the BIC functions follow a logarithmic trend for low sample sizes, and have a transition around $N = 50$. **b.** The three stages of numerosity perception: errorless (orange), Weber-like (pink), and density-based (green). Image adapted from Burr et al. (2018). 25
- 3.1 **Task, models and participants' answers. a.** The five models that participants had to choose among (bottom) and an example of a trial from the task (top). The trial shown was generated by a model with three well-separated components. **b.** Average confusion matrix across participants and trials: as in Fig. 2.2c, the effect of the simplicity bias is evident from the fact that the distribution of answers for models with more parameters is skewed towards lower-value answers. 50

- 3.2 **Parameters and comparison of the “computation” and “free penalty” descriptions.** **a.** Sensitivities to likelihood (β_L) and dimensionality (β_D) for the “computation” description across the 46 participants. For each plot, the red vertical line represents the posterior mean of the population average μ_β , while the distance between the red line and each of the gray lines represents the population standard deviation σ_β . The dots represent the posterior mean of the values of each participant, and the errorbars are the s.e.m.. The values of both β_L and β_D are sorted independently in each panel for ease of visualization. **b.** Result of the LOO CV comparison between the BIC “computation” description and the “free penalty” description. The x-axis shows the Expected Log Predictive Density (ELPD) of each description, while the errorbars represent the standard error of the ELPD differences with the best model, $\text{ELPD}_{\text{best}} - \text{ELPD}_M$. **c.** Inferred values of the perceived number of parameters $\hat{K}_M$. The green line is the posterior mean of the population average $\mu_{\hat{K}_M}$. The green shaded area is the mean $\pm$ s.e.m.. The gray lines are the values of $\hat{K}_M$ for each participant. 52
- 3.3 **Comparison between “computational”, “free parameters”, and the two “correction” descriptions.** The x-axis shows the Expected Log Predictive Density (ELPD) of each description, while the errorbars represent the standard error of the ELPD differences with the best model, $\text{ELPD}_{\text{best}} - \text{ELPD}_M$. 56
- 3.4 **Parameters of the two “correction” descriptions.** **a.** Inferred sensitivities to likelihood (β_L), dimensionality (β_D) and naive symmetry ($\beta_{S_{\text{naive}}}$) for the “naive correction” description. For each plot, the red vertical line represents the posterior mean of the population average μ_β , while the two gray lines represent the population standard deviation σ_β . The dots represent the posterior mean of the values of each participant, and the errorbars are the s.e.m.. The values of all β are sorted. The shaded gray area indicates where the inferred values are negative. **b.** Same as for point **a**, but for the “correction” description. 57

- 3.5 **Parameters and comparison of the “FIA” description.** **a.** Inferred sensitivities to likelihood (β_L), dimensionality (β_D), symmetry (β_S) and remainder term (β_R) for the “FIA” description. For each plot, the red vertical line represents the posterior mean of the population average μ_β , while the two gray lines represent the population standard deviation σ_β . The dots represent the posterior mean of the values of each participant, and the errorbars are the s.e.m.. The values of all β are sorted. The shaded gray area indicates where the inferred values are negative. **b.** Result of the LOO CV comparison between the BIC “computation” description, the “correction” description and the “FIA” description. The x-axis shows the Expected Log Predictive Density (ELPD) of each description, while the errorbars represent the standard error of the ELPD differences with the best model, $\text{ELPD}_{\text{best}} - \text{ELPD}_M$ 61
- 3.6 **Failure of Laplace approximation for close components.** **a.** Comparison between the marginal likelihood and the Laplace approximation for a model with $K_M = 2$ on data from two clusters at varying distances. On the x-axis, the distance between the means of the two clusters. The inserts show some of the resulting compositions. In black, the marginal likelihood computed numerically. In red, the corresponding Laplace approximation up to the $O(1)$ term, multiplied by 2 (to account for the presence of two peaks in the likelihood landscape). The y-axis is in logarithmic scale. Notice that, since we simply multiply by 2, here we are using the S_{naive_M} term. The approximation fails when the clusters become too close. Interestingly, the approximation starts working better again in the proximity of $\Delta\mu = 0$, particularly if we remove the $\times 2$ factor, thus applying the S_M term. **b** Likelihood landscape for the clusters in **a** at $\Delta\mu = 0.1$, i.e., where the Laplace approximation explodes. The two peaks generated by the symmetry are merging, thus creating a ridge. The y-axis scale factor is 10^{18} . Note that only the landscape around the peak is shown. 62

- 3.7 **Comparison between the marginal likelihood and Laplace approximation.** **a.** Ratio between the marginal likelihood, integrated numerically, and the Laplace approximation for a model with $K_M = 3$ on data from three clusters at varying distances. The distances are always computed relatively to a central cluster. As in Fig 3.6a, the three clusters are always the same and only their relative position changes. On the x-axis, the distance between the central cluster and the cluster on the left. On the y-axis, the distance between the central cluster and the cluster on the right. The distances are shown up to a maximum of $\Delta\mu = 0.5$, even if the values could go as far as $\Delta\mu = 2$, to focus on the cases of close clusters. The inset shows the value of $\mathcal{N}_{\text{peaks}}$ for each value of $\Delta\mu_1$ and $\Delta\mu_2$. The colors follow the same scale as the main figure, indicating six (yellow), three (purple), or one (dark purple) peaks. **b.** Same as **a**, but the Laplace approximation is multiplied by $\mathcal{N}_{\text{peaks}}$ 63

- 4.1 **Experimental setup and audio-visual stimulus conditions.** **A.** Rendering of a rat extending its head through a viewing hole in the wall of a sound-isolated operant box, equipped with a display and a speaker for presentation of the audio-visual stimuli. The hole was placed in front of the stimulus display and the speaker (in cyan). The hole also allowed access to three nose-poke ports used to trigger stimulus presentation and report the temporal frequency of the visual stimuli. **B.** Schematic representation of a behavioral trial. As in previous studies ([Caramellino et al., 2021](#); [Matteucci et al., 2021](#)), the rats learned to protrude their head through the viewing hole and trigger stimulus presentation by touching the central nose-poke port. The visual stimuli were circular, outward-moving sinusoidal gratings that rats classified according to their temporal frequency (TF). Gratings with $TF < 2.12$ Hz had to be classified as “Low TF”; those with $TF > 2.12$ Hz as “High TF”. Each class was associated with one of the two lateral nose-poke ports. Auditory stimuli did not provide any information about the correct classification of the visual stimuli. **C.** Stimulus conditions presented to the rats during the test phase of the experiment. These included: 1) visual gratings paired with fixed amplitude sounds (green); 2) visual gratings paired with AM sounds (cells of the square matrix: blue, gray and red refer, respectively, to congruent, incongruent and anti-congruent audiovisual conditions; the shades of gray, from light to dark, indicate AM sounds with progressively larger intensity; see Methods and [A.2 Fig](#) for details); 3) unimodal purely visual gratings (purple); and 4) unimodal purely auditory AM sounds (cyan). **D.** Details on the auditory stimuli. Green dot: white noise burst with a fixed maximal amplitude. Gray dots: amplitude-modulated white noise bursts, each with a different temporal frequency of the sinusoidal envelope. Purple dot: absence of an auditory stimulus. See Methods and [A.2 Fig](#) for details on estimation of average perceived sound intensity. Error bars: s.e.m. across rats. For the modulated noise bursts, different shades of gray are used to group stimuli based on their average sound intensity (same as in **C**). Insets: sound amplitude control signal for white noise burst (green), the absence of auditory stimulus (purple) and one example of modulated noise burst (gray).

- 4.2 **Discrimination accuracy of two example rats across the different phases of the training procedure.** **A** The line plot (top) shows the percentage of correct choices of example rat 1 in classifying the temporal frequency of the drifting gratings across more than 70 daily training sessions. The color of the dots indicates the training phase (from 2 to 5; see caption), while the intensity of the gray patches denotes the various stages of contrast reduction that were performed during the training (see legend and the Materials and Methods). The dashed horizontal line represents the criterion accuracy (70%) required to progress to the next training phase. To advance, rats had to achieve performance above criterion for at least two consecutive days within the same training phase. The bar plot (bottom) shows the number of behavioral trials that were collected for the rat in each daily session. **B**. Same as **A**, but for example rat 2. 75
- 4.3 **Task-irrelevant sounds strongly modulate rat classification of the temporal frequency of drifting gratings.** **A**. Group average proportion of “high TF” choices ($n=10$ rats) as a function of the TF of the visual gratings, when the latter were paired with: 1) fixed amplitude sounds (green); 2) congruent AM sounds (blue); and 3) anti-congruent AM sounds (red). The error bars denote s.e.m. over the group of 10 rats. The curves are the result of logistic regressions to the data points. The inset illustrates which of the stimulus conditions shown in Fig 4.1C contributed to the data points/curves. See Table 4.1 for a statistical comparison among the three curves. **B**. Same as in **A**, but for stimulus conditions where the gratings were paired with: 1) fixed amplitude sounds (green; same data as in **A**); 2) AM sounds (black); and 3) no sounds (purple). See Table 4.2 for a statistical comparison among the three curves. 76

4.4 Schematic of an ideal observer model where neuronal populations encode linearly visual temporal frequencies and are suppressed by concomitant sounds in a way that depends on their intensity.

A. Hypothesized tuning of a visual cortical neuron in the ideal observer model. The neuron responds to the visual gratings with an average firing rate that grows linearly as a function of the grating TF. Due to neuronal noise, the responses to repeated presentations of the same grating are distributed according to a Gaussian function centered on the average firing rate elicited by the grating. More specifically, the red and cyan dots refer to those responses that are elicited by the gratings that, in our task, had to be classified as belonging to the “low TF” and “high TF” classes, respectively. When only the gratings are shown (purely visual condition), the neuronal response evoked by any given TF reaches its maximal possible value (purple line). When the gratings are paired with concomitant sounds, the responses to all the TFs are inhibited by a factor that depends on sound intensity, resulting in a reduction of the slope of the linear dependency between firing rate and TF. Since the AM sounds have a lower average intensity than the fixed amplitude sound (see Fig 4.1D, Materials and Methods, and A.2 Fig), the resulting compression of the firing rate is larger for the latter (green line) than for the former (gray line). **B.** Illustration of the ideal observer model when a population of multiple neurons is considered. The drawing shows a hypothetical neuronal representational space made of two units. Each neuron behaves as shown in A. In the example, two TFs yields two distributions of population vectors (red and cyan dots). When the gratings are paired with a sound with strong intensity (e.g., the fixed amplitude noise burst), all responses are inhibited and, as a result, the distributions are compressed toward the origin of the representational space (green curves). Since rats are trained to discriminate the gratings under these conditions, they learn a decision boundary (dashed line) that optimally separates such compressed distributions. When the animals are presented with unimodal, purely visual gratings, the sound-induced inhibition is released and the response distributions shift towards higher firing rate values (purple curves). The animals, however, still rely on the previously learned decision boundary to discriminate these stimulus conditions. As a result, the overall proportion of “high TF” choices increases (i.e., more red dots fall on the “false alarm” side of the decision boundary).

- 4.5 **An ideal observer model where task-irrelevant sounds compress rat visual perceptual space according to their average intensity accounts for rat perceptual choices.** **A.** Group average proportion of “high TF” choices ($n=10$ rats) as a function of the TF of the visual gratings, when the latter were paired with: 1) fixed amplitude sounds (green); 2) AM sounds with increasingly larger TFs (progressively darker shades of gray, matching those in the stimulus matrix shown in Fig 4.1C); and 3) no sounds (purple). The error bars denote s.e.m. over the group of 10 rats. The curves are group-level predictions of the proportions of “high TF” choices according to the Bayesian ideal observer model defined in Eq (4.1), with a single value of the sensitivity parameter σ and 5 different values of the scaling factor γ , one for each level of sound intensity (more precisely, each curve is the mean of the distribution of psychometric curves induced by the posterior distribution of the σ and γ parameters obtained for a given sound intensity level). For this model and the data as presented in the plot, $R^2 = 0.9973 \pm 0.0002$ (posterior mean $\pm$ st. dev.). The inset shows the difference between the expected log predictive density (ELPD) of the reference model with 1 σ and 5 γ values (i.e., the one yielding the psychometric curves) and variants of the model having a given number of free σ and γ parameters (as indicated on the ordinate axis). In all comparisons, the reference model was the one with lowest ELPD, i.e., with the higher predictive power. **B.** Relationship between the magnitude of the scaling factor γ in the ideal observer model and the intensity of the sounds that were paired with the visual gratings. Triangle, square and cross symbols refer to individual rats, while circles are group averages. Sound conditions are labeled according to the same color code as in **A** and Fig 4.1C. Vertical error bars denote the std. dev. of the posterior distribution of the group-level γ parameters (see Methods). The horizontal error bars denote s.e.m. over the group of 10 rats. 82
- 4.6 **Outline of the hypothesized interaction between auditory and visual cortex underlying the sound-driven compression of rat visual perceptual space.** Activation of A1 (in cyan) by sound bursts leads to an inhibition of the responses evoked by circular drifting gratings in V1 (in green) via cortico-cortical projections (orange line). This produces a decrease in the gain of the tuning curves of V1 neurons over the temporal frequencies of the gratings. 89

- 5.1 **Inferred responses in time of the LTSI (continuous lines) and THIN (dashed lines) neuronal population to binary-only predictors.** Each row represents one of the variables included in the model. The left column shows the response after the start of the trial. The right column shows the response around the reward delivery. The time of reward delivery is marked by a vertical dashed line. On each box of the plot, the colored line represents the posterior mean of the inferred function, and the shaded area represents a 90% CI. 121
- 5.2 **Inferred responses in time of the LTSI (continuous lines) and THIN (dashed lines) neuronal population to binary and continuous predictors.** Details as in Fig. 5.1. 123
- 5.3 **Posterior correlations of spectral coefficients across behavioral predictors in LTSI.** The heatmaps display the pairwise Pearson correlation coefficients between the population-level basis function weights (z_k) for the Go cue (left) and Outcome (right) kernels. Rows represent pairs of predictors, while columns correspond to the indices of the Hilbert space basis functions, ordered from low to high frequency. The color scale is common to both panels, centered at zero (white), with red indicating positive correlation and blue indicating negative correlation. 124
- 5.4 **Posterior correlations of spectral coefficients across behavioral predictors in THIN.** The heatmaps display the pairwise Pearson correlation coefficients between the population-level basis function weights (z_k) for the Go cue (left) and Outcome (right) kernels. Rows represent pairs of predictors, while columns correspond to the indices of the Hilbert space basis functions, ordered from low to high frequency. The color scale is common to both panels, centered at zero (white), with red indicating positive correlation and blue indicating negative correlation. 125
- 5.5 **Model comparison between binary-only and complete models.** Pareto-Smoothed-Important-Sampling Leave-One-Out Cross-Validation (PSIS LOO CV) comparison between the binary-only model ("binary") and the model with both binary and discrete predictors ("full") for LTSI (left) and THIN (right). The x-axis shows the Expected Log Predictive Density (ELPD) of each description. The errorbars represent the standard error of the ELPD differences with the best model, $\text{ELPD}_{\text{best}} - \text{ELPD}_M$. The vertical dashed line marks the ELPD value of the best model. 126

5.6	Posterior predictive simulation of the model with continuous predictors on THIN data Each box represents one trial of subject 1 of the THIN population. Only the first 60 trials of the dataset are shown. On each box, the gray line represents the true neuronal activity, the red line represents the average of the posterior draws, and the gray shaded area represents one standard deviation from the average. For each trial, 1500×4 simulations were drawn, one for each sample of the four MCMC chains. A yellow line on each box shows the reward delivery time. In areas outside of both GP windows, where no simulation was possible, a dashed purple line represents the best guess for the data (i.e., a flat line at 0, since the data was standardized), and the light purple shaded area shows the inferred noise of the neural activity.	128
5.7	Posterior Predictive Check (PPC) of the average LTSI population responses by predictor and condition. Each row represents one of the variables included in the model. The two leftmost columns show the trials aligned with the Go cue. The two rightmost columns show the trials aligned to the choice outcome. The times of alignment are marked by a vertical dashed line. The first of each pair of columns represents the trials where the variable was 1 (for binary predictors) or greater than 0 (for continuous predictors). The second represents the trials where the variable was 0 (for binary predictors) or less than 0 (for continuous predictors). On each box of the plot, the gray line represents the trials where the corresponding variable followed the corresponding condition, averaged. The gray shaded area represents the standard error of the mean (SEM). The colored lines represent the posterior expectation of the mean response. Shaded colored regions indicate the Posterior Standard Deviation of the Mean.	129
5.8	Posterior Predictive Check (PPC) of the average THIN population responses by predictor and condition. Details as in Fig. 5.7.	130
A.1	Rats are not sensitive to the temporal frequency of the unimodal, purely auditory stimuli. Group average proportion of “high TF” choices (n=10 rats) as a function of the TF of the unimodal auditory stimuli. Error bars are SEM.	171
	Sound power computation	172

- A.2 The auditory stimuli and their intensity.** Each row refers to a distinct auditory stimulus used in the experiment: 1) the white noise burst with fixed maximal amplitude (top row; green); 2) the amplitude modulated white noise bursts, with the 9 different temporal frequencies of the sinusoidal envelopes (middle rows; gray); and 3) the absence of the auditory stimulus (bottom row; purple). **A.** Waveforms of the auditory stimuli used to drive the speakers that delivered the sounds to the rats (note that, for the sake of visualization, the sounds are not shown at their actual sampling frequency of 44.1 KHz, but they have been downsampled to 500 Hz). **B.** Intensities of the auditory stimuli in dB as a function of time, as computed based on the waveforms shown in **A** and the measured intensity of 10 s long white noise bursts with different amplitudes (see the Materials and Methods for details). The panel referring to the TF of 0.25 Hz illustrates how the average sound intensity experienced by a rat in a given trial was computed, based on the reaction time RcT of the animal. The latter was obtained by subtracting the estimated motor response time $MoReT$ to the measured response time RT . **C.** Distributions and estimates of the average sound intensities experienced by each rat across all trials recorded for any given stimulus condition. Scatter plots show the estimated sound intensity experienced by the rats in each individual trial. Violin plots represent the full probability density distributions of these estimates while box plots show their median and interquartile range. The dashed line indicates the across-rat average intensity per stimulus condition. 173
- A.3 Schematic of Bayesian hierarchical inference for our ideal observer model.** On the left, each row is labeled as representing the population parameters priors ($p(\eta)$), the subject-level parameters priors ($p(\theta | \eta)$) or the likelihood ($p(D | \theta)$). On the right, each bubble represents a parameter of the model. The name of the parameter and its distribution are reported. The rectangular plates indicate that multiple iid variables have been grouped. The group name and the corresponding number of variables are indicated on the bottom left of each plate. Each arrow represents the dependencies of the model. Gray coloring indicates that the variable is conditioned on observations. 174

- A.4 **Prior predictive simulation of the 5γ 1σ model.** In black, the average proportion of answers "High" of the rats. Error bars denote the s.e.m. across all rats. Each colored line is the average predicted psychometric curve obtained for one of the five noise conditions, generated by parameters sampled from its respective prior distributions. The γ and σ parameters are drawn from the population average parameters; the ε parameters, given the absence of a hierarchical structure, are averaged across all rats. The shaded areas represent one (dark area) and two (dim area) standard deviations from the mean. The five curves all look almost identical (up to sampling noise) as the priors for the noise conditions were the same. 175
- A.5 **Posterior predictive simulation of the 5γ 1σ model.** In black, the average proportion of answers "High" of the rats. Error bars denote the s.e.m. across all rats. Each colored line is the average predicted psychometric curve obtained for one of the five noise conditions, generated by the parameters sampled in the MCMC chain, following the respective posterior distributions (same as the solid lines displayed in Fig. 4.5A). The γ and σ parameters are drawn from the population average parameters; the ε parameters, given the absence of a hierarchical structure, are averaged across all rats. The shaded areas represent one (dark area) and two (dim area) standard deviations from the mean of the posterior. 176
- A.6 **Posterior prediction of the linear relationship between sound intensity and perceptual gain factor in an ideal observer model.** The solid line represents the population-level gain parameter μ_γ , modeled as a linear function of the average sound pressure level I . The slope of the line is $(-8.6 \pm 1.4) \times 10^{-3}$, the intercept 1.636 ± 0.083 (posterior mean $\pm$ st. dev.). The color gradient along the line maps the specific sound conditions: green for the fixed amplitude condition, grays for the V+AM conditions, and purple for the V-only condition. The dashed lines represent one standard deviation of μ_γ . The shaded area around the regression line represents the posterior mean of the between-subject standard deviation, σ_γ . This area illustrates the estimated population-level variability. The symbols (triangle, square, cross) denote the relative average perceived sound pressure (x axis) and the posterior means (y axis) for individual rats. The inset shows the relative distance in Expected Log-Predictive Density (ELPD) between the linear model and our best model (5γ , 1σ), error bars denote the standard deviation of the difference. 177

List of tables

4.1	Two-way ANOVA table (A) and Tukey post-hoc tests (B) for the comparison among rat psychometric curves shown in Fig 4.3A. A repeated-measures two-way ANOVA was run on the proportion of “High TF” choices reported in Fig 4.3A, having as factors the experimental condition (“V + fixed amplitude sound”, “V + AM congruent” and “V + AM anti-congruent”) and the grating TF.	77
4.2	Two-way ANOVA table (A) and Tukey post-hoc tests (B) for the comparison among rat psychometric curves shown in Fig 4.3B. A repeated-measures two-way ANOVA was run on the proportion of “High TF” choices reported in Fig 4.3B, having as factors the experimental condition (“V + fixed amplitude sound”, “V + AM sound” and “V only”) and the grating TF.	78
A.1	Posterior summary statistics for the key population-level parameters of the AIC “computation” description. The columns represent the posterior mean (mean), posterior standard deviation (sd), the 3% and 97% boundaries of the Highest Density Interval (hdi_3% , hdi_97%), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (mcse_mean , mcse_sd) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (ess_bulk , ess_tail) (Vehtari et al., 2021), and the R-hat convergence diagnostic (r_hat) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019).	161
A.2	Posterior summary statistics for the key population-level parameters of the BIC “computation” description. Details are as in Table A.1.	162
A.3	Posterior summary statistics for the key population-level parameters of the AIC “independent” description. Details are as in Table A.1.	162
A.4	Posterior summary statistics for the key population-level parameters of the AIC “heuristic scaling” description. Details are as in Table A.1.	162

A.5	Posterior summary statistics for the key population-level parameters of the BIC “heuristic scaling” description. Details are as in Table A.1.	162
A.6	Posterior summary statistics for the key population-level parameters of the AIC “free scaling” description. Details are as in Table A.1.	163
A.7	Posterior summary statistics for the key population-level parameters of the BIC “free scaling” description. Details are as in Table A.1.	163
A.8	Posterior summary statistics for the key population-level parameters of the L-only description. Details are as in Table A.1.	163
A.9	Posterior summary statistics for the key population-level parameters of the W-I description. Details are as in Table A.1.	163
A.10	Posterior summary statistics for the key population-level parameters of the BIC “computation” description. The columns represent the posterior mean (mean), posterior standard deviation (sd), the 3% and 97% boundaries of the Highest Density Interval (hdi_3% , hdi_97%), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (mcse_mean , mcse_sd) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (ess_bulk , ess_tail) (Vehtari et al., 2021), and the R-hat convergence diagnostic (r_hat) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019).	169
A.11	Posterior summary statistics for the key population-level parameters of the “free penalty” description. Details are as in Table A.10.	169
A.12	Posterior summary statistics for the key population-level parameters of the “naive correction” description. Details are as in Table A.10.	170
A.13	Posterior summary statistics for the key population-level parameters of the “correction” description. Details are as in Table A.10.	170
A.14	Posterior summary statistics for the key population-level parameters of the “FIA” description. Details are as in Table A.10.	170

A.15	Posterior summary statistics for the key population-level parameters of the model. The columns represent the posterior mean (mean), posterior standard deviation (sd), the 3% and 97% boundaries of the Highest Density Interval (hdi_3% , hdi_97%), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (mcse_mean , mcse_sd) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (ess_bulk , ess_tail) (Vehtari et al., 2021), and the R-hat convergence diagnostic (r_hat) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019). Note that the μ_ϵ values are not population parameters in the same hierarchical sense as the others; instead, they represent the simple average of the individual rat ϵ values estimated, as the ϵ parameters were modeled without a group-level distribution.	178
A.16	Posterior summary statistics for the key population-level parameters of the model with binary predictors only on the LTSI population. The columns represent the posterior mean (mean), posterior standard deviation (sd), the 3% and 97% boundaries of the Highest Density Interval (hdi_3% , hdi_97%), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (mcse_mean , mcse_sd) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (ess_bulk , ess_tail) (Vehtari et al., 2021), and the R-hat convergence diagnostic (r_hat) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019).	182
A.17	Posterior summary statistics for the key population-level parameters of the model with binary predictors only on the THIN population. Details are as in Table A.16.	183
A.18	Posterior summary statistics for the key population-level parameters of the model with binary and continuous predictors only on the LTSI population. Details are as in Table A.16.	184
A.19	Posterior summary statistics for the key population-level parameters of the model with binary and continuous predictors only on the THIN population. Details are as in Table A.16.	185

Chapter 1

Introduction

1.1 Internal models of the brain

A model is a simplified representation of an idea, object, system or event that captures its essential properties and dynamics. It's not an exact replica of reality, but it reduces its description to the fundamental concepts necessary in a certain context. Already in the early period of philosophy, the fundamental human desire to explain the cosmos in a simple, unified way established the use of models as the primary conceptual tool for understanding the universe. This trend got stronger and more formal as the sciences started to develop. The role of model-making became so central that, in the words of John von Neumann, "sciences do not try to explain, they hardly even try to interpret, they mainly make models" ([Von Neumann, 1955](#)). Indeed, the true power of these models, from the hydrogen atom to complex global climate simulations, lies in their utility: they are indispensable for reducing complex phenomena to fundamental relationships, allowing us to interpret them, and most importantly to make predictions about future or untested scenarios.

The capacity of models to formulate predictions makes them crucial for more than just scientific applications. Any system interacting with its environment and trying to achieve or maintain a desired state, like keeping steady its internal temperature or moving an object, must possess an internal representation of the outside world ([Conant and Ross Ashby, 1970](#); [Francis and Wonham, 1976](#)). The system cannot rely solely on simple, instantaneous corrections to errors as they appear. Instead, it needs to incorporate the underlying dynamics of the external environment and the processes that generate external inputs. This internal model allows the system to generate a response that matches the structure of the incoming inputs, making it possible to cancel out persistent fluctuations rather than merely reducing them. In this sense, modeling the dynamics of the environment isn't just a helpful feature for

a system, but a necessary structural component for achieving precise, robust interaction with the world outside itself.

While this internal model principle (IMP) was first postulated for the design of artificial systems, it applies to biological organisms as well (Bin et al., 2022). Humans and other animals are prime examples of complex systems that must successfully navigate and regulate their interactions with a dynamic and often unpredictable environment. A huge range of tasks necessary for survival, from estimating the force needed to grasp a tool, to planning a navigational route, or anticipating the trajectory of a moving object, are impossible without the capacity to internally simulate the consequences of actions and the state of the environment. The IMP suggests that to achieve this, the brain must construct and maintain internal models of the world.

Indeed, even before the IMP formalized their importance, the notion that internal models explain the brain's computational power was already gaining traction (Craig, 1943). Ever since, a vast amount of evidence for such models has been found, spanning from low-level motor control to high-level abstract reasoning. Perhaps the most direct biological realization of the IMP is the concept of *forward models* in the motor system (Sperry, 1950; Welniarz et al., 2021). To overcome the significant delays present in sensory feedback, which can make reactions unstable and slow, the brain is thought to use a neural simulator. When a motor command is sent to the muscles, a copy of that command is processed by a forward model which predicts the future sensory consequences of the action. This prediction can be used to anticipate and correct errors before sensory feedback even arrives, enabling the fluid and rapid movements characteristic of skilled behavior. Moreover, by comparing the predicted sensation with the ones physically perceived, the brain can distinguish between those generated by the body itself and those caused by the external world, a mechanism believed to explain why it is impossible to tickle oneself (Blakemore et al., 2000).

In the broader context of acting on the external environment, *cognitive maps* are one of the earliest and most compelling pieces of evidence for the use of internal models in the brain. Indeed, simple stimulus-response or habit-learning theories can't account for complex behaviors observed in animals navigating an environment (typically a maze): rats, for example, proved to be able to find novel shortcuts to a reward when the learned path was blocked. This understanding of the spatial surroundings requires a map-like representation of the maze layout, a cognitive map (Tolman, 1948). This map is a model of the spatial environment, allowing the animal to make predictions about the consequences of its navigational choices. Plausible neural realizations of these maps have been found in the hippocampus and entorhinal cortex, with the discovery of *place cells* and *grid cells* (Hafting et al., 2005; O'Keefe and Dostrovsky, 1971). Moreover, the hippocampus has been shown to extend

beyond purely spatial information, encoding abstract variables like reward value, suggesting that cognitive maps may represent relational knowledge more broadly (Knudsen and Wallis, 2022; Viganò et al., 2021).

Evidence for the use of internal models has been found in the domain of decision making as well. Reinforcement Learning (RL), the framework where an agent learns through trial and error to select the optimal action for any given situation, is a powerful way to account for animal and human behavior (Dayan and Daw, 2008; Glimcher, 2011; Maia and Frank, 2011). Classical RL can accurately describe habitual, reflexive actions through association of particular states (of the subject and environment) to future rewards. However, in situations that require flexible and adaptive strategies, the behaviors observed in experiments can only be explained through the use of an internal model. These cases are well described by *model-based reinforcement learning*, where instead of learning associations of states and rewards, the subject learns an internal model of the task dynamics, encapsulating the probability of transitioning from one state to another and how rewards change over time. For example, rats will quickly stop pursuing a food rewarding action if they are later made sick from that same food, even without further trial-and-error: this flexible shift cannot be accounted for by purely habitual descriptions of behavior (Daw et al., 2005; Dickinson and Balleine, 2002).

This kind of models highlight a fundamental feature of the environment in which animals and humans operate: it is noisy, ambiguous and probabilistic. The same can be said about the perception mechanisms that inform the brain about that environment. For an agent to be successful, its internal models must therefore include a way to quantify this uncertainty. The Bayesian Brain hypothesis (Knill and Pouget, 2004) gives a possible answer to how the brain overcomes this challenge: it proposes that the nervous system handles uncertainty by representing information not as single, fixed values, but as probability distributions. A belief about any aspect of the world is captured by a distribution that reflects the range of possibilities and the likelihood of each of them. This framework recasts internal models as probabilistic models. Cognition, from perception to decision-making, can then be understood as a process of “inverting” these models, as they tell the brain how to link what is perceived with its most likely hidden causes, all while naturally and continually representing the uncertainty inherent in the task. To achieve this, the brain is believed to use Bayesian inference.

1.2 Bayesian models in cognition

Discovered in the mid 18th century by reverend Thomas Bayes, mysteriously abandoned and later rediscovered by Pierre Simon Laplace, Bayes’ theorem is one of the bases of modern

probability. It is a tool to update the degree of belief in a hypothesis in light of new evidence. It tells us that, upon observing some data, the probability assigned to a certain hypothesis should depend both on how likely the hypothesis is to generate that data and on how confident we were about the hypothesis to begin with. The rule is expressed mathematically as:

$$P(H | D) = \frac{P(D | H)P(H)}{P(D)} \quad (1.1)$$

Where $P(H | D)$ is the *posterior probability* of the hypothesis H given the data D , $P(D | H)$ is the *likelihood* of the data given the hypothesis, $P(H)$ is the *prior probability* of the hypothesis and $P(D)$ the *evidence* of the data.

The definition of probability in this framework is fundamentally different from the classical (or frequentist) notion. While the frequentist view defines probability as the long-run frequency of an event, the Bayesian definition uses it to quantify uncertainty regarding that event. This perspective is powerful because it allows us to assign probabilities to singular events that cannot be repeated, such as the probability that the orange spot you are seeing in the grass is a tiger. More importantly, it describes a subjective evaluation of belief. This makes it a natural framework for describing the internal representation of uncertainty in humans and animals.

Bayesian inference is the process of evaluating the degree of belief assigned to a particular hypothesis. Probabilistic models are the cornerstone of this process: they define the space of possible hypotheses (represented as parameters of the model) and specify how probable the observed data is under each of them. For example, to link the result of a coin toss to how fair the coin is, we use a Bernoulli distribution (the model) defined by a parameter p , representing the tendency of the coin to land on head. The model treats each value of p (e.g., $p = 0.5$ for fair) as a distinct hypothesis. Bayesian inference assigns a degree of belief to each hypothesis based on the results of the tosses and the initial prejudice on the coin fairness.

According to the Bayesian Brain Hypothesis, the brain uses the same machinery to make sense of the world: in a given context, it produces a probabilistic model of the world that defines all the possible hypotheses. Each of these hypotheses has a prior, that encodes the belief of the brain at a certain moment. In the example of the coin, the prior would represent how much we believe the coin to be fair before a toss. When new data is gathered (i.e., when the brain perceives a stimulus, as observing the result of the next toss), the model assigns to the data a likelihood for each of the hypotheses and updates its beliefs.

This deceptively simple scheme, the same used by statisticians and scientists, is capable of explaining many biases and behaviors in humans and animals. For example, it has been used to explain why owls underestimate the position of sound sources when they are not directly in front of their head ([Fischer and Peña, 2011](#)), or why low contrast stimuli appear

slower to humans (Weiss et al., 2002). It accounts very well for settings where different cues have to be combined to form a unified percept: when vision and touch provide conflicting information about object properties (e.g., size or texture), humans weight each modality according to its reliability, giving more weight to the more precise sense (Ernst and Banks (2002); see also Alais and Burr (2004); Helbig and Ernst (2008); Kording et al. (2004) for integration between other sensory modalities). Internal Bayesian models seem to be behind human understanding of other people's goals (Alanqary et al., 2021; Bhui and Gershman, 2020; Ying et al., 2023), the development of a theory of mind, i.e., the ability to understand that others have their own distinct mental states (Asakura and Inui, 2016), and even complex phenomena like skepticism can arise from optimal Bayesian inference (Alon et al., 2024), together with many other cognitive biases and abilities (Griffiths et al., 2024; Ma et al., 2023b).

1.3 Cognitive application of Bayesian methods¹

A simple inferential scheme, like the ones we just saw, is not always enough to make sense of the world. When different observations come from different sources, the natural behavior would be to use different inferences for each of them; however it can happen that the different sources describe the same phenomenon. For example, different cats have more or less similar shapes, so we could build a probabilistic model of how cats look and find the parameters of the model (their average length and how much it varies, their spectrum of eye-colors, and so on) applying Bayesian inference on every cat observed. In a similar way, one can build a model for dogs. However, cats and dogs, despite being different animals, have many matching features, as the number of legs or the position of the eyes or the presence of a tail. Ignoring these similarities, and considering cats and dogs as completely separate concepts, leads to a lot of redundancy in their models. Moreover, when we meet a new mammal, we will need to build an entirely new model for it, ignoring all the information we already have from cats and dogs. On the other hand, if we use the same model for both cats and dogs, we will be able to recognize their common features, and possibly also the features of new mammals; however, we will lose a lot of detail in the definition of each animal. The best solution is to structure the models in a hierarchy: there is a model, with its own parameters, for dogs, one for cats and so on, but all their parameters will be themselves collected in a "hyper-model", describing all mammals. So when we see a new cat, we update our belief of how cats look, but we also update a little bit our belief of how all mammals look. And when

¹or: a sneaky introduction to the statistical tools used in this thesis

we see a new mammal, we are able to understand quickly, more or less, how other animals of that species look, without the need of many observations.

This hierarchical organization is extremely powerful and useful in many practical and scientific applications (all of the statistical models used in this thesis, for example, follow the same structure: this allows us to describe at once the behavior of individual experiment participants and of their overall population). The use of similar structures has been observed in perception and cognition as well. Bayesian hierarchical models explain, among other things, how humans learn categories as those in the cats and dogs example ([Kemp et al., 2007](#)), how they make decisions in situations with little information ([Lee, 2006](#)), how they interpret complex moving scenes ([Bill et al., 2022](#); [Yang et al., 2021](#)) and how they infer the intentions of other people ([Diaconescu et al., 2014](#)).

Models like these, while practically advantageous, introduce a great deal of complexity, that translates to computational demands. Often it is not easy, if not outright impossible, to update beliefs using the exact Bayesian formula. In statistics, the solution is to rely on estimation methods, particularly those based on random sampling, also known as Monte Carlo methods. The fundamental advantage of these methods is that performing computations with samples from a probability distribution is simpler and quicker than using the distribution itself.

Imagine, for example, to read a text that is missing a word. Guessing that word with a Bayesian computation would require calculating the likelihood and prior probability of every single word in the vocabulary. A more practical approach is to use a method that only relies on a few words. One such method is a stochastic search: we start from a single, random word, and evaluate how well it fits (its likelihood) and how often it is used in general (its prior). We then consider a slightly different, similar word and decide if its combined fit is better than the first word. If the second word is a better option, or not much worse, we accept it and repeat the process by proposing a new word similar to it. Otherwise we discard it and try a different word similar to the original one. After a certain time repeating this process we will have a bunch of words, some considered many times, some just a few times and some one time only or immediately discarded. This will give us an idea of what the best word is for that context and of how certain we can be about it. It might also tell us if there are semantically different words that all have a good chance of being the right one. This kind of process gives us a very close approximation of the true posterior probability, i.e., the mathematically right degree of belief, for every word. The same fundamental logic applies in more general situations, even highly complex ones, which is why this method, called Markov Chain Monte Carlo (MCMC, [Gelman et al., 2014a](#)), is so powerful and widely used in statistical applications (including the analyses presented in this thesis).

This and similar sampling algorithms have natural psychological interpretations, offering insight into the mechanisms by which brains, despite their limited capacity, implement inference with complex internal world models (Griffiths et al., 2024). One instance, similar to the word example just discussed, is the Remote Associates Test (Mednick and Halpern, 1968): a task in which participants have to identify a target word related to three provided words. People's responses to the test are consistent with a stochastic search through a set of possible answers that matches MCMC (Bourgin et al., 2014; Smith et al., 2013). In a similar way, children seem to develop intuitive theories of the world by a process that resemble closely this kind of stochastic search algorithms (Ullman et al., 2012). Biases like perceptual multistability (the phenomenon where our perception of a stimulus spontaneously switches between different interpretations) have also been argued to originate from the same stochastic sampling process (Gershman et al., 2012). Related sampling models can provide an explanation for memory biases (Abbott and Griffiths, 2011), as the primacy effect (better memory for items at the start of a list) and the recency effect (better memory for items at the end of a list), in addition to various other cognitive capabilities (Chater et al., 2020; Daw and Courville, 2008; Lloyd et al., 2019; Sanborn and Chater, 2016; Zhu et al., 2020).

All these examples paint a clear picture: the brain bases many of its functions, from perception and cognition to its underlying computational strategies, on probabilistic models. Even when we might expect a deterministic view of reality, such as in the perception of physics, the brain appears to rely on Bayesian-like representations (Sanborn et al., 2013). Theories like the Free Energy Principle go even further, framing all of perception and action as a continuous process of testing and updating a probabilistic world model (Clark, 2013; Friston, 2010; Friston et al., 2006). This raises a fundamental question: if indeed the brain represent so many aspects of reality as probabilistic models, how does it know if it is using the "right" ones? And when multiple models describe equally well the sensory evidence, how does it choose among them?

1.4 Model Selection

We have seen how Bayesian inference allows humans to optimally integrate multiple sensory cues. The same inferential structure naturally extends to situations where there are multiple interpretations for the cause of the cues. For instance, when we see a mouth moving and simultaneously hear speech, the brain must choose between two competing hypotheses: a common source (one person talking) or separate sources (watching one person while hearing another) (Magnotti et al., 2013). This inferential framework, called Bayesian Causal Inference (BCI), is practically no different from the ones we have already seen: each possible

cause have a prior, and internal models tells us the probability of the sound and movements perceived in the case of a common source and in the case of separate ones. From these we can attribute to each cause a degree of belief. The difference from the instances seen before is that here the inference allows us not (only) to retrieve the content of the source (the words spoken), but to distinguish between two different internal models (common and separate causes) for the same perceptual stimuli.

BCI provides a powerful, principled explanation for a host of well-known perceptual effects that result from the brain's decision to either integrate or segregate conflicting sensory signals. In the example we presented above, it can explain why the same word is perceived as different sounds when shown together with different mouth movements (an illusion known as McGurk effect, see [McGurk and MacDonald, 1976](#)): if the brain decides the signals came from the same source, it "compromises" and integrates them, resulting in the illusion ([Lindborg and Andersen, 2021](#); [Magnotti et al., 2020](#)). Similarly, BCI explains how the brain combines sensory cues that may come from the same or different locations ([Körding et al., 2007](#); [Rohe and Noppeney, 2015](#)). It is able to explain why a smaller object with equal weight as a bigger one is perceived as heavier (size-weight illusion): competing hypotheses for the density relationship between the objects generate the illusion ([Peters et al., 2016](#)). A similar competition between internal models explains the rubber-hand illusion, where touch is felt on an artificial limb ([Samad et al., 2015](#)).

In all these examples, the Bayesian framework necessitates assigning prior probabilities to each possible explanation. However, the framework itself does not specify the source of these priors: a point often handled experimentally by either inferring subject-specific values or setting them based on qualitative assumptions. While the choice is certainly often guided by learned priors (e.g., our experience makes us expect that mouth movement directly causes the sound of words), the recurring characteristics of the models favored by the brain indicate the existence of underlying principles that actively constrain the selection among competing interpretations.

Indeed we know that, in general, the internal models utilized by the brain tend strictly toward simplicity. This preference for parsimony is observable across the hierarchy of neural processing, from low-level perception to high-level cognition ([Feldman, 2016](#)). For instance, the brain actively resolves the ambiguity inherent in visual input by seeking the simplest organizational structure. This involves grouping similar or proximal elements and recognizing consistent patterns, a phenomenon famously captured by the Gestalt laws of organization ([Hatfield and Epstein, 1985](#); [Koffka, 2013](#)). For instance, when the visual system encounters an object partially hidden by an occluder (such as a horse behind a tree) it "completes" the hidden geometry. We perceive a whole horse rather than two disjoint

horse-halves, despite the latter being perfectly consistent with the sensory data: interestingly, this also happens when the halves are far away, i.e., when the hypothesis of two horses should be stronger (Nanay, 2018). The brain rejects the complex, arbitrary juxtaposition of parts in favor of the simpler, unified object as it seeks to minimize the cost, in term of independent parts necessary to build the perceived image, of its sensory stimuli (Adelson et al., 1996; Van Lier et al., 1995).

This drive for simplicity extends well beyond vision into higher-order cognition and causal reasoning. One example is the ability to learn concepts, as it depends directly on their intrinsic complexity: to understand something we need to be able to extract a simplified abstraction from examples (Feldman, 2003). The same is true for memorization, as the simpler a concept is, the more it can be compressed and easily stored in memory (Ma et al., 2014).

An equivalent preference towards simplicity has been observed in the evaluation of causal explanations: the interpretations chosen by humans tend to be based on few causes, even when the likelihood of more complex explanation with multiple causes is higher (Lombrozo, 2007; Pacer and Lombrozo, 2017). This ubiquitous application of Occam's razor, the heuristic that "entities should not be multiplied beyond necessity", suggests that the brain faces a fundamental trade-off: it must select a model that explains the sensory data well (high likelihood) while simultaneously avoiding models that are unnecessarily complex. The same bias towards simplicity has been observed in young children (Bonawitz and Lombrozo, 2012), suggesting that the choice between causes is not simply guided by learned priors, but depends on intrinsic qualities of the explanations considered for the observations. These qualities, as simplicity, are not something we need to add on top of the probabilistic mechanisms we found in humans; rather, they are part of the statistical framework themselves (Wojtowicz and DeDeo, 2020). In fact, the same trade-off between simplicity and complexity found in humans arises naturally from Model Selection (MS), the statistical field that studies how to choose the best model to describe noisy data.

There are many frameworks in MS, all based on different principles, i.e., on different definition of what the "best" model is. The Minimum Description Length (MDL), for example, states that the best model provides the shortest possible description of the data while maintaining a short description itself (Grünwald, 2007; Rissanen, 1978). This principle is strongly reminiscent of the way the brain deals with concept learning and memorization, as we just saw. Bayesian Model Selection (BMS), instead, identifies the best model as the one with the highest posterior probability given the data, calculated using Bayes theorem (Jeffreys, 1935; MacKay, 1992); given the numerous instances of Bayesian inference in the brain, this framework also shows many parallels with the way the mind works. A third

approach is to assess how well a model will perform on future observation. Methods like Cross-Validation (CV) ([Arlot and Celisse, 2009](#); [Stone, 1974](#); [Vehtari et al., 2017](#)) implement this by directly testing the model's generalizability through repeated partitioning of the data into training and testing sets to estimate how well the model would perform on data it has not yet encountered.

All this and many other methods don't explicitly require to choose the simplest model. However, they naturally embody Occam's razor. In some cases this is evident from the structure of the method: to have a short description, a model needs to be simple, making the Minimum Description Length an expression of Occam's razor. In other cases this link is less clear, as for Bayesian Model Selection. BMS implies that every model has a fixed amount of probability (100%) to distribute among its predictions. A complex model is flexible: it is capable of explaining a vast number of distinct datasets. However, because it covers so much ground, it must spread its probability very thinly across all those possibilities. A simple model, by contrast, is rigid. It can only explain a narrow range of outcomes. Because it focuses on fewer possibilities, it assigns a much higher probability to those specific instances. Therefore, if the observed data falls within that narrow range, the simple model is the mathematical winner: it predicted the event with high confidence, whereas the complex model assigned to it a lower probability.

Cross Validation reaches the same preference for simplicity through the mechanism of overfitting. A complex model is highly flexible, allowing it to describe not just the general patterns in the training data, but also the random noise and accidental coincidences. While this leads to a high score on the training set, it causes the model to fail when faced with new, unseen data where those specific coincidences do not exist. A simple model, by contrast, is too rigid to capture the noise. It is forced to describe only the stable, underlying rules. Since these rules are consistent across both the training and the unseen data, the simple model makes better predictions on the rest of the data.

This link to Occam's razor can be made even more obvious by using approximations such as the Bayesian Information Criterion (BIC, [Schwarz, 1978](#)), the Akaike Information Criterion (AIC, [Akaike, 1998](#)), the Fisher Information Approximation (FIA, [Balasubramanian, 1997](#)) or one of the many other MS criteria ([Gelman et al., 2014b](#); [Konishi and Kitagawa, 2008](#)). In all these frameworks, derived from the same optimality principles we saw above, there is a clear and explicit balancing between a measure of goodness-of-fit and penalties for its complexity.

As we mentioned, when humans evaluate competing explanations, they show a very similar balance: while diagnosing diseases from some symptoms, for example, they consistently prefer a single-cause explanation over a multiple-cause one, even when base rates

make the complex scenario mathematically more likely ([Lombrozo, 2007](#)). In perceptual tasks, they tend to group different clusters of data under the same latent cause unless the distribution of the clusters doesn't diverge significantly, effectively minimizing the number of hypothesized categories ([Gershman and Niv, 2013](#)). Similarly, when performing "intuitive regression" to fit patterns, participants often reject complex polynomial functions in favor of linear ones, tolerating a significant amount of error to maintain structural simplicity ([Johnson et al., 2014](#)).

This parallelism with theoretical prescriptions makes it reasonable to use Model Selection as a normative framework to investigate the mechanisms behind human choices. This has been done by [Genewein and Braun \(2014\)](#), in an experiment where they asked participants to track hidden trajectories: the choices of the participants, overwhelmingly favoring a smooth line over a rough one, were studied using Bayesian Model Selection and the Minimum Description Length, showing that a balance between a theoretical definition of goodness-of-fit and of complexity penalty coincided well with the choices of the participants. Similarly, [Blanchard et al. \(2018\)](#) demonstrated that the human simplicity bias is not merely a preference for fewer parameters: similarly to BMS, it penalizes models based on their flexibility (i.e., the volume of the hypothesis space they could occupy) even when they have the same number of parameters. [Griffiths and Tenenbaum \(2005\)](#) analyzed how humans decide if some data (such as the activation of a gene in rats) is caused by another event (such as a treatment) or arises from random chance. They compared human evaluations to the predictions of Bayesian Model Selection, modeling the task as a structural decision between a graph with a causal link and a simpler graph without one. They showed that human judgments strongly resemble the theoretical framework: specifically, participants preferred the simpler "chance" model when the number of observations was low, even if the correlation between cause and effect was perfect, effectively imposing a penalty on the complex model that disappears only when sufficient data is available.

While previous studies established the general link between simplicity and Bayesian selection, [Piasini et al. \(2025\)](#) introduced a significantly more granular and quantitative resolution by decomposing the abstract notion of "complexity" into distinct model properties. Using the Fisher Information Approximation (FIA), an approximation of BMS, they mathematically separated the Bayesian penalty into four measurable components: dimensionality (number of parameters), volume (size of the hypothesis space), boundary (proximity of solutions to the edge of the parameter space), and robustness (sensitivity to small parameter changes/curvature). By testing participants on visual tasks designed to isolate each feature, they found that human decisions are quantitatively guided by penalties on all four geometric

properties, demonstrating a direct, quantitative relevance of theoretical notions of model complexity to human behavior.

1.5 Thesis Outline

In this thesis we expand this relatively young area of research to further investigate how and if the choice between explanations in humans parallels its theoretical counterparts. Despite the comparisons drawn between theory and decision-making in literature, in fact, it is not yet clear if this similitude spurs from the application of the same theoretical computations in the brain, or if instead cognition employs heuristics whose results mimic optimal ones. Moreover, we don't know if the brain employs a general definition of simplicity, as it happens in MS frameworks as AIC or BIC, or if it elastically fits its judgment of complexity to the different kinds of interpretations it considers.

Beyond the problem of model selection, we also apply the Bayesian modeling methods presented in this introduction to help explain biases in behavior and point toward neural causes for them. Finally, we apply these statistical tools to propose an improvement in the study of neural population activity.

Overall we explore the role of models in the brain on three different levels: how the brain chooses between models, how internal models can account for perceptual biases and what variables are encoded in the brain to construct an internal model.

The first two chapters investigate how human choose between multiple probabilistic models to explain noisy data. In Chapter 2 we explore how this choice is affected when changing the amount of data available. In theoretical frameworks, a smaller sample size leads to a stronger bias towards simpler models: while we observe the same qualitative effect in humans, the quantitative trend shows a clear divergence from optimality, hinting at the fact that known perceptual biases of numerosity perception condition the model selection process. In Chapter 3, following a similar approach to [Piasini et al. \(2025\)](#), we characterize how the brain evaluates the complexity of a model. We show that an apparent deviation from theoretical optimality is well explained by including in the definition of complexity characteristics unique to the model used, suggesting that the brain doesn't use a common definition of complexity for all kinds of model, as some frameworks of MS do.

In Chapter 4, we apply the Bayesian Inference framework to show how an internal model in rats can explain the influence of irrelevant sounds on a visual task. In the task, rats were trained to distinguish between stimuli with high and low visual frequency. The presence of unrelated noises seemed to affect their performance. We used a Bayesian model to show that

this effect is caused by a compression of the perceived frequencies, possibly induced by a direct inhibition of the auditory cortex on the visual cortex.

In Chapter 5, we leverage the statistical techniques established in previous chapters to develop a novel Bayesian method for analyzing neuronal population data. We use data from a value-based decision-making task where mice learn to push or pull a joystick based on the history of rewarded trials, a process that requires the animal to construct and update an internal model of reward contingency. We investigate the encoding of behavioral and contextual variables within two populations of inter-neurons. We report the initial, encouraging results from this method, which is currently in its early stages of development.

Chapter 2

Sample size dependence of Occam's razor in human decision-making

This chapter is a reproduction of the paper:

Rinaldi F. G., Piasini E. (2026) “Sample size dependence of Occam's razor in human decision-making”, biorXiv.

2.1 Introduction

To make sense of the world, humans and animals evolved to find interpretations or explanations to the sensory data they gather. These interpretations allow them to find meaning in what is perceived, and to predict what will be perceived in the future. This results in an understanding of reality structured by predictive models: models whose components allow to dissect and abstract the comprehension of the environment, and to generate new data to anticipate possible futures.

However, in a world of noisy data and noisier perceptions, often more than one interpretation is possible: many internal models could explain the same sensory stimulus. Solving this task is of primary importance for living organisms, as often the competing interpretations can lead to completely different actions (or inaction) and make the difference between death and survival.

Previous research highlighted that humans, when facing tasks that require to choose among different interpretations for some data, seem to adhere to Occam's razor principle: the idea that when more than one explanation fits similarly well a dataset, the simpler one is to be preferred ([Blanchard et al., 2018](#); [Gershman and Niv, 2013](#); [Griffiths and Tenenbaum, 2005](#); [Johnson et al., 2014](#); [Pacer and Lombrozo, 2017](#)). For example, when shown some dots sampled with a certain noise from an unknown curve, and asked to guess the shape

of the curve, participants tended to draw simple, smooth curves rather than squiggly ones (Genewein and Braun, 2014). Similarly, when asked to choose between explanations for an alien's disease, humans preferred those with few causes, even if a more complex, multi-cause alternative was significantly more probable (Lombrozo, 2007).

This simplicity bias is in general an effective strategy. In fact, Occam's razor principle guides most theoretical methods for finding the best model to describe some noisy data, a process known in statistics as Model Selection (Myung, 2000). In these frameworks, the principle takes shape as a balancing between a measure of goodness-of-fit and penalties for how complex the model is (Akaike, 1998; Balasubramanian, 1997; de Mulatier and Marsili, 2025; MacKay, 1992; Schwarz, 1978; Spiegelhalter et al., 2002; Watanabe, 2013). Humans seem to parallel this approach. When asked to choose between two competing interpretations to explain some noisy visual data, participants adopt strategies that reflect the same balancing, and, moreover, that align with theory driven notions of complexity (Piasini et al., 2025). This observed alignment with theoretical Model Selection is helpful to shed light on what kind of strategies humans intuitively employ when faced with this kind of problems. It is still unclear, however, if the alignment stems from the application of similar theoretical principles, or if humans learned heuristics that mimic the theoretical definitions of complexity while adopting a more elementary approach.

In theoretical frameworks the balancing emerges from the application of optimality principles. For example, minimizing the distance between the unknown true generating distribution $g(x)$ for some data X and a parametric distribution $f(x | \theta_M)$ from a candidate model M results in the well-known Akaike Information Criterion (AIC): $M_{best} = \max_M \{L_M - K_M\}$ where the maximum log-likelihood of the model, L_M , is the goodness-of-fit measure, and its number of parameters, K_M , is the complexity penalty (Akaike, 1998). Similarly, in the Bayesian Model Selection (BMS) framework the optimal model is the one with the highest posterior probability given the observed data: $P(M | X) \propto \int P(X | \vec{\theta}_M) P(\vec{\theta}_M) d\theta_M$, where $\vec{\theta}_M$ are the model parameters (MacKay, 2003). While not explicitly introducing any measure of complexity, for large N the logarithm of the integral can be approximated with an asymptotic expansion as a sum of terms, each growing slower than the previous as N increases. Keeping only the first two terms of the expansion, one obtains the Bayesian Information Criterion (BIC): $\log P(M | X) \simeq L_M - D_M + O(1)$, where we have again the log-likelihood term and a penalty known as Dimensionality, given by $D_M = \frac{\log(N)}{2} K_M$, with N number of data-points (Fig. 2.1a) (Schwarz, 1978). Continuing the expansion produces more penalty terms, corresponding to more nuanced notions of complexity that can be interpreted with the tools of information geometry (Balasubramanian, 1997; MacKay, 2003; Piasini et al., 2025).

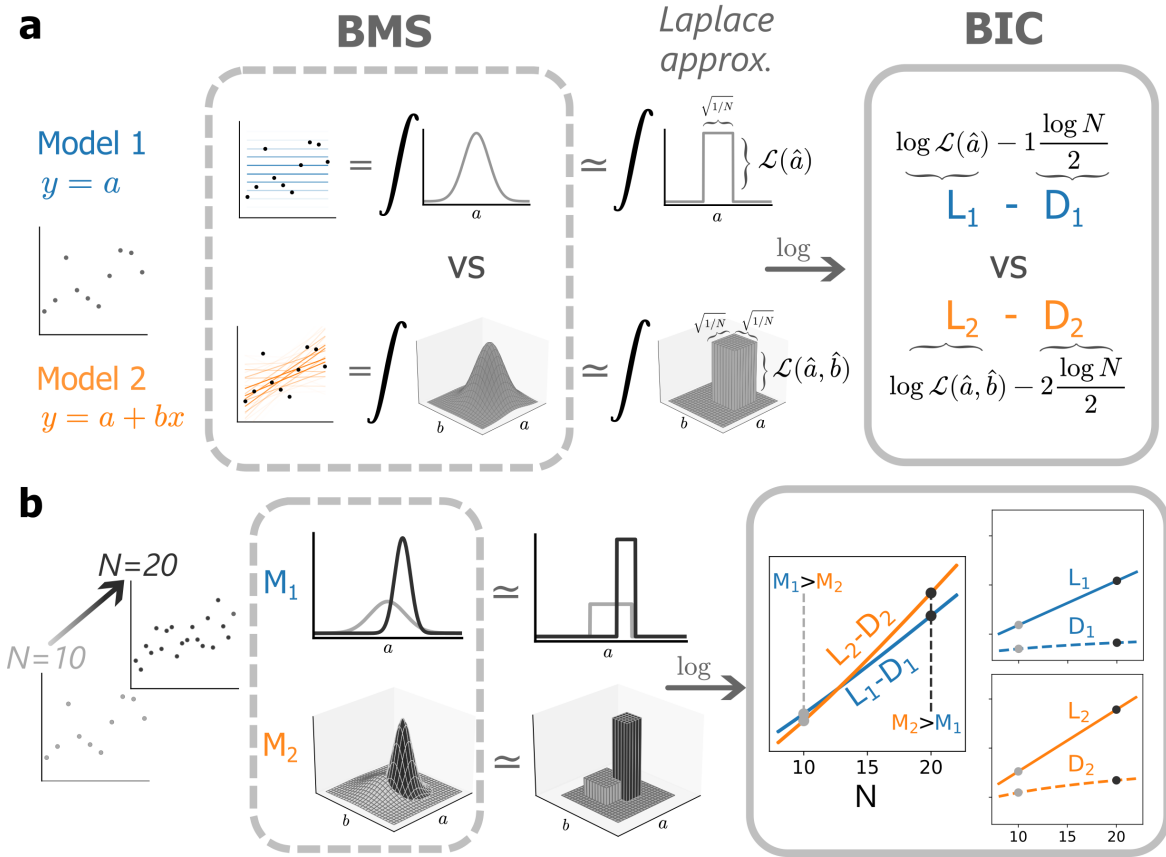


Fig. 2.1 Balancing of goodness-of-fit and complexity in Bayesian Model Selection. **a.** Bayesian Model Selection (BMS) between a model with one parameter (Model 1, linear model with fixed slope = 0, blue) and a model with two parameters (Model 2, linear model, orange). Given the same prior for both models and uniform prior across their parameters, the best model will be the one with the higher likelihood averaged across all the possible choices of parameters (dashed box, left). This quantity, the marginal likelihood, corresponds to the integral of the likelihood times the prior (dashed box, right). By using Laplace approximation, this integral can be decomposed in the product of a “height” term (proportional to the maximum likelihood of the model) and “width” terms (proportional to the inverse square root of the number of data-points N). In the logarithmic space, these terms correspond to the log-likelihood L_M and the dimensionality D_M of the Bayesian Information Criterion (BIC, solid-lined box). **b.** Effect of increasing the sample size. The height of the likelihood landscape is exponential in the number of points N (since it is the product of the single-point likelihood of N data-points), while the widths of the peaks decrease as $N^{-1/2}$. In the logarithmic space, this translates to a log-likelihood term that grows linearly in N and a dimensionality that grows logarithmically. Since the goodness-of-fit grows faster than the complexity penalty, the simplicity bias gets relatively weaker for higher sample sizes, allowing for the selection of more complex models (solid-lined box).

In all these frameworks, the strength of the simplicity bias changes when changing the amount of data available for the inference. For independently drawn observations, the likelihood is the product of the probability of each data-point, and thus the log-likelihood will be the sum of their log-probabilities. L_M , therefore, grows on average linearly with the number of observations N . On the other hand the complexity penalties are either independent or sublinearly dependent on the quantity of observed data. The result is that the importance of complexity relative to goodness of fit becomes smaller as the amount of available data increases. This is a major component in every implementation of Occam's razor: when a larger sample is available, a rational decision-maker can afford the use of more complex models (Fig. 2.1b).

As we have seen, the presence of a simplicity bias in humans has been established in previous work, and its strength has been quantitatively measured. However, it is not known how this strength and its importance relative to goodness-of-fit depend on the amount of data that is available to human observers, as only few and conflicting studies touched on the matter (Griffiths and Tenenbaum, 2005; White, 2003). Moreover, if such a scaling exists, it is not known whether it is a side effect of an approximate computation, such as a Bayesian integration, or whether it emerges from an explicit heuristic mandating that complexity should become less and less important for larger sample sizes. In the latter case, one would expect that the scaling would be affected by cognitive processes and distortions related to numerosity estimation (Burr et al., 2018; Cheyette and Piantadosi, 2020; Krueger, 1972; Merten and Nieder, 2009); while if it is the outcome of an approach based on direct computations, we expect this not to happen.

Here we investigate the existence and nature of this scaling. To do so, we focus on AIC and BIC as normative frameworks to describe human behavior in a visual model selection task. The complexity penalties of these frameworks, K_M and D_M respectively, depend in a simple and well understood way on the sample size N (Fig. 2.1). Moreover, dimensionality D_M was shown to be the most important term in shaping human strategies, making it an ideal factor for this study (Piasini et al., 2025). If humans intuitively perform a computation based on optimality principles (for instance, if they perform an approximate Bayesian integration), we expect their behavior to be best described by a linearly growing goodness-of-fit counterbalancing a constant (K_M) or sublinearly growing (D_M) penalty, as seen in the theoretical frameworks. If instead there is no scaling in N , and human choices are only guided by an empirical estimate of the true distribution of sampled data, we expect the best description to have a fixed penalty and a goodness-of-fit independent on N , such as the divergence between the empirical measure of the data and a candidate model distribution. Finally, if human strategies are based on an heuristic that explicitly requires a scaling based

on the estimated number of data-points, we would expect the growth of both the goodness-of-fit and the penalty to be distorted compared to their theoretical counterparts, consistently with the well known sublinear effects that can be observed in human numerosity perception (Anobile et al., 2015; Burr et al., 2018; Fechner, 1860; Merten and Nieder, 2009).

Our results show that human behavior is best described by a complexity penalty whose importance relative to the goodness-of-fit grows sublinearly as the sample size N grows. By using Gaussian Processes to estimate the precise shape of the scaling, we show that it closely matches a logarithmic trend, consistently with the prescription of Weber’s law (Fechner, 1860), up to high values ($\simeq 50$) of N , where a transition happens. This trend could match the transition observed in visual numerosity perception between a regime of direct number estimation and a regime of estimation from density (Anobile et al., 2015; Burr et al., 2018), reinforcing the idea that the scaling stems from a heuristic-mandated weighting based on estimated sample size.

2.2 Results

We designed a simple decision-making task to investigate the strategies of intuitive model selection employed by human participants. Participants were shown noisy visual data and had to guess which one among a set of possible models generated it. To test a wide range of dimensionalities D_M , we used a set of 5 models, each with a different number of parameters. Each model was a mixture of K one-dimensional Gaussian components, with $K \in [1, 5]$ (Fig 2.2a, down). The standard deviation of all components was fixed ($\sigma = 0.1$), and within a model all components had the same weight $\pi = \frac{1}{K}$, thus limiting the parameters of the model to the K means of its components. These constraints allowed us to make the task more manageable for the participants, as well as focus on a single kind of parameter, avoiding the possibility that different kinds of parameters might have a different influence on human strategies. On each trial, data was generated by one of the models, selected at random with uniform probability. The positions of the K means were randomly generated on a horizontal line $[-1, 1]$ with uniform probability, and N observations were sampled from the resulting distribution. These observations were represented as white vertical lines on a gray background. To avoid underestimation of the sample size due to overlapping lines, each line was marked by a white circle (Fig 2.2a, top). To study the effect of the amount of available data we showed trials with ten different sample sizes N , equally spaced between 16 and 70. The amount of trials generated by each model for each value of N was balanced, to ensure a uniform exposure to all the models. Overall, the experiment consisted of 900 trials, shown in

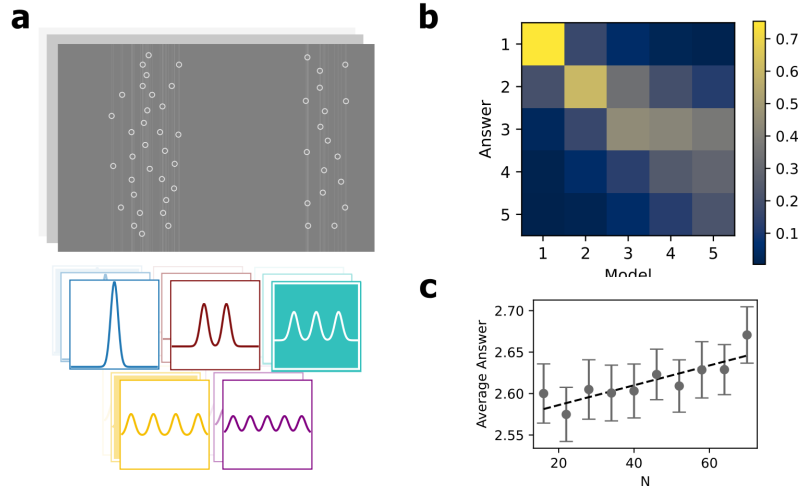


Fig. 2.2 Task, models and participants' answers. **a.** The five models that participants had to choose among (bottom) and an example of a trial from the task (top). The trial shown was generated by a model with three components; two had close means and generated the cluster on the left, one had an isolated mean and generated the cluster on the right. **b.** Average confusion matrix across participants and trials: while for models with low dimensionality ($K_M = 1, 2$) the answers mostly correspond to the true model, when the number of parameters increases, the answers are skewed towards values lower than the true ones, showing the effect of the simplicity bias. **c.** Average answer across all participants and all trials with N data-points. The values consistently fall below the true mean $\langle K_M \rangle = 3$, once again indicating the presence of a simplicity bias. However, the upward trend suggests a gradual weakening of the bias as N increases, as quantified by the Weighted Least Squares regression (dashed line; slope: 0.0012, one-sided p-value 0.04), supporting the hypothesis of a scaling in N .

blocks of 30. Within a block, all trials had the same sample size. We collected data from 100 participants (see Methods for more details).

An initial analysis of the data revealed that, on average, task performance was lower on trials where the true value of K was higher, and participants' error patterns were skewed towards reporting a value of K lower than the true one (Figure 2.2b). Moreover, the average value of K reported by participants increased as a function of N (Figure 2.2c). Overall, these patterns suggest that participants exhibit a simplicity bias (preferring models with lower K), and that this bias becomes weaker as the amount of available data increases.

To test these hypotheses further, we turned to a statistical analysis pipeline that we designed and preregistered before collecting the data (Rinaldi, 2025). The pipeline uses a Bayesian multinomial logistic regression (Gelman et al., 2014a) to measure the degree to which each participant's choices were affected by the goodness-of-fit and the complexity penalty of models (Piasini et al., 2025). Specifically, we assigned a measure of sensitivity β to each of the terms: the probability to choose a specific model was therefore dependent on

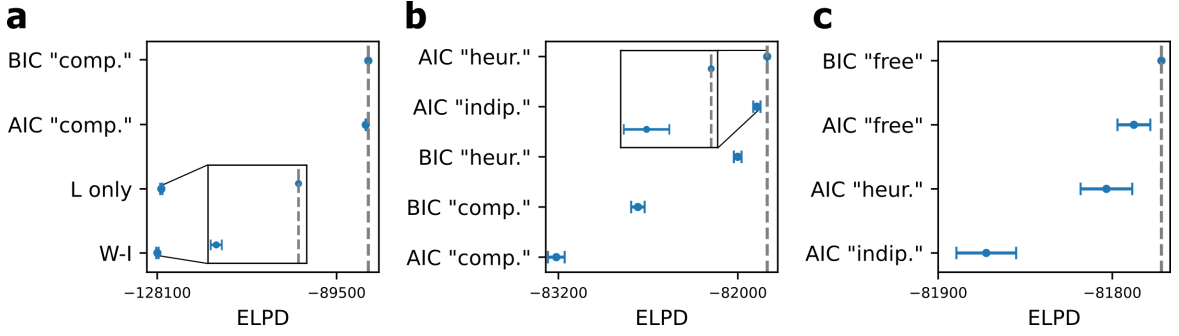


Fig. 2.3 Model comparison between candidate descriptions. **a.** suPareto-Smoothed-Important-Sampling Leave-One-Out Cross-Validation (PSIS LOO CV) comparison between the two Goodness-of-fit descriptions (W-I and Likelihood-only) and two descriptions based on AIC and BIC. The x-axis shows the Expected Log Predictive Density (ELPD) of each description. The errorbars represent the standard error of the ELPD differences with the best model, $ELPD_{\text{best}} - ELPD_M$. The inset shows the PSIS LOO comparison between the Likelihood-only and the W-I descriptions, showing that the Likelihood-only description has a significantly higher ELPD than the W-I one. **b.** ELPD comparison between the “computation”, “independent” and “heuristic scaling” description. The “computation” ones perform significantly worse than the others. The inset highlights that the difference between the AIC “heuristic” and AIC “independent” descriptions is within three standard deviations. **c.** ELPD comparison between the AIC “heuristic”, AIC “independent” and “free scaling” descriptions. The “independent” description is the only one to perform significantly worse than the others.

the sum of the product of each term (goodness-of-fit or penalty) with its sensitivity parameter:

$$p(\text{choice} = M \mid X) \propto \exp(\beta_G G_M + \beta_P P_M)$$

Where G_M is the goodness-of-fit of the model M for the data X and P_M is its complexity penalty.

We first verified whether the behavior of participants displays a simplicity bias in our setup. We compared regressions based on AIC-like or BIC-like decision rules ($G_M = L_M$ and respectively $P_M = K_M$, the number of parameters, or $P_M = D_M$, the dimensionality) with a regression that only uses the log-likelihood L_M , without any penalties. The comparison was performed using the Pareto smoothed importance sampling Leave-One-Out Cross-Validation (PSIS-LOO CV, Vehtari et al., 2024). This cross-validation method allows us to approximate the Expected Log Predictive Density (ELPD) of a fitted regression, a statistical measure of its predictive power on data it wasn’t trained on (Vehtari et al., 2017). As shown in Fig. 2.3a, AIC and BIC greatly outperform the predictive power of a likelihood-only regression. This confirms that our participants follow Occam’s razor while performing the task.

Moreover, we found that the inferred sensitivity to dimensionality D_M in the BIC-like description (and similarly the sensitivity to K_M in the AIC-like one) is greater than the sensitivity to likelihood, as well as the theoretically optimal value of 1 (See Tables A.1 and A.2). This over-weighting of dimensionality has also been observed by Piasini et al. (2025)¹.

In a similar manner to the previous comparison, we tested whether a more elementary, intuitive measure of goodness-of-fit would better align to human strategies. We compared the previous regressions with a decision rule based on a set of heuristics tailored to the nature of the sensory stimuli. Since the models that our participants compare generate clusters of points, we used as features the cluster width (W) and the imbalance of points between clusters (I): a good model would be able to partition the data into similarly wide and similarly populated clusters. Each term had its own sensitivity parameter β , allowing the regression to flexibly find the right weights for the two goodness-of-fit metrics. The ELPD of the resulting regression (W-I regression) was lower not only than the ones incorporating a complexity penalty (Fig. 2.3a), but also of the likelihood-only inference (Fig. 2.3a, inset), proving that, compared to these heuristic features, L_M captures more of the information used by participants to drive their choices.

We therefore proceeded to study whether and how goodness-of-fit and complexity scale with sample size. We developed a total of five regression variants for AIC and BIC-based terms. The variants are divided into three kinds (“computation”, “independent”, and “heuristic scaling”), one for each possible behavior of the scaling (Fig. 2.4). The two “computation” variants (AIC_{comp} and BIC_{comp}) represent decision strategies based on a direct computation, similar to the ones that lead to the AIC and BIC criteria and that could for instance be performed through probabilistic computations implemented at the neuronal level (Koblinger et al.; Ma and Jazayeri). These variants make use of the theoretical values for the goodness-of-fit and complexity penalty (Fig. 2.4, left). These are the variants used in the previous comparison. The “independent” variant (AIC_{ind}) matches a strategy that doesn’t depend on the number of points, but only on their density. Both the goodness-of-fit and the penalty are independent of the sample size. To achieve it, we substitute the log-likelihood L_M with the average point-wise log-likelihood $l_M = \frac{L_M}{N}$ (Fig. 2.4, center). This quantity corresponds to the Kullback-Leibler divergence between the empirical measure of the data and the candidate model distribution, up to an additive term (the entropy of the empirical measure of the data, which doesn’t change between models and thus doesn’t affect the probability $P(\text{choice} = M | X)$). In this case, we only consider one variant in common for AIC and BIC, since, if we remove the dependence on N of the BIC dimensionality D_M , it will be the same

¹In Piasini et al. (2025), the dimensionality is computed slightly differently, as it is: $D_M = K_M \log(\frac{N}{2\pi})/2$. However, the over-weighting of dimensionality persists even when adjusting for this difference, i.e., when multiplying the sensitivity to dimensionality β_D estimated in (Piasini et al., 2025) by $\log(\frac{N}{2\pi})/\log(N)$.

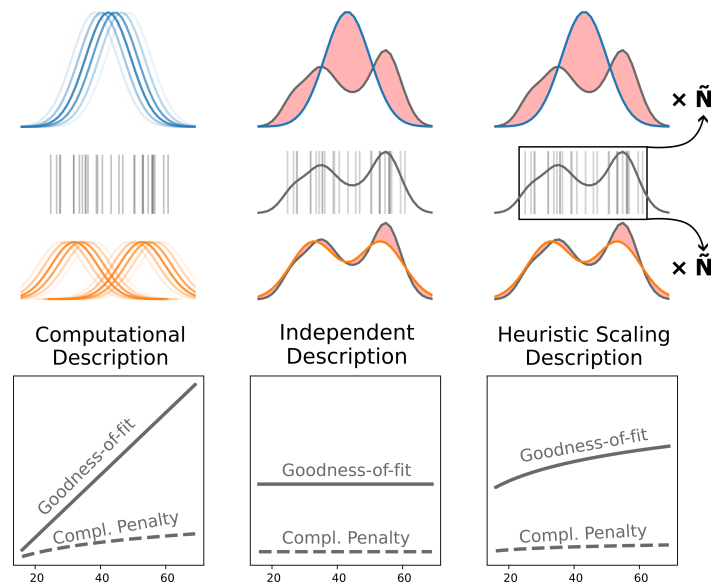


Fig. 2.4 **Graphical representation of the three descriptions we used to model the participants' strategies and dependence of their terms on the sample size N .** Only the BIC versions are shown. The “computation” description (top left) assumes that participants' choices depend on a computation like Bayesian integration: as shown in Fig 2.1a, the integration sums the likelihoods for each possible parameter value, here represented as different values of the components' means. The resulting goodness-of-fit scales linearly in N , while the resulting complexity penalty scales logarithmically (bottom left). The “independent” description (top center) assumes that participants only rely on the density of the data (gray line) and compare it with the best-fitting shape of the models. In the figure, the red areas represent the difference between the empirical data distribution and the model distributions. Both the goodness-of-fit and the dimensionality are constant in N (bottom center). The “heuristic scaling” description (top right) assumes that participants rely on a process similar to that in the “independent” description, but try to recover the optimal scaling by weighting the result by the perceived number of data-points $\tilde{N}$. The resulting goodness-of-fit scales as the perceived numerosity of the data, which we assume to be $\tilde{N} = \log(N)$ (Fechner, 1860), while the penalty scales as $\log(\tilde{N})$.

as the AIC penalty K_M up to a multiplicative factor. Finally, to check whether cognitive distortions affect human strategies, in the two “heuristic scaling” variants (AIC_{heur} and BIC_{heur}) we substitute the dependence on N in each term with a dependence on $\tilde{N} = \log N$. This compressed dependence represents the effect of a sublinear Weber-like perception of numerosity (Merten and Nieder, 2009). Therefore, the goodness-of-fit term will be $\tilde{N}l_M$ and the BIC dimensionality penalty $\log(\tilde{N})\frac{K_M}{2}$, while the AIC penalty remains unchanged (Fig. 2.4, right). The details of all these variants, with the exception of the “independent” one, were specified in our preregistration document (Rinaldi, 2025).

We used LOO-CV again to perform the comparison (Fig 2.3b). The “computation” variants are outperformed by all other variants. This suggests that participants don’t follow a decision rule based on theory-like computations, such as (approximate) Bayesian integration. However, it is harder to adjudicate between the other hypotheses and to draw a definite conclusion on the scaling of the simplicity bias with the sample size N . The AIC “heuristic” variant, where the importance of goodness-of-fit scales as $\tilde{N} = \log(N)$, is the one with the highest predictive power, but it is closely followed by the AIC “independent” variant, where there is no scaling at all ($\Delta ELPD = 68.99 \pm 24.32$; Fig 2.3b, inset).

To better understand the scaling trend, we developed two more variants for AIC and BIC, the “free scaling” variants (AIC_{free} and BIC_{free}). These variants are operationally the same as the “heuristic scaling” variants, but instead of substituting the dependence on N of the terms with a dependence on $\tilde{N} = \log N$, we use a generic function $\tilde{N} = f(N)$. The function is implemented as a Gaussian Process (GP) (MacKay et al., 1998), inferred together with the other parameters of the regression. This allows us to reconstruct the actual shape of the scaling, and thus to distinguish between a flat dependence (as in the “independent” variant) or a sublinear one (as in the “heuristic scaling” variant). The inferred curves for the “free scaling” variants are shown in Fig 2.5a. They both follow a sublinear trend up to $N \simeq 50$, where the $\tilde{N}$ of AIC_{free} flattens and $\tilde{N}$ BIC_{free} starts to decrease. These trends suggest that participants follow a “heuristic scaling” type of behavior up to high values of N , where the strategy seems to change.

A possible explanation for this transition, at least for the $\tilde{N}$ of AIC_{free} , is a strategy change from a numerosity-based heuristic to a density-based one, and therefore from a strategy dependent on the number of data-points to an independent one. This kind of change is reminiscent of the observed switch from Weber-like numerosity perception to a density-based estimation for large numbers of data-points (Fig 2.5b, Anobile et al., 2015; Burr et al., 2018). However, the specific value (or range of values) of N where this switch occurs is dependent on the details of the experimental setup: in Anobile et al. (2015), the transition threshold depended on the eccentricity of the cluster of data-points or on the distance between clusters.

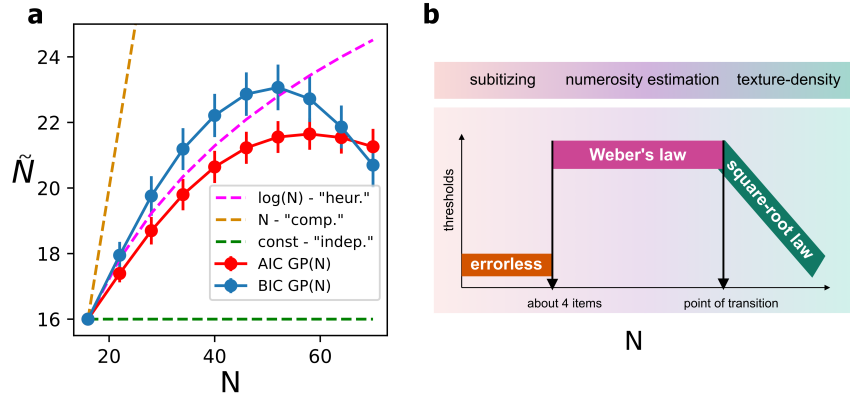


Fig. 2.5 **Scaling of the perceived sample size.** **a.** The inferred functions $\tilde{N} = GP(N)$, computed for both the AIC (red) and BIC (blue) “free scaling” descriptions, compared with the linear scaling of the “computation” descriptions (gray), the flat scaling of the “independent” description (green) and the logarithmic scaling of the “heuristic scaling” descriptions (pink). Both the AIC and the BIC functions follow a logarithmic trend for low sample sizes, and have a transition around $N = 50$. **b.** The three stages of numerosity perception: errorless (orange), Weber-like (pink), and density-based (green). Image adapted from [Burr et al. \(2018\)](#).

In our setup, not only we don’t control for the distance of the stimulus from the center of vision, as the participants are free to move their gaze, but different trials display clusters in different positions and often more than two clusters are presented at once, making it difficult to link in a more quantitative way the two perceptual phenomena. Moreover, our analysis doesn’t allow us to infer the Weber’s fractions and compare them to the literature, as their value is impossible to disentangle from that of the sensitivity to goodness-of-fit β_L .

2.3 Discussion

This study investigated how sample size influences the human decision-making process when choosing an explanation for noisy data. In particular, we studied how the amount of available data influences the trade-off between favoring well-fitting explanations and penalizing their complexity. While previous research has established that humans have a simplicity bias, aligning with the principles of Occam’s razor, it remained unknown whether this bias adapts to the sample size (N) as prescribed by normative statistical theories ([Akaike, 1998](#); [Schwarz, 1978](#)). Our results confirm previous findings about the presence of this simplicity bias ([Piasini et al., 2025](#)). Moreover, we demonstrate that human choice strategies do change with the number of data points, but not following the linear scaling of the relative importance

of the goodness-of-fit that would be expected from an approximate Bayesian computation or other optimal statistical methods. Instead, our results strongly support the existence of a sub-linear scaling, where the influence of goodness-of-fit grows more slowly than the amount of data. The shape of this sublinear scaling (logarithmic for smaller sample sizes and flattening for larger ones) parallels the biases found in the perception of numerosity (Anobile et al., 2015; Burr et al., 2018). This suggests that an estimation of the quantity of data points might be used by participants to heuristically guide the balance between goodness-of-fit and complexity.

Inference frameworks such as Bayesian Model Selection provide a principled reason for why the penalty for complexity should become less important as more data becomes available (Balasubramanian, 1997; MacKay, 2003). The maximum log-likelihood, representing goodness-of-fit, grows linearly in N for independent data-points, while complexity penalties grow much slower (as in BIC) or not at all (as in AIC). This change of balance naturally arises from the application of optimality principles and ensures that with enough evidence, a more complex model can be justifiably chosen. The absence of a similar scaling in humans suggests a fundamentally different, and potentially more resource-rational, computational strategy.

Our results are consistent with strategies that don’t compute goodness-of-fit through an explicit calculation of the maximum likelihood, by summing the log-likelihood of individual points. Human behavior is instead well described by a process that incorporates a global evaluation of goodness-of-fit for the whole dataset. This evaluation could, for instance, be based on the difference between the data density and the model shape, which could make it more computationally efficient. At the same time, our participants implicitly assigned a weight, or a measure of trustworthiness, to this quantity based on the perceived sample size of the data. This interpretation is in line with numerous instances where the brain has been shown to extract population statistics from groups of objects, and to be more reliable in these group estimates compared to single object estimates (Cepeda et al., 2025; Whitney and Yamanashi Leib, 2018).

Our results don’t allow us to tell with a high level of confidence whether the same weighting happens for the complexity penalty. Formally, the model that best describes our data is the “free scaling” BIC, where the complexity penalty scales logarithmically in the perceived number of datapoints $\tilde{N}$. However, this model is closely followed in predictive power by the “free scaling” AIC, where the complexity penalty does not depend on $\tilde{N}$ at all (only the goodness-of-fit does). To some extent, it is unsurprising that it is harder to assess the dependence of the penalty (that is, to differentiate AIC from BIC) than that of the goodness-of-fit (differentiating, for instance, the distinct variants of AIC), because the

logarithmic dependence of the BIC complexity on $\tilde{N}$ compresses the range over which it varies, and consequently its sensitivity to the specific values taken by $\tilde{N}$. However, it is clear from all of our penalty-incorporating descriptions that the weight of the penalty is higher than what would be expected from the normative theory. In agreement with previous findings on a different task (Piasini et al., 2025), our participants are therefore more skeptical of complex models than an optimal strategy would recommend.

The main goal in our experimental design was to allow for a quantitative characterization of the way in which sample size affects intuitive model selection. This entailed some necessary limitations for the present study. First, our experiment uses simple, abstract one-dimensional stimuli, focusing exclusively on model dimensionality as complexity penalty. As shown in (Piasini et al., 2025), more nuanced model features related to complexity can have a non-negligible effect on human choices. We don't know what form of sample size scaling, if any, might affect these features. It also remains an open question whether the specific sublinear scaling we observed applies to different or more naturalistic choices of noisy data, and more generally to other sensory modalities. Second, our task was limited to low-dimensional models ($1 \leq K \leq 5$), motivated by the fact that many interesting perceptual and cognitive phenomena can be effectively conceptualized and studied empirically using relatively simple parametric descriptions of the sensory data (De Valois and De Valois; Geisler, 2011; Gold and Shadlen; Green and Swets; Knill and Richards; Ma et al., 2023b; Whitney and Yamanashi Leib, 2018). While we expect that the qualitative features of our conclusions (the amount of available data affects intuitive model selection; its impact follows a sublinear trend) to have very broad applicability, and their quantitative details to generalize to tasks involving models of moderate dimensionality, making quantitative predictions for settings involving very high dimensional models (such as those commonly used in machine learning) is a harder task. In fact, the theoretical frameworks needed to characterize the complexity of very high dimensional models may require a more powerful set of analytical tools than the ones we incorporated in our analyses (see for instance (Chen et al.; Karakida et al.; Sun and Nielsen)). Establishing whether this regime is relevant for intuitive model selection, and how to tackle it with our approach, is an intriguing direction for future research.

A different type of limitation is that, while our data supports a heuristic-based origin for the logarithmic scaling, it cannot entirely rule out an alternative “computation” description. It could be argued that participants may actually carry out an approximate form of Bayesian integration, but the likelihood is computed on a subset of the available data-points obtained through a sub-sampling process, possibly guided by attentional or processing bottlenecks. If the size of this subset scales logarithmically with the total sample size, it could produce (depending on which points are sampled) the same sub-linear growth in the goodness-of-fit

term we observed. In this view, numerosity perception and likelihood calculation would not be feeding one into the other, but would rather both be outcomes of the same underlying attentional sampling mechanism, thus displaying similar scaling. This interpretation, while conceivable, posits a complex integration process on top of a specific, arbitrarily-defined sampling step, and is therefore less parsimonious than the one we choose to put forward. The heuristic description we used, where a global perceptual estimate directly modulates the trade-off, provides a simpler and more direct cognitive explanation that also naturally explains the scaling shift observed at high N .

In conclusion, this study provides the first quantitative characterization of how the human simplicity bias scales with the amount of available evidence. The sublinear scaling we found represents a qualitative departure from the linear scaling prescribed by normative statistical frameworks such as AIC and BIC. Our results demonstrate that this sub-linear trend is not arbitrary but closely follows the logarithmic compression characteristic of numerosity perception. We therefore conclude that human model selection is not guided by a formal statistical computation — not even in an approximate sense — but by a resource-rational heuristic that repurposes the brain's native perceptual machinery to solve abstract inferential problems.

2.4 Methods

2.4.1 Behavioral experiment

In our preregistered behavioral task, participants were shown a certain number N of white vertical lines with varying horizontal positions on a gray background. As a visual aid, each line was paired with a small circle, in order to help the participants discern overlapping lines: the horizontal position of the circle corresponding to that of the paired line, while the vertical position was a random coordinate on the line sampled in such a way as to not overlap with the circles of other nearby lines (Fig. 2.2a, top).

On each trial, the horizontal position $x \in [-1, 1]$ of the vertical lines was generated according to the following rules:

1. A random number $K \in [1, 5]$ was sampled: this was the number of latent sources that the data would be generated from;
2. The position μ_i of each of the K latent sources was sampled from a uniform distribution in $[-1, 1]$;

3. The position of each vertical line was sampled from the resulting gaussian mixture distribution $p(x) = \sum_{i=1}^K \frac{1}{K} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu_i)^2}{2\sigma^2}$, where $\sigma = 0.1$ was fixed (Fig. 2.2a, bottom).

Therefore, on each trial there were K clusters of vertical lines, sometimes with a certain degree of overlap between clusters. The participants had to report the number of latent sources, i.e., identify the model generating the data. They did so by pressing on their keyboard “W” to report 1, “D” to report 2, “K” to report 3, “O” to report 4, and Space to report 5. The key-to-number mapping was designed to allow a comfortable positioning of the hands on the keyboard (assuming a standard QWERTY layout) and hence avoid biases due to uncomfortably placed keys. To avoid confusion, this mapping was constantly indicated on the screen, the number selected by the participant was highlighted upon selection, and participants had the possibility to correct mistakes by going back to the previous trial at any moment by pressing “B”.

The participants were instructed, during a brief introductory tutorial, on the properties of the Gaussian mixture models we used. In particular, they were shown different examples of data generated by the GMMs to illustrate the fact that each component had the same variance (and thus would generate clusters of approximately a fixed width) and that components of the same model had the same weight (and thus would generate clusters of approximately the same size). To better help them internalize these concepts, participants were told that:

- the vertical lines were trails left on the ground by N wolves belonging to different packs: each area (i.e., each trial shown) could host from one to five packs;
- wolves tend to walk roughly in a file, so while walking each pack would leave a trail of approximately the same width, independently of the number of wolves in the pack;
- if more than one pack roamed an area, the sizes of the packs would approximately be the same.

Each participant was tested on 10 different values of N (number of vertical lines displayed): 16, 22, 28, 34, 40, 46, 52, 58, 64, and 70. For each value of N , participants were shown 90 trials. Unbeknownst to the participants, each trial was repeated 3 times in different moments of the experiment, in order to account for decision noise and verify participant consistency. Therefore, there were 30 distinct trials for each value of N , equally divided among the 5 values of K , for a total of 300 trials repeated 3 times each.

The trials were pre-generated with custom code and stored in a JSON-formatted data file. The file contained 30000 trials, 3000 for each N , of which 600 for each K . At the beginning

of the experiment, 300 trials are selected at random from the file: 30 trials for each N , of which 6 for each K .

The task was divided into 30 small blocks of 30 trials each. In each block, all distinct trials for a given value of N were shown, in random order. The first block contained all trials with $N = 70$, the second block those with $N = 64$, and so on, in a progressively decreasing sequence of N until the last block, which contained the trials with $N = 16$. This sequence of blocks with decreasing values of N was then repeated 3 times. The participants were not informed of this repetition, and the trials were shuffled between blocks. The task was self-paced, and the participants were free to take breaks as they saw fit. Following the introductory tutorial, the participants performed a small number (60) of training trials. After each training trial, the participant was told whether their answer was correct. If it was not correct, we provided an assessment of how likely their answer was (*likely*, *fairly likely*, *unlikely*, *very unlikely*), according to the mathematical model used to generate the data for different values of K . This evaluation was based on the likelihood of the vertical positions of the lines according to a model with K components, where K corresponded to the answer given, and μ_i of each component was chosen to maximize the likelihood.

After the training phase, no more feedback was given to the participants on their performance.

We ran the experiment on the online platform Pavlovia (pavlovia.org), with participants recruited on Prolific (prolific.com). We collected data from at least 100 participants who passed a pre-established performance threshold (25% or higher on each of the three sequences of blocks of the experiment; this threshold and the number of participants were fixed at preregistration). We discarded the data collected from all other participants. The final dataset included 100 participants (47% female; mean age = 34.16, std = 10.78, range 19-69). The average time required to complete the task was 75 minutes.

2.4.2 Experimental data analysis

To describe the behavior of our participants, we assumed that each of them sampled their choice for each trial from a probability distribution determined by a measure of goodness-of-fit G and a complexity penalty P for the models considered. We defined different sets of these terms, each connected to a possible explanation for their decision strategies. All the resulting probability distributions shared the same structure:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_G G_M - \beta_P P_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_G G_m - \beta_P P_m)} \quad (2.1)$$

Where β_G and β_P represent how sensitive the participant is to the goodness-of-fit or the complexity penalty of the models, respectively, and $\mathcal{M}$ is the set of five models the participants could choose from, as described in the previous section.

We used hierarchical Bayesian multinomial logistic regression to measure the parameters of each of the descriptions obtained from the different terms, and compared them using LOO to find the one fitting human behavior better. Here we explain each of the descriptions and relative terms used, the details of the regressions, and of the comparison. All the descriptions, except for the “independent” one, were preregistered. The “independent” description was added upon realizing that it could be a reasonable and parsimonious candidate to describe human behavior, despite the scaling trend of the average answer observed in Fig 2.2c.

Log-likelihood description

To represent the hypothesis that participants lacked a simplicity bias, and only used a goodness of fit measure, we defined a model selection probability that only depends on the maximum log-likelihood L_M of each model for the trial data. The maximum log-likelihood for the Gaussian mixture models presented to the participants is given by:

$$L_M = \ln p(X_t | \vec{\mu}_M^*; \sigma = 0.1) = C + \sum_{i=1}^N \ln \left(\frac{1}{K} \sum_{k=1}^K \exp \left(-\frac{(x_i - \mu_{M_k}^*)^2}{2\sigma^2} \right) \right) \quad (2.2)$$

Where X_t is the data of trial t , x_i is the horizontal position of a singular data-point, N is the number of data-points, K is the number of components of the gaussian mixture distribution, $\vec{\mu}_M^*$ is the vector of component means that maximizes the likelihood for model M and $C = -\frac{N}{2} \ln(2\pi\sigma^2)$ is a model-independent constant. The resulting probability to choose the model M is:

$$P(\text{choice} = M | X) = \frac{\exp(\beta_L L_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m)} \quad (2.3)$$

To find the values of μ^* of each trial for each model, we used a modified Expectation-Maximization algorithm (EM algorithm, [Dempster et al., 1977](#)). Since the standard deviation and weight of the components were fixed, to find the means we followed these steps:

1. Generate an initialization of $\vec{\mu}$;
2. Compute the *responsibility* γ_{nk} of each component k for each data-point x_n :

$$\gamma_{nk} = \frac{\exp(-||x_n - \mu_k||^2/2\sigma)}{\sum_j \exp(-||x_n - \mu_j||^2/2\sigma)}$$

3. Compute each μ_k as the mean of the data-points weighted by the responsibility of k :

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} x_n$$

4. Repeat from step 2. until convergence.

Since the EM algorithm is only guaranteed to converge to a local maximum of the likelihood, we repeated the process with many different initializations, including:

- The positions of the true means that generated the data if $K = K_{\text{true}}$;
- Every subset of size K of the original means if $K < K_{\text{true}}$;
- The union of the true means and every possible subset of size $K - K_{\text{true}}$ of the true means if $K > K_{\text{true}}$. The means in the subsets were given a small displacement to avoid degeneracies;
- The centroids of a K-means clustering ([MacQueen, 1967](#)) with K components.

Cluster width and imbalance description

To represent the hypothesis that the log-likelihood doesn't accurately capture how humans measure a model's goodness-of-fit, and more generally that a different approach might better describe human behavior, we defined a model selection probability that depends on heuristics tailored to the stimuli. Each trial consisted of vertical lines grouped in clusters, generated by gaussian components with equal standard deviation and weight. Therefore, a natural way to determine if a K-component model is a good fit for a trial is to check if it produces clusters with a similar number of points and similar widths. More specifically, we considered the maximum width of the cluster (W) and the maximum imbalance in the number of points between clusters (I). These measures don't penalize the complexity of a model, but only tell us how close the model's partition of the data is to an optimal partition, according to the properties of Gaussian Mixtures. Therefore, similarly to the previous framework, this is a goodness-of-fit-based description.

To compute the terms, we clustered the data of each trial using the same EM algorithm described for the Log-likelihood description. This algorithm allows for overlapping clusters and favors balanced clusters when possible. The maximum width W_M of a model was defined as the maximum weighted standard deviation among all its clusters. The weighted standard deviation of each cluster was calculated using the responsibilities γ_{nk}^* corresponding to the

maximum likelihood means μ^* of the model. Therefore, data-points with a higher probability of belonging to a cluster contribute more to its size calculation.

$$W_M = \max_{k=1\dots K} \sqrt{\frac{\sum_{n=1}^N \gamma_{nk} (x_n - \mu_{M_k}^*)^2}{\sum_{n=1}^N \gamma_{nk}}}$$

The maximum imbalance of points I_M is a measure of how evenly the data-points are distributed among its clusters. It was defined as the maximum size of any cluster compared to the optimal balanced size N/K . A perfectly balanced model, where all clusters are of equal size, will have a balance value of 1. A value greater than 1 indicates an imbalance, with some clusters being disproportionately more populated than others. The balance measure was computed by summing the responsibilities γ_{nk} of each cluster to determine its effective number of points, and then finding the largest of these normalized and scaled cluster sizes:

$$I_M = \max_{k=1\dots K} \left(\frac{\sum_{n=1}^N \gamma_{nk}}{N/K} \right)$$

The resulting probability to choose model M is:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_W W_M + \beta_I I_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_W W_m + \beta_I I_m)} \quad (2.4)$$

AIC and BIC “computation” description

We defined two descriptions of behavior based on two well known Model Selection frameworks, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria are approximations of difficult-to-compute quantities. We used them as both a baseline to represent the hypothesis that participants display a simplicity bias in their strategies, and more specifically, to represent the hypothesis that strategies were based on a computation that parallels the ones used in theoretical model selection.

In the case of BIC, the computation being approximated is a Bayesian integration. BIC is an approximation of Bayesian Model Selection (BMS), based on the definition of “best model” as the one among a set with the highest posterior probability:

$$P(M|X) = \frac{P(X|M)P(M)}{P(X)} \quad (2.5)$$

In our experiment, the model prior is uniform, so the only quantities needed to find the highest posterior probability are the marginal likelihoods of the data for each model:

$$P(X | M) = \int_{\mathbb{R}^K} p(X | \vec{\theta}_M) p(\vec{\theta}_M | M) \quad (2.6)$$

where $\vec{\theta}_M$ is the K -dimensional vector of model parameters. Using the Laplace approximation on the integral yields:

$$P(X | M) \simeq \left(\sqrt{\frac{2\pi}{N}} \right)^K p(X | \vec{\theta}_M^*) p(\vec{\theta}_M^* | M) |J(\vec{\theta}_M^*)|^{-1/2} \quad (2.7)$$

Applying the logarithm, and ignoring the $O(1)$ terms, gives us BIC (Fig. 2.1a):

$$\log P(X | M) \simeq \log \left(p(X | \vec{\theta}_M^*) \right) - \frac{K}{2} \log(N) = L_M - D_M \quad (2.8)$$

Therefore, if participants base their choices on a Bayesian integration, or on an approximation of this computation, we expect their answers to display a scaling in N similar to that of a BIC-like description.

In the case of AIC, the criterion is derived from an estimate of the Kullback-Leibler divergence between the best fitting distribution $f(X | \theta_M^*)$ belonging to a parametric model M and the true distribution $g(X)$ that generated the data:

$$KL(g(X) || f(X | \theta_M^*)) = E_{g(X)} \left[\log \frac{g(X)}{f(X | \theta_M^*)} \right] \quad (2.9)$$

Where $E_{g(X)}$ is the expected value under the distribution $g(X)$. Using the properties of the logarithm, we can rewrite the divergence as:

$$KL(g(X) || f(X | \theta_M^*)) = C(X) - E_{g(X)} [\log f(X | \theta_M^*)] \quad (2.10)$$

Where $C(X)$ is a quantity that only depends on the data: when comparing two models, this quantity is the same for both of them, thus we can ignore it. The remainder $D(\theta_M^*, X)$ is still dependent on the true distribution $g(X)$, a quantity that is usually inaccessible during model comparison. Akaike showed that under some assumptions, $D(\theta_M^*, X) \simeq -\log f(X | \theta_M^*) + K_M$, where K_M is the number of parameters of the model. Therefore, finding the model that minimizes the KL divergence corresponds to finding the one that maximizes $L_M - K_M$.

While not directly derived from a computation that could be intuitively performed by participants, AIC finds the model with the maximum expected out-of-sample likelihood,

i.e., the model that, on average, better describes unseen data. This makes it so that AIC approximates more computationally intensive methods, such as Cross Validation (Hastie, 2009; Stone, 1974). Therefore, if participants base their choices on a computation that tries to find the highest predictive power among models, we expect their answers to display a scaling in N similar to that of an AIC-like description.

To compute the terms of both AIC and BIC “computation” descriptions, we used the same calculation of L_M presented for the Log-likelihood description. The penalty terms are trivial to compute. The resulting probabilities to choose model M are:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L L_M + \beta_K K_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m + \beta_K K_m)} \quad (2.11)$$

for AIC, and

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L L_M + \beta_D D_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m + \beta_D D_m)} \quad (2.12)$$

for BIC.

AIC “independent” description

To represent the hypothesis that participants lack any scaling, we defined a description of their behavior that doesn’t depend on the number of data-points N . On average, the log-likelihood grows linearly in the number of data-points. Indeed

$$L_M = \log(p(X \mid \theta_M^*)) = \log \left(\prod_{i=1}^N p(x_i \mid \theta_M^*) \right) = \sum_{i=1}^N \log(p(x_i \mid \theta_M^*)) = N l_M \quad (2.13)$$

Where $l_M = \langle \log(p(x_i \mid \theta_M^*)) \rangle_e$, where $\langle \cdot \rangle_e$ is the empirical mean. This quantity also corresponds to the negative of the Kullback-Leibler divergence between the empirical measure of the data $p_e(x) = \sum_{i=1}^N \delta(x - x_i)/N$ (where $\delta(x - x_i)$ is Dirac’s delta) and the candidate model

distribution, up to a model-independent additive value:

$$KL(p_e(x) || p(x | \theta_M^*)) = \mathbb{E}_{x \sim p_e(x)} \left[\log \frac{p_e(x)}{p(x | \theta_M^*)} \right] \quad (2.14)$$

$$= \int p_e(x) \log \frac{p_e(x)}{p(x | \theta_M^*)} dx \quad (2.15)$$

$$= \frac{1}{N} \sum_{i=1}^N \log \frac{p_e(x_i)}{p(x_i | \theta_M^*)} \quad (2.16)$$

$$= -\frac{1}{N} \sum_{i=1}^N \log p(x_i | \theta_M^*) - H(p_e(x)) \quad (2.17)$$

$$= -\frac{L_M}{N} - \log(N) \quad (2.18)$$

Where $H(p_e(x)) = \log(N)$ is the entropy of the empirical measure.

We used l_M as the goodness-of-fit term of this description. For the penalty, we used the complexity penalty of AIC, K_M , as it is independent of N . The disproportion in relative size of goodness-of-fit and penalty after these adjustments is balanced in the regression by the β sensitivities, which will therefore result bigger for the goodness-of-fit in this description compared to the previous ones.

The resulting probability to choose model M is:

$$P(\text{choice} = M | X) = \frac{\exp(\beta_l l_M + \beta_K K_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_l l_m + \beta_K K_m)} \quad (2.19)$$

AIC and BIC “heuristic scaling” descriptions

To represent the hypothesis that participants adopt strategies based on a heuristic rebalancing of the relative weight of goodness-of-fit and complexity penalty, based on the perceived size of the data, we defined two descriptions where the terms depend not on N , but on $\tilde{N} = \log(N)$. The choice for a sublinear function was driven by the literature on numerosity perception, where a compression on the perceived quantity of data is well documented (Anobile et al., 2012, 2015; Cheyette and Piantadosi, 2020; Krueger, 1972). We specifically chose the logarithm function as a compressive function to parallel Weber's law (Fechner, 1860), that has been shown to govern numerosity perception in humans (Burr et al., 2018; Merten and Nieder, 2009).

The goodness-of-fit term with logarithmic scaling was defined as $\tilde{N} \cdot l_M = \log(N) \cdot l_M$, where $\log$ is the natural logarithm. For the AIC-based description, the complexity penalty remained unchanged. For the BIC-based description, it was defined as $\log(\tilde{N}) \frac{K_M}{2} =$

$\log(\log N) \frac{K_M}{2}$. Since the values of N we used were all greater than Euler's number e , there were no undefined or negative values of $\log(\tilde{N})$.

The resulting probabilities to choose model M are:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L \log(N) l_M + \beta_K K_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_L \log(N) l_m + \beta_K K_m)} \quad (2.20)$$

for AIC and

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L \log(N) l_M + \beta_D \log(\log N) \frac{K_M}{2})}{\sum_{m \in \mathcal{M}} \exp(\beta_L \log(N) l_m + \beta_D \log(\log N) \frac{K_m}{2})} \quad (2.21)$$

for BIC.

AIC and BIC “free scaling” descriptions

To find the actual shape of the scaling of our participants, we defined a modified version of the “heuristic scaling” descriptions. In these versions, $\tilde{N}$ was not defined as the logarithm of N , but as a free function to be inferred. For both AIC and BIC, we had two versions: one with a common $\tilde{N}$ for the entire population and one with a specific value of $\tilde{N}$ for each participant. The details of the inference are presented below.

The resulting probabilities to choose model M are:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L f(N) l_M + \beta_K K_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_L f(N) l_m + \beta_K K_m)} \quad (2.22)$$

for AIC and

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L f(N) l_M + \beta_D f(\log N) \frac{K_M}{2})}{\sum_{m \in \mathcal{M}} \exp(\beta_L f(N) l_m + \beta_D f(\log N) \frac{K_m}{2})} \quad (2.23)$$

for BIC.

Lapse rate

We extended all the defined probabilities to account for lapses by defining a lapse rate ε for each participant. Therefore, on each trial, there is a probability $(1 - \varepsilon)$ that the participant will perform the task and act following their strategy, and a probability ε that they will answer randomly. The general structure of the probability to choose model M will therefore be:

$$P(\text{choice} = M \mid X) = \varepsilon \cdot \frac{1}{K} + (1 - \varepsilon) \cdot \frac{\exp(\beta_G G_M - \beta_P P_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_G G_m - \beta_P P_m)} \quad (2.24)$$

Where $\frac{1}{K}$ represent the uniform probability to randomly choose model M .

To be able to get a reliable model comparison, we also added a second lapse rate δ (see [Model comparison](#)). While ϵ is the lapse rate of a single participant, the same δ is in common between all participants. Philosophically, this term represents the subject-independent, global probability that external events will lead to errors in the execution of the task. From a practical point of view, this term can be included in ϵ by slightly changing its definition: if we define as $\tilde{\epsilon}$ the lapse of a participant, their overall lapse will become $\epsilon = \delta + \tilde{\epsilon}$. Therefore, adding δ doesn't affect the above formula, but only alters the prior distribution of ϵ .

2.4.3 Bayesian regression

In the formulas for each description of possible strategies we defined, the probability to choose a model M , $P(M | X)$, depends on the subject-specific parameters $\theta^s = \{\vec{\beta}^s, \epsilon^s\}$ (where $\vec{\beta}^s$ represents the parameters of a specific strategy for subject s). To infer their values we used a hierarchical Bayesian approach ([Gelman et al., 2014a](#)). We used the same regression structure for all strategies other than the “free scaling” ones, which are described in more detail in the next section.

For all the descriptions, denoting by C the recorded choices of all participants, and by C_t^s the recorded choice of participant s on trial t , we define the likelihood function of the associated parameters θ^s is:

$$\mathcal{L}^s(\theta^s) = p(C^s | \theta^s; X^s) \quad (2.25)$$

$$= \prod_{t=1}^{N_t} p(\text{choice} = C_t^s | \theta^s; X_t^s) \quad (2.26)$$

$$= \prod_{M=1}^5 \prod_{t: C_t^s = M} p(\text{choice} = M | \theta^s; X_t^s) \quad (2.27)$$

We also define a prior for each of the parameters:

$$p(\theta^s) = p(\epsilon^s) \prod_{\text{terms}} p(\beta_{\text{term}}) \quad (2.28)$$

where the terms are specific to each description.

By applying Bayes' theorem, we compute a joint posterior probability distribution for the parameters:

$$p(\theta^s | C^s) \propto p(C^s | \theta^s) p(\theta^s) \quad (2.29)$$

To extend the same computation to the entire population we use a hierarchical approach. This allows us to infer population-level parameters and, at the same time, regularize the inference procedure, with a regularization strength that depends on the data (Gelman et al., 2014a): since the participants belong to the same population and are exposed to the same experimental conditions, we assume that the parameters describing them will be related. In other words, the data we collect about one participant tells us something about that specific participant, but should also inform us about human behavior in general, and thus about the behavior of other participants. We can build this assumption into the model by imposing that the parameters associated to different participants are all sampled from the same distribution, leaving the parameters of this higher-level probability distribution to be also inferred from the data. This procedure has the additional advantage of partially constraining the subject-level parameters to reasonable values given what we observed in the other participants, thus acting as a regularizer for the inference process. Therefore, our posterior will be:

$$p(\theta, \zeta \mid C) \propto p(C \mid \theta)p(\theta \mid \zeta)p(\zeta) \quad (2.30)$$

where $\theta = \{\theta^s\}_{s=1}^{N_s}$ are the subject-specific parameters for all N_s participants, and ζ are the common (population) parameters.

To ensure a more robust inference, the terms of each description have been standardized to a mean of zero and a standard deviation of one before fitting the model. The “free scaling” descriptions’ terms, as previously stated, could not be standardized, since they are computed during the inference process. Therefore, we standardized L_M before multiplying it by $\frac{f(N)}{N}$. For the dimensionality term, we standardized $\frac{K_M}{2}$ before multiplying it by $\log(f(N))$. In the following, we will refer to the parameters as $\hat{\beta}$ when they are associated with a standardized term and β when they are not.

In the likelihood-only description, we also added to each likelihood, before standardization, a small noise $\in [-0.5, 0.5] \cdot 10^{-5}$: this small shift ensured a stable inference in those trials where more than one model had the exact same likelihood. Indeed, for those trials, the probabilities assigned by the models ($p \propto \exp(\beta_L L_M) = 0.2$ if all the models have the same likelihood) are independent of β_L . This caused the same importance weights to be assigned to all the parameters of the MCMC chain for these trials (see [Model comparison](#)), which in turn made the Pareto inference of the PSIS LOO computation unreliable.

The specific priors we assign to the parameters ($p(\theta \mid \zeta)$ in Eq (2.30)) are:

$$\delta \sim \text{Beta}(\alpha = 1, \beta = 99) \quad (2.31)$$

$$\varepsilon^s \sim \text{LogitNormal}(\mu_{\text{logit } \varepsilon}, \sigma_{\text{logit } \varepsilon}) \quad \forall s \quad (2.32)$$

$$\hat{\beta}_{\text{term}}^s \sim \mathcal{N}(\hat{\mu}_{\beta_{\text{term}}}, \hat{\sigma}_{\beta_{\text{term}}}^2) \quad \forall \text{ terms}, s \quad (2.33)$$

Our population parameters, ζ , are therefore the logit mean $\mu_{\text{logit } \varepsilon}$ and standard deviation $\sigma_{\text{logit } \varepsilon}$ of the ε^s parameters, and $\hat{\mu}_{\beta_{\text{term}}}$ and $\hat{\sigma}_{\beta_{\text{term}}}^2$ for the β^s parameters of each term. Following standard practice, we assign weakly informative priors $p(\zeta)$ to the population-level parameters:

$$\mu_{\text{logit } \varepsilon} \sim \mathcal{N}(\mu = -3, \sigma = 1.5) \quad (2.34)$$

$$\sigma_{\text{logit } \varepsilon} \sim \text{HalfNormal}(\sigma = 0.5) \quad (2.35)$$

$$\hat{\mu}_{\beta_{\text{term}}} \sim \mathcal{N}(\mu = 0, \sigma = 10) \quad \forall \text{ terms} \quad (2.36)$$

$$\hat{\sigma}_{\beta_{\text{term}}} \sim \text{Exponential}(\text{scale} = 1) \quad \forall \text{ terms} \quad (2.37)$$

$$(2.38)$$

The same priors were assigned to all descriptions, with the only exception of the Log-likelihood description, that had a prior

$$\hat{\mu}_{\beta_l} \sim \mathcal{N}(\mu = 0, \sigma = 1.5) \quad (2.39)$$

for its only β term, since less informative priors lead to less robust inferences.

Finally, the global likelihood function was obtained as a product of the subject-level likelihood functions:

$$\mathcal{L}(\theta) = p\left(C \mid \left\{\vec{\beta}_s, \varepsilon_s\right\}_{s=1}^{N_s}, \delta\right) \quad (2.40)$$

$$= \prod_{s=1}^{N_s} \prod_{t=1}^{N_t} p(\text{choice} = C_t^s \mid \theta^s; X_t^s) \quad (2.41)$$

Gaussian Processes

In the “free scaling” descriptions, the scaling of both goodness-of-fit and penalty terms is a generic function $f(N)$. To infer the shape of this function from the data, we employed a Gaussian Process (GP) (MacKay et al., 1998). GPs are a principled, non-parametric method

for defining a prior over functions, allowing us to treat the whole function as a parameter of the regression.

We had two versions of “free scaling” description, for both AIC and BIC. In the population-scaling version, we had only one scaling function $f(N)$ for all participants. In this version, the GP prior was defined as:

$$f(N) \sim \mathcal{GP}(\mu_{f(N)}, K(N, N')) \quad (2.42)$$

where $\mu_{f(N)}$ is the mean of the prior for each value of N (typically set to $\mu_{f(N)} = 0$) and $K(N, N')$ is the Exponentiated Quadratic covariance kernel. A GP kernel is a function that allows us to compute the covariance between any two values of N , N_i , and N_j :

$$k(N_i, N_j) = \eta^2 \exp\left(-\frac{(N_i - N_j)^2}{2l^2}\right) \quad (2.43)$$

where η is the marginal standard deviation of the Gaussian process, which controls the overall vertical amplitude of the function, and l is the length scale, which determines its smoothness. We placed weakly informative priors on these population-level hyperparameters:

$$\eta \sim \text{HalfNormal}(\sigma = 1) \quad (2.44)$$

$$l \sim \text{Gamma}(\alpha = 5, \beta = 2) \quad (2.45)$$

A challenge in this approach is potential model degeneracy: if both the regression coefficients (β_L or β_D) and the function $f(N)$ (modeling, together, the goodness-of-fit term and the dimensionality penalty) are flexible, the model can become unidentifiable. For instance, one can double the scale of the GP function while halving the β_L coefficient without changing the likelihood. To resolve this non-identifiability issue, we enforce the constraint $GP(N = 16) = 16$. This is equivalent to using the GP function to model scaling in a relative sense, leaving the absolute magnitude of the goodness-of-fit and dimensionality terms to be controlled by the regression coefficients β .

To implement this constraint, we defined the GP as a conditional multivariate normal distribution. Given the full covariance matrix K over all unique values of N , we partition it as follows:

$$K = \begin{pmatrix} K_{00} & K_{0x} \\ K_{x0} & K_{xx} \end{pmatrix} \quad (2.46)$$

where K_{00} is the variance at the reference point ($N = 16$), K_{xx} is the covariance matrix of the other points, and K_{x0} contains the cross-covariances. The values of the GP for all values

of N other than 16, $f(N_x)$, are then drawn from a multivariate normal distribution whose parameters are conditioned on the fixed value $f(16) = 16$:

$$\mu_{\text{cond}} = \mu_{f(N_x)} + K_{x0}K_{00}^{-1} \cdot (16 - \mu_{f(N_0)}) = 16 \quad (2.47)$$

$$\Sigma_{\text{cond}} = K_{xx} - K_{x0}K_{00}^{-1}K_{0x} \quad (2.48)$$

Where, to impose $GP(16) = 16$, we put $\mu_{f(N)} = 16$ for all N .

To ensure numerical stability and improve sampling efficiency, we adopt a non-centered parameterization. Instead of sampling directly from this correlated conditional distribution, we sampled a vector of independent standard normal variables z and transformed them using the Cholesky decomposition (Higham, 2009) of the conditional covariance matrix, $L = \text{chol}(\Sigma_{\text{cond}})$, so $f(N_x) = \mu_{\text{cond}} + L \cdot z$. Moreover, we rescaled the input values of the GP (i.e., of the kernel function) as $\hat{N} = (N - 16)/16$.

In the subject-scaling version of the descriptions, each participant had their own function $f_s(N)$ to account for individual differences in the scaling. As for the other parameters of the description, we developed a hierarchical formulation: we have a common, population-level scaling function $f_\mu(N)$ acting as the population mean for the subject-level scaling functions.

For these subject-scaling versions, the definition of the AIC and BIC descriptions differs.

In the AIC description, the definition of $f_\mu(N)$ was the same as the definition of $f(N)$ seen for the population-scaling versions. To model the subject-level functions we used a non-centered parametrization:

$$f_s(N) = f_\mu(N) + \eta_s g(N) \quad (2.49)$$

Where $g_s(N)$ is a zero-mean GP with an Exponentiated Quadratic kernel, marginal standard deviation set to 1, and length scale l_s . The population-level priors for these functions are:

$$\eta_{\text{participants}} \sim \text{HalfNormal}(\sigma = 1) \quad (2.50)$$

$$l_{\text{participants}} \sim \text{Gamma}(\alpha = 5, \beta = 2) \quad (2.51)$$

Where both $\eta_{\text{participants}}$ and $l_{\text{participants}}$ are common to all participants. As for the case of $f_\mu(N)$, to maintain inference stability, the $g_s(N)$ are anchored, this time at $GP_s(N = 16) = 0$, using the same decomposition of the covariance matrix seen above.

In the BIC description, the function $f_s(N)$ is the argument of a logarithm in the dimensionality term $D_M = \log(f_s(N))^{\frac{K_M}{2}}$. While this didn't prove to be a problem in the population-scaling version of the description, in the subject-scaling version a definition of $f_s(N)$ as the one used for AIC leads to undefined values.

To avoid this problem, we instead inferred the functions $\phi_\mu(N) = \log(f_\mu(N))$ and $\phi_s(N) = \log(f_s(N))$. We used the same scheme defined for the AIC description, with the only difference that we use stricter priors for the marginal standard deviations of the Gaussian processes:

$$\eta_\mu^\phi \sim \text{HalfNormal}(\sigma = 0.1) \quad (2.52)$$

$$\eta_s^\phi \sim \text{HalfNormal}(\sigma = 0.1) \quad (2.53)$$

$$(2.54)$$

This was necessary since the exponentiation $f_s(N) = \exp(\phi_s(N))$ can lead to extremely high values of f and destabilize the inference.

Sampling the posterior distribution.

We performed inference with a Markov Chain Monte Carlo (MCMC) method, using the NUTS algorithm (Hoffman and Gelman, 2014), as implemented in PYMC version 5.25.1 (Salvatier et al., 2016). This method samples parameters from the posterior distribution $p(\theta | C)$: with a high enough number of samples, the distribution of the samples matches closely the distribution of the posterior. We sampled 4 independent chains of 10000 draws each, 5000 of which we used for tuning the algorithm and subsequently discarded, and 5000 as actual samples of the target posterior distribution. For computational convenience, we applied a thinning process to reduce the number of posterior samples, keeping only 1000 for the analyses. The target acceptance probability parameter for NUTS was set at 0.95, and no divergences were detected during sampling (see Tables A.1 to A.9).

2.4.4 Model comparison

To verify which of the descriptions better fitted human behavior, we compared them with a Leave-One-Out Cross-Validation (LOO) estimate of the Expected Log pointwise Predictive Density (ELPD). LOO estimates how well, on average, the model (i.e., a description) describes data it hasn't been trained on. To do so, LOO computes how well the result of a regression on all the trials (across all participants) except one predicts the outcome of that trial, and averages this value over all the trials. The ELPD, therefore, automatically balances goodness of fit with model complexity, penalizing overly complex models that tend to overfit the noise in the data. More formally, the LOO estimate of the ELPD for a statistical model

$\mathcal{M}$ for which we can sample the posterior distribution with MCMC is defined as:

$$\text{ELPD}_{\text{LOO}}(\mathcal{M}) = \sum_{i=1}^{N_T} \log \left(\frac{1}{N_{\text{samples}}} \sum_{j=1}^{N_{\text{samples}}} p_{\mathcal{M}}(c_i | \theta_{(j; C_{-i})}) \right) \quad (2.55)$$

where N_T is the total number of trials in the data, N_{samples} is the number of posterior samples in the Markov chain for model $\mathcal{M}$, $p_{\mathcal{M}}$ is the likelihood function of the model, c_i is the data (i.e., the human choice) from the i -th trial and $\theta_{(j; C_{-i})}$ is the j -th posterior sample of a Markov chain generated from all data except c_i (note that each posterior distribution sample, including $\theta_{(j; C_{-i})}$, is a high-dimensional vector containing one value for each parameter in the model). In Eq (2.55), the argument of the logarithm represents the likelihood of the description $\mathcal{M}$ for the choice c_i , averaged over the posterior distribution over θ (where this posterior is obtained without looking at data-point c_i). Indeed, since the training of $\mathcal{M}$ results in a posterior distribution rather than a single parameter vector, all quantities related to $\mathcal{M}$ must be averaged over this distribution. For MCMC training, this average is computed as a sum over all sampled parameters, given that the samples follow the posterior distribution: $\frac{1}{N_{\text{samples}}} \sum_{j=1}^{N_{\text{samples}}} p_{\mathcal{M}}(c_i | \theta_{(j; C_{-i})}) \simeq \int p_{\mathcal{M}}(c_i | \theta) p(\theta | C_{-i}) d\theta = \langle p_{\mathcal{M}}(c_i | \theta) \rangle_{p(\theta | C_{-i})}$. In practice, computing Eq (2.55) requires training the model N_T times, one per trial: given the large number of trials in our experiment, we approximated the leave-one-out procedure with Pareto-smoothed importance sampling (PSIS-LOO).

Importance Sampling LOO is based on the fact that, under the hypothesis of conditional independence of the data given the parameters,

$$p(\theta_j | C) = \frac{p(C | \theta_j) p(\theta_j)}{p(C)} = p(c_i | \theta_j) \frac{p(C_{-i} | \theta_j) p(\theta_j)}{p(C)} \quad (2.56)$$

$$= p(c_i | \theta_j) \frac{p(C_{-i} | \theta_j) p(\theta_j)}{p(C_{-i})} \frac{p(C_{-i})}{p(C)} \quad (2.57)$$

$$= p(c_i | \theta_j) p(\theta_j | C_{-i}) \frac{p(C_{-i})}{p(C)} \quad (2.58)$$

$$\propto p(c_i | \theta_j) p(\theta_j | C_{-i}) \quad (2.59)$$

Therefore,

$$p(\theta_j | C_{-i}) \propto r_i^j p(\theta_j | C) \quad (2.60)$$

where

$$r_i^j = \frac{1}{p(c_i | \theta_j)} \quad (2.61)$$

which implies that we can approximate a Monte Carlo average over $p(\theta_j|C_{-i})$ by one over $p(\theta_j|C)$, where each sample is weighted by the importance weight r_i^j . So instead of writing

$$\int p(c_i|\theta)p(\theta|C_{-i})d\theta \simeq \sum_{\substack{j=1 \\ \theta_j \sim \theta|C_{-i}}}^{N_{\text{samples}}} p(c_i|\theta_j) \quad (2.62)$$

we can write

$$\int p(c_i|\theta)p(\theta|C_{-i})d\theta \simeq \left(\sum_{\substack{j=1 \\ \theta_j \sim \theta|C}}^{N_{\text{samples}}} p(c_i|\theta_j)r_i^j \right) \left(\sum_{\substack{j=1 \\ \theta_j \sim \theta|C}}^{N_{\text{samples}}} r_i^j \right)^{-1} \quad (2.63)$$

Using the r_i^j as weights, we can approximate the sum on N_{samples} in 2.55 as:

$$\sum_{j=1}^{N_{\text{samples}}} p_{\mathcal{M}}(c_i | \theta_{(j;C_{-i})}) \simeq \frac{\sum_{j=1}^{N_{\text{samples}}} r_i^j p_{\mathcal{M}}(c_i | \theta_{(j;C)})}{\sum_{j=1}^{N_{\text{samples}}} r_i^j} = \frac{N_{\text{samples}}}{\sum_{j=1}^{N_{\text{samples}}} \frac{1}{p_{\mathcal{M}}(c_i|\theta_j)}} \quad (2.64)$$

The problem with this approximation is that $p_{\mathcal{M}}(c_i | \theta_j)$ can be arbitrarily small, leading to instabilities, as a single element of the chain can greatly skew the result. To avoid this, a Pareto distribution is fitted on the M biggest r_i^j (corresponding to the smallest $p_{\mathcal{M}}(c_i | \theta_j)$) for each i , and the weights are substituted by new weights sampled from the fitted distribution, with a threshold to avoid sampling extremely high weights.

Necessity of δ in the lapse rate In our dataset, a few participants lapsed (or were considered lapsing by the description) only one time. For the trained model, this means that that participant has $\varepsilon > 0$. However, during the computation of *LOO*, the exclusion of that trial results in $\varepsilon \simeq 0$. This means that that trial alone guides the value of the inferred ε : it's a highly influential observation for the regression. These kinds of observations are not problematic for LOO, but can greatly affect the performance of the PSIS approximation. To avoid this kind of trial to be too influential, we included the parameter δ : this parameter gives ε a lower bound, computed across all participants, thus avoiding the possibility that removing a single trial will influence too much the inferred lapse rate.

Necessity for the added noise to L_M for the likelihood-only description For the likelihood-only description, for some trials all the probabilities $p_{\mathcal{M}}(c_i | \theta_j)$ are the same, independently of θ_j , for all j . This causes problems when fitting the Pareto distribution, as the weights are

all equal and cannot be meaningfully interpreted as being Pareto distributed. The fitted Pareto distribution has an infinite variance, making the Pareto Smoothed Importance Sampling unreliable. To avoid this problem, we added a small noise to all the likelihoods, thus breaking this symmetry and introducing differences in the values of the weights.

Chapter 3

Characterization of Complexity in Intuitive Model Selection

3.1 Introduction

The previous chapter established that changing the amount of observed data affects the strength of the simplicity bias in humans. While this is qualitatively in line with the prescriptions of theoretical frameworks for model selection, quantitatively this effect proved to be far from the normative predictions. The measure of goodness-of-fit that the participants rely on grows sublinearly with the number of available datapoints: the log-likelihood, the measure used in the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), as well as in other frameworks, grows instead linearly on average. We were able to draw a possible connection between this observed compression of evidence and the perception of numerosity, which is affected by similar biases to the ones observed. Our data turned out to be most informative about the properties of the goodness-of-fit measure used by participants; by contrast, we could not draw strong conclusions about the nature of the complexity penalty. Indeed, when accounting for the sublinear compression of the perceived number of datapoints, AIC and BIC-like descriptions of human behavior showed similar performance in predicting unseen data. Therefore, we could not conclude if a better account of human choices is given by a fixed, AIC-like penalty or a variable, BIC-like one. Moreover, since we used AIC and BIC as normative frameworks, we only focused on penalties dependent on the number of parameters and didn't explore the possible effect of finer measures of model complexity.

In general, a model's complexity is best understood as its *flexibility*, defined as the range of possible datasets it can effectively describe. A highly flexible model can account for a

vast array of outcomes. Conversely, a simpler, more constrained model is parsimonious, making specific predictions about a narrow range of data. When an observation falls within the narrow range predicted by the simple model, it provides stronger evidence for that model. This is because it is more probable that the data was generated by a process that consistently produces such results, rather than by a highly flexible process that could have generated many other outcomes but, by chance, happened to produce the specific one we observed.

This idea is mathematically embodied in Bayesian Model Selection (BMS) (MacKay, 1992). BMS is, in principle, a simple application of Bayes' rule: given some data $\mathbf{X}$ and a set of candidate models $\{M_1, \dots, M_K\}$, the best model is the one with the highest posterior probability $P(M_i | \mathbf{X}) \propto P(\mathbf{X} | M_i)P(M_i)$. If all the candidate models are equally likely *a priori*, as it is often the case, the main quantity of BMS is the marginal likelihood $P(\mathbf{X} | M)$, a distribution that roughly tells us: given a model, what is the kind of data I could expect it to generate? Importantly, the marginal likelihood is normalized on the space of data $\mathbf{X}$, so if it is significantly greater than zero in many regions of this space (i.e., if the model is flexible), its value will be overall low in all regions. On the other hand, a model that only fits well a small region of the space of data will have a high value in that area, thus outperforming the more flexible model.

This interpretation of complexity as flexibility, particularly in the Bayesian framework, has been shown to account well for human behavior (Blanchard et al., 2018; Piasini et al., 2025). In Chapter 2 we ruled out Bayesian integration (i.e., the computation of the marginal likelihood $P(\mathbf{X} | M)$) as a viable explanation for strategies employed by humans; however, BMS can still help formulate more nuanced and principled definitions of complexity that might affect participants' behavior. While the marginal likelihood doesn't explicitly include complexity terms, such as those of BIC and AIC, using approximations to compute it reveals the familiar balance between goodness-of-fit and penalties. The Fisher Information Approximation (FIA, Balasubramanian, 1997; Piasini et al., 2025), for example, is an extension of BIC, and therefore shares the same Log-likelihood and Dimensionality terms. However, it also includes more nuanced terms, all related to geometrical definitions of complexity, that better allow to approximate the true value of $P(\mathbf{X} | M)$ (Balasubramanian, 1997; Piasini et al., 2025). Therefore, compared to BIC, it allows for a better account of complexity as flexibility of the model, other than simply as its number of parameters.

When using the terms of FIA to describe human behavior, Dimensionality proved to be by far the most relevant one (Piasini et al., 2025). Nevertheless, the other terms all had an appreciable effect. More so, one of the terms used in Piasini et al. (2025), the Boundary term, was specific to the particular configuration of the task presented to the participants, hinting at

the fact that humans might not simply rely on task-agnostic definitions of complexity, but adapt their simplicity bias to the characteristics of the problem.

In this chapter, we aim to better characterize what are the model properties that drive the simplicity bias. Our goal is twofold. First, we want to analyze more thoroughly the role and properties of the Dimensionality term. We established that, compared to theory, this term is over-relied upon by humans; what we do here is check if participants indeed use the number of model parameters as a penalty, or if they show biases and divergences from optimality also in the application of this term. Second, we want to verify whether theoretical considerations based on the principles of Bayesian Model Selection (and its Fisher Information Approximation) help better describe human behavior, in order to identify other relevant definitions of complexity that our participants could base their choices on.

Our results show that the perceived dimensionality presents a sublinear compression not dissimilar to the one observed for the goodness-of-fit. We find the cause for this compression not to be related to numerosity perception (that in the regime of very small quantities is linear, [Kaufman et al., 1949](#)) but rather in the interplay of new complexity terms that need to be considered to properly account for human choices. In particular, a term arising from the symmetries of the model presented in our experiment proves to have the strongest and most interpretable effect, suggesting that the penalties utilized by participants *do* depend on the specific properties of the system analyzed. Moreover, we use the terms derived from FIA, as done in [Piasini et al. \(2025\)](#), to verify the effect of more nuanced definitions of complexity. The presence of a new term, R_M , does indeed enhance the predictive performance of the description, but the results are very subject-dependent, and the average sensitivity to the new term is comparable to 0. The distribution of sensitivities across participants suggests that the form of the term needs correction, just like the dimensionality did. This intuition is strengthened by the fact that the presence of this new term, while in general improving the approximation of BMS compared to BIC, only does so in certain regimes of the data, and fails completely in others.

3.2 Sublinear dimensionality

For this study, we used a decision-making task similar to the one in Chapter 3. Participants were presented with noisy visual data and had to understand which of a set of possible models generated it. The models were again one dimensional gaussian mixtures with $K \in [1, 5]$, fixed variance and weights, and variable position of the components. The random data generated by the model was displayed as vertical white lines on a gray background. There were three main differences compared to the previous experiment:

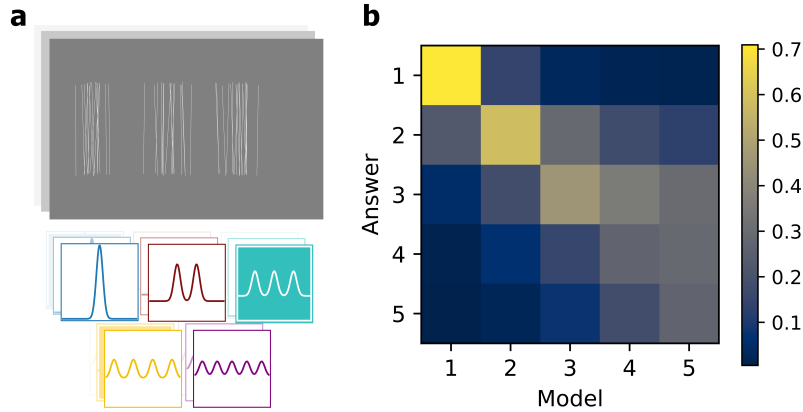


Fig. 3.1 Task, models and participants' answers. **a.** The five models that participants had to choose among (bottom) and an example of a trial from the task (top). The trial shown was generated by a model with three well-separated components. **b.** Average confusion matrix across participants and trials: as in Fig. 2.2c, the effect of the simplicity bias is evident from the fact that the distribution of answers for models with more parameters is skewed towards lower-value answers.

- Since we are not interested in the scaling effect here, the number of vertical lines was fixed to $N = 50$.
- All participants were presented with the same 300 trials, repeated three times each, for a total of 900 trials. The order in which the trials were presented was randomized and different from participant to participant.
- In this version of the experiment, we didn't include the dots that helped better recognize the number of vertical lines. Instead, the lines had a random tilt that helped to better disentangle overlapping datapoints (Fig. 3.1a).

The 900 trials were shown in 10 blocks of 90 trials each, and presented in such a way that the same trial was never repeated in the same block. 46 participants took part in the experiment (43% female; mean age = 31.21, std = 10.93, range 20-62).

We used a Bayesian multinomial logistic regression to measure the degree to which each participant's choices were affected by the goodness-of-fit and the complexity penalty of models. Specifically, as in the previous chapter, we assigned a measure of sensitivity β to each of the terms: the probability to choose a specific model was therefore:

$$p(\text{choice} = M \mid X) \propto \exp(\beta_G G_M + \beta_P P_M)$$

Where G_M is the goodness-of-fit of the model M for the data $\mathbf{X}$ and P_M its complexity penalty. We first trained a description corresponding to the “computation” BIC one from Chapter 3:

this description uses as goodness-of-fit the maximum log-likelihood L_M of the model and as penalty the dimensionality $D_M = K_M \frac{\log(N)}{2}$, where K_M is the number of parameters of the model. The penalties of AIC and BIC for data with a fixed sample size are equivalent, up to a constant multiplicative factor: in this version of the experiment, they would give the same β_G and $\beta_P^{AIC} = \beta_P^{BIC} \frac{\log(50)}{2}$. Therefore, we didn't perform the inference with the AIC-like description.

Unsurprisingly, the inferred values of β confirm the results obtained in Chapter 3: the sensitivity to D_M is stronger than the one to L_M (Fig. 3.2a). The results on the variable- N experiment suggest that this difference is at least partially due to the compression of the perceived sample size, but for simplicity, we didn't include the compression in the fit since, as for the dimensionality argument, the difference would only be a multiplicative factor.

To verify if participants actually use a penalty that linearly grows with the number of parameters in the model, as both D_M and K_M do, we defined a “free penalty” description. In this description the goodness-of-fit is the log-likelihood L_M , and the complexity penalty is a modified dimensionality $\hat{D}_M = \hat{K}_M \frac{\log(N)}{2}$ where the value of the perceived number of parameters $\hat{K}_M$ is inferred from the participant's answers (see [Supplementary Material](#) for details).

We compared the “computation” description with the “free penalty” description using Pareto Smoothed Importance Sampling Leave One Out Cross Validation (PSIS-LOO CV, [Vehtari et al., 2017, 2024](#)). The description with the variable penalty has a significantly higher Expected Log Predictive Density (ELPD), proving a much higher predictive power on data it wasn't trained on (Fig. 3.2b). We inferred both the population-level perceived number of parameters $\mu_{\hat{K}_M}$ and the subject-level one $\hat{K}_M$. The population-level penalty clearly shows a sublinearity in the perceived number of model parameters. However, this sublinearity can't be explained with perceptual biases of numerosity, as we did for the compression observed in the perceived sample size $\hat{N}$. For numbers between 1 and 4, the perception of numerosity is in the *subitizing* range: these values are perceived by participants very quickly, and without error ([Kaufman et al., 1949](#)). Moreover, even if for 5 elements there can be a degree of error in perception, this error is not appreciably sublinear.

If we consider the trend observed for single participants, we get a wide range of different behaviors. While the majority of the participants show a sublinear trend, more or less consistent with the one observed for the mean, some participants show a superlinear trend or even non-monotonic trends that decrease with K_M for small values of K_M and increase for larger values. This could mean that participants genuinely differ in how they weigh model dimensionality, but it could also mean that our description of participant behavior is missing an important source of variability. To explore this possibility, we analyze the origin

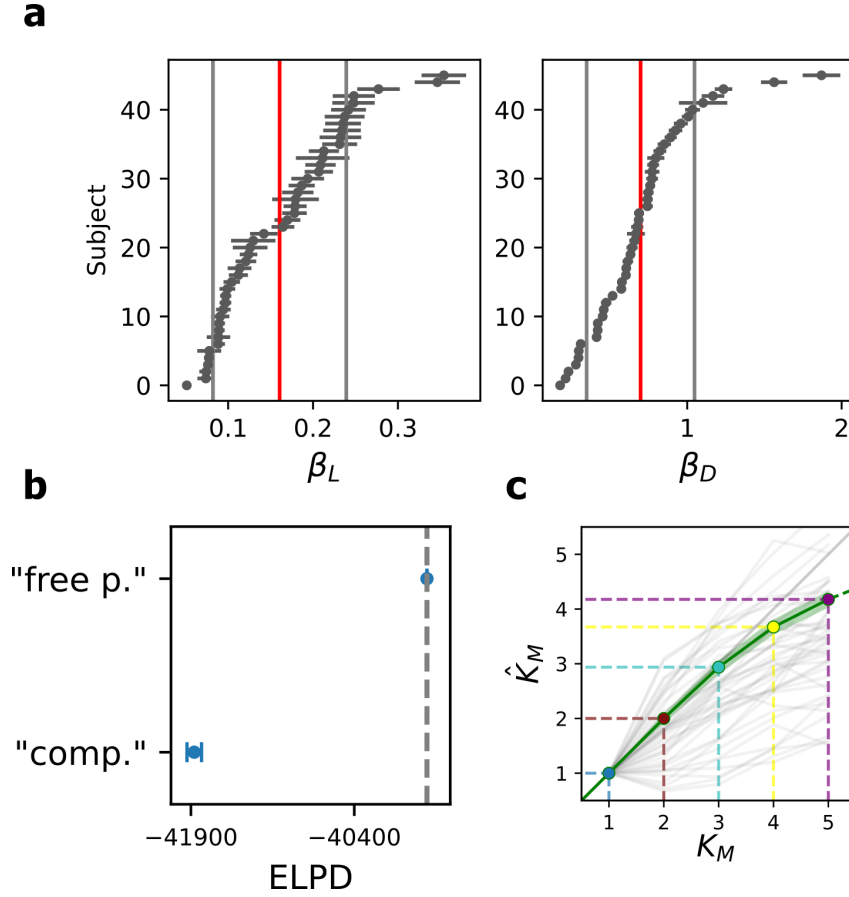


Fig. 3.2 Parameters and comparison of the “computation” and “free penalty” descriptions. **a.** Sensitivities to likelihood (β_L) and dimensionality (β_D) for the “computation” description across the 46 participants. For each plot, the red vertical line represents the posterior mean of the population average μ_β , while the distance between the red line and each of the gray lines represents the population standard deviation σ_β . The dots represent the posterior mean of the values of each participant, and the errorbars are the s.e.m.. The values of both β_L and β_D are sorted independently in each panel for ease of visualization. **b.** Result of the LOO CV comparison between the BIC “computation” description and the “free penalty” description. The x-axis shows the Expected Log Predictive Density (ELPD) of each description, while the errorbars represent the standard error of the ELPD differences with the best model, $\text{ELPD}_{\text{best}} - \text{ELPD}_M$. **c.** Inferred values of the perceived number of parameters $\hat{K}_M$. The green line is the posterior mean of the population average $\mu_{\hat{K}_M}$. The green shaded area is the mean $\pm$ s.e.m.. The gray lines are the values of $\hat{K}_M$ for each participant.

of the penalty term in BIC and look for possible corrections that would explain this range of behaviors.

3.3 Laplace approximation of Bayesian Model Selection

As mentioned in the introduction, and touched upon in Chapter 3, the Bayesian Information Criterion is obtained as an approximation of the marginal likelihood $P(\mathbf{X} | M)$ in Bayesian Model Selection. The marginal likelihood is given by:

$$P(\mathbf{X} | M) = \int_{\mathbb{R}^{K_M}} d\theta_M P(\mathbf{X} | \vec{\theta}_M) P(\vec{\theta}_M) \quad (3.1)$$

Where $\vec{\theta}_M$ is the vector of K_M parameters of M , $P(\vec{\theta}_M)$ is their prior and $P(\mathbf{X} | \vec{\theta}_M)$ the likelihood. For the Gaussian Mixture Models that our participants have to compare, it corresponds to:

$$p(\mathbf{X} | K_M) = \int_{-1}^1 d\mu_1 \dots \int_{-1}^1 d\mu_K P(\mathbf{X} | \vec{\mu}; \sigma, K_M) P(\vec{\mu}) \quad (3.2)$$

Where the likelihood, since the data is independent and identically distributed, is:

$$P(\mathbf{X} | \vec{\mu}; \sigma, K_M) = \prod_{x_i \in \mathbf{X}} \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{j=1}^{K_M} \frac{1}{K_M} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma^2}\right) \quad (3.3)$$

with $\sigma = 0.1$, while the prior is $P(\vec{\mu} | K_M) = \frac{1}{T^{K_M}}$ ¹, with $T = 2$ being the length of the space $[-1, 1]$ where the component means can be located.

This integral is both analytically and numerically difficult to solve, as its result depends on a number of terms that grows exponentially in the number of points (see [Supplementary Material](#)). To overcome this analytical difficulty, the marginal likelihood is obtained by using the Laplace approximation ([MacKay, 2003](#), chapter 27). In general, when solving an integral with the form $I = \int f(\theta) \exp(n\psi(\theta)) d\theta$, if ψ is regular and it has a single global maximum in θ^* , we can use this approximation, which becomes increasingly precise for $n \rightarrow \infty$. The Taylor expansion of ψ around θ^* gives us:

$$\exp(n\psi(\theta)) \simeq \exp\left(n\psi(\theta^*) - \frac{n(\theta - \theta^*)^2}{2} |\psi''(\theta^*)|\right) \quad (3.4)$$

¹More accurately, the prior is $P(\vec{\mu} | K_M) = \prod_{k=1}^{K_M} (\Theta(\mu_k + 1) - \Theta(\mu_k - 1)) / T$, where $\Theta(x)$ is Heavyside step function. For simplicity, in the text we accounted for the step function by changing the extremes of integration from $\mathbb{R}^{K_M}$ to $[-1, 1]^{K_M}$.

Where $-|\psi''(\theta^*)| = \psi''(\theta^*)$ since $\psi(\theta^*)$ is the maximum and thus has negative second derivative. The second term has the form of a Gaussian with mean θ^* and variance $(n|\psi''(\theta^*)|)^{-1}$. Therefore, if we expand $f(\theta)$ to the first order around θ^* , we can solve I as a gaussian integral:

$$I \simeq \exp(n\psi(\theta^*))f(\theta^*) \int \exp\left(-\frac{n(\theta - \theta^*)^2}{2}|\psi''(\theta^*)|\right) d\theta \quad (3.5)$$

$$= \sqrt{\frac{2\pi}{n|\psi''(\theta^*)|}} \exp(n\psi(\theta^*))f(\theta^*) \quad (3.6)$$

If we do the same for a K -dimensional function, we get:

$$I \simeq \left(\frac{2\pi}{n}\right)^{\frac{K}{2}} \det(\hat{H})^{-\frac{1}{2}} \exp(n\psi(\theta^*))f(\theta^*) \quad (3.7)$$

Where $\hat{H} = -\nabla^2\psi(\theta^*)$ is the Hessian matrix of ψ in θ^* .

For our GMMs, $n = N$, $\psi(\vec{\mu}) = \log P(\mathbf{X} | \vec{\mu}; \sigma, K_M)/N$ and $f(\vec{\mu}) = P(\vec{\mu})$. By applying the Laplace approximation, we therefore obtain:

$$P(\mathbf{X} | M) \simeq \left(\frac{2\pi}{N}\right)^{\frac{K_M}{2}} \det(\hat{F})^{-\frac{1}{2}} \exp(\log P(\mathbf{X} | \vec{\mu}^*))P(\vec{\mu}^*) \quad (3.8)$$

Where $\hat{F} = -\langle \hat{H}\{\log P(x_i | \vec{\mu})\} |_{\mu^*} \rangle_e$ is the empirical Fisher Information Matrix:

$$\hat{F}_{i,j}(\vec{\mu}) = -\frac{1}{N} \sum_{x_i} \frac{\partial^2}{\partial \mu_i \partial \mu_j} \log P(x_i | \vec{\mu}) \quad (3.9)$$

If we take the logarithm of $P(\mathbf{X} | M)$:

$$\log P(\mathbf{X} | M) \simeq \log P(\mathbf{X} | \vec{\mu}^*) - \frac{K}{2} \log\left(\frac{N}{2\pi}\right) - \frac{1}{2} \log(\det \hat{F}) + \log P(\vec{\mu}^*) \quad (3.10)$$

by remembering that we are assuming that N is large and therefore discarding the terms that do not grow in N we get BIC:

$$\log P(\mathbf{X} | M) \simeq \log P(\mathbf{X} | \vec{\mu}^*) - \frac{K}{2} \log(N). \quad (3.11)$$

3.4 Corrections to BIC

There is a problem with the application of the Laplace approximation to the GMMs. As we said, the approximation only works if there is a single global maximum to the likelihood function. However, when $K \geq 2$, for every configuration of the means $\vec{\mu} = \{\mu_1, \mu_2, \dots\}$ there is at least one more configuration with the same likelihood, for example $\vec{\mu} = \{\mu_2, \mu_1, \dots\}$. In practice, this means that if we have two clusters of data, we can assign the first cluster to either the first component or to the second one, without changing the likelihood. This symmetry holds for every pair of means μ_i^* and μ_j^* : therefore, the likelihood function has $K!$ identical peaks, and only one of them is accounted for in the Laplace approximation. A first, naive solution is therefore to apply the approximation to each of the $K!$ peaks. Since for many configurations (i.e., when there are K distinct clusters of data) the peaks are well separated and all identical, we can simply use Laplace's result and multiply it by $K!$. The resulting approximation of the marginal likelihood is:

$$P(\mathbf{X} | M) \simeq K! \left(\frac{2\pi}{N} \right)^{\frac{K_M}{2}} \det(\hat{F})^{-\frac{1}{2}} \exp(\log P(\mathbf{X} | \vec{\mu}^*)) P(\vec{\mu}^*) \quad (3.12)$$

When considering $\log P(\mathbf{X} | M)$, this gives us a new “symmetry” term $S_{naive_M} = \log K_M!$ added to FIA. Notice that this new term is additive, not subtractive, and thus has the opposite sign compared to the dimensionality D_M : its effect is therefore that of damping the dimensionality penalty. Moreover, since $\log K!$ grows superlinearly in K , the difference $D_M - S_M$ grows sublinearly in K_M (for low values of K_M), a trend that matches qualitatively our observations for the average participant's behavior.

We first added this new term to BIC, instead of using the entire FIA; we will consider the complete expansion later in the chapter. The resulting “naive correction” description is the same as the “computation” one, with the addition of the new S_{naive_M} term (see [Supplementary Material](#)).

The ELPD for this new description marks a considerable improvement over the previous “computation” one. However, its predictive accuracy on unseen data does not match that of the “free penalty” description (Fig. 3.3). By examining the resulting β parameters (Fig. 3.4a), we can observe once again a segregation of participants into sublinear and superlinear response trends. The value β_S is positive for the majority of the participants, causing the S_{naive_M} term to contribute an overall sign opposite to that of D_M ; however, β_S is negative for a minority of them, resulting in the opposite effect. Moreover, unlike the “computation” description, the coefficient for dimensionality, β_D , is negative for some participants. These

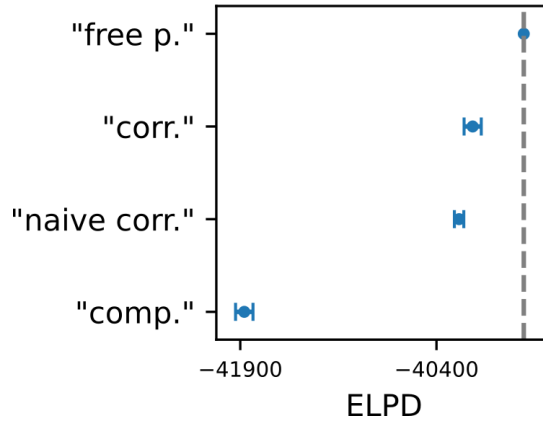


Fig. 3.3 **Comparison between “computational”, “free parameters”, and the two “correction” descriptions.** The x-axis shows the Expected Log Predictive Density (ELPD) of each description, while the errorbars represent the standard error of the ELPD differences with the best model, $ELPD_{\text{best}} - ELPS_M$.

negative values recast the influence of dimensionality from a penalizing factor to one that favors the selection of higher-dimensional models.

This finding aligns with observations from the “free penalty” description, specifically identifying the participants whose estimated dimensionality ($\hat{D}_M$) initially decreased before beginning to increase (a U-shaped trend). Because the new description incorporates two distinct terms (D_M and S_{naive_M}) that jointly influence the response to the number of parameters, the β values can now appropriately account for and frame this complex non-monotonic effect.

However, these increased flexibility and performance of the description come at the cost of a reduced overall interpretability. Before we had a clear penalizing effect coming from the dimensionality, easily identifiable in the positive values of the participants’ β_D . In the new description, instead, these values paint a more complex picture, one that, without the more complete view given by the “free penalty” $\hat{D}_M$, could lead us to interpret the dimensionality as a penalizing factor only for some participants.

To make more sense of these terms, we refined the definition of the S_{naive_M} term. As previously stated, the factor $K_M!$ only fully accounts for parameter permutation symmetry when the data genuinely contains K_M or more distinct, well-separated clusters. This assumption often holds if the true generating model has $K_{\text{true}} \geq K_M$ components. However, the correction breaks down when $K_{\text{true}} < K_M$. Consider a scenario where the true generating process has only $K_{\text{true}} = 1$ component, resulting in a single data cluster. If we attempt to fit a model with $K_M = 5$ components, the maximum-likelihood means ($\vec{\mu}^*$) will collapse: all five $\vec{\mu}^*$ will converge to the exact same, or extremely close, location. In this collapsed state, the

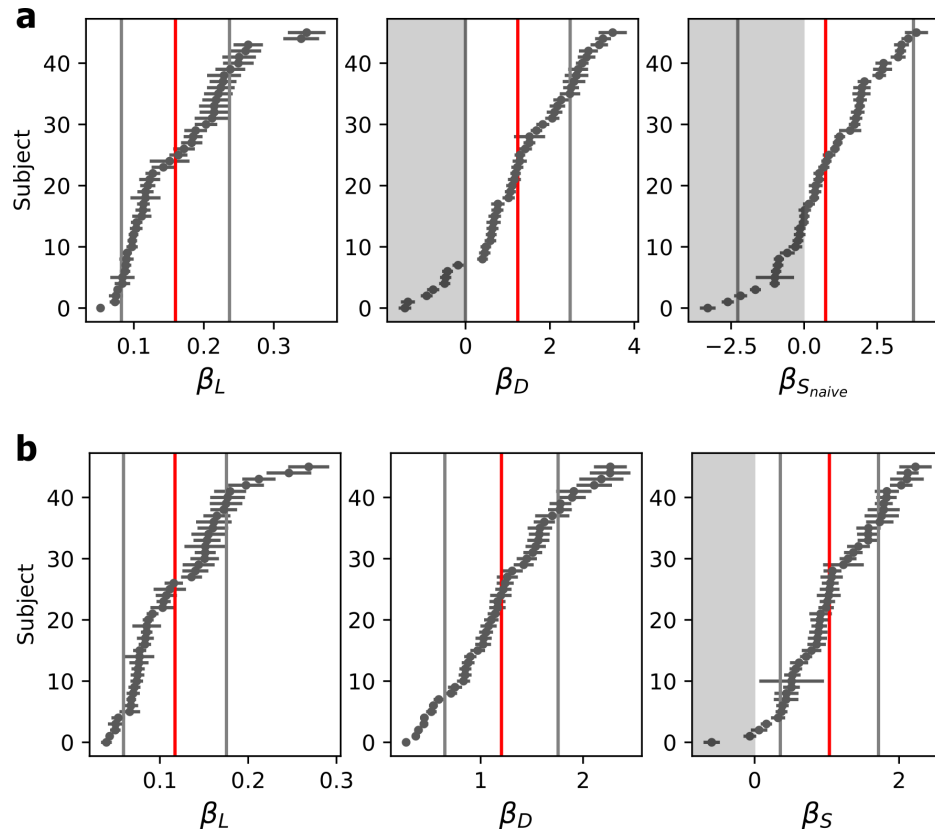


Fig. 3.4 Parameters of the two “correction” descriptions. **a.** Inferred sensitivities to likelihood (β_L), dimensionality (β_D) and naive symmetry ($\beta_{S_{naive}}$) for the “naive correction” description. For each plot, the red vertical line represents the posterior mean of the population average μ_β , while the two gray lines represent the population standard deviation σ_β . The dots represent the posterior mean of the values of each participant, and the errorbars are the s.e.m.. The values of all β are sorted. The shaded gray area indicates where the inferred values are negative. **b.** Same as for point **a**, but for the “correction” description.

symmetry effect is nullified. Permuting the component labels (the $\vec{\mu}^*$ vectors) does not lead to a distinct assignment of data points to clusters, and therefore, does not generate a new peak in the likelihood landscape. For the $K_M = 5$ and $K_{\text{true}} = 1$ example, we would observe only one peak in the likelihood, rendering the multiplication by $5!$ incorrect.

To properly account for this and all intermediate scenarios where only a subset of the means collapse, we must calculate the multiplicity of the means. Instead of a fixed $K_M!$ factor, the required correction involves determining the number of distinct means among the maximum-likelihood estimates. Specifically, we need to compute the multiplicity m_i for each distinct maximum-likelihood mean, μ_i . If five of the means converge to the same value, for instance, then that mean has a multiplicity of $m = 5$. This approach allows the symmetry correction to scale dynamically with the effective number of distinguishable components.

To compute the multiplicities, we set a threshold of similarity between two means, dependent on the distance between the means computed with the Expectation-Maximization algorithm and the number of points assigned to each mean (see [Supplementary Material](#)). The corrected number of identical maxima in the likelihood is therefore:

$$\mathcal{N}_{\text{peaks}} = \binom{K}{m_1, m_2, \dots} = \frac{K!}{m_1! m_2! \dots} \quad (3.13)$$

and the new correction obtained is $S_M = \log(\mathcal{N}_{\text{peaks}})$.

The resulting description, the “correction” description, is formally the same as the “naive correction” one, but with the modified term S_M . The ELPD is improved, even without adding any new term (Fig 3.3). Nevertheless, it still is not as high as the most flexible and purely descriptive “free penalty” description. Notice that the new definition of S_M has a value that doesn’t only depend on the model, but on the specific configuration of the datapoints. It is therefore describing the effect of the penalty (or rather the damping of the penalty) in a way that neither the “naive correction nor the “free penalty” description could, as both depended only on the model and not on the data, and thus were the same in all trials.

By examining the parameters of this description (Fig 3.4b), two key changes emerge when compared to the “naive correction”. First, the values of the dimensionality coefficients β_D are uniformly positive across participants. This result restores the more straightforward interpretation of dimensionality as a purely penalizing term, as in the case of the “computation” description. Second, the coefficients for the symmetry-related term (β_S) are positive for nearly all participants. When considering a varying multiplicity for the number of peaks, the behavioral differences among participants prove to be mostly quantitative (differing in the magnitude of the β_S value) rather than qualitative (differing in the sign of the β_S value).

3.5 Remainder term R_M

In this section, we consider the meaning of the remaining term of FIA, $R_M = -\frac{1}{2} \log(\det \hat{F})$, and its potential role in intuitive model selection. The observed Fisher Information Matrix $\hat{F}$ (Efron and Hinkley, 1978; Gelman et al., 2014a) is a $K_M \times K_M$ matrix that tells us how curved the (log) likelihood is at its peak. A high curvature indicates that the likelihood function is very sensitive to small changes in the parameter values. This implies that the data strongly favors a particular set of parameter values. On the other hand, a low curvature tells us that a wide range of values is nearly equally likely, resulting in less precise estimates. More specifically, the diagonal elements inform us on how precise the estimate is for a single parameter, while the off-diagonal ones indicate how intertwined the information is between two distinct parameters.

When inverting $\hat{F}$, as in the Laplace approximation, we obtain an estimate of the variance-covariance matrix of the parameter estimates. In the approximation, this matrix corresponds to an estimate of the variance of the likelihood around its peak. Since the Laplace approximation describes the shape of the peak as a multivariate Gaussian, the eigenvalues of the inverse of $\hat{F}$ indicate the principal axes of the K_M -dimensional ellipsoid associated with the Gaussian, while the square roots of its eigenvalues, if multiplied by $\sqrt{N}$, give the standard deviations along these directions.

When we limit the Laplace approximation to BIC, not only do we ignore the effect of the prior (that in our case simply becomes a correction to D_M , so it doesn't affect the overall description of participants' behavior in this experiment), but we treat each peak of the likelihood as if its width is the same along all directions. For our models, this is often not the case. If, for example, we have data generated by five components, four of which having the same mean, we will get two clusters with very different amounts of points. When we describe the shape of the likelihood curve for the model $K_M = 2$ with BIC, we are behaving as if the two clusters give the same information about the position of the two means, while the one with more points will arguably allow for a more precise estimate of it than the smaller one. Therefore, even if the likelihood and the dimensionality are the two main factors that decide between two models in most scenarios, adding this new term to the description of human behavior could improve it.

As previously noted, Piasini et al. (2025) used the Fisher Information Approximation to describe human choices. The specific models they employed, two-dimensional Gaussians with a parametric constraint on the mean position, allowed the use of Jeffreys' prior for the model parameters (Jeffreys, 1946). Unlike our experiment setup, their models' structure meant that this invariant prior mapped transparently onto the task structure. In contrast, our task's structure caused the uniform prior to be the only choice with this prerequisite. This

difference makes it so that the terms of the Laplace approximation are not the same in the two setups. Nevertheless, [Piasini et al. \(2025\)](#) demonstrated that human participants rely on these non-dimensionality definitions of complexity.

To investigate whether this reliance holds true in our task as well, we defined a new “FIA” description. This description incorporates all the terms derived from the Fisher Information Approximation: the maximum log-likelihood (L_M), the dimensionality (D_M), the symmetry term (S_M), and the remainder term (R_M). The β_R coefficient associated with this new R_M term shares the same population parameters and priors as the other β parameters within our Bayesian regression framework (see Section 3.7).

The addition of the R_M term further improves the predictive power of the description, though the magnitude of this effect is more limited compared to previous additions (Fig. 3.5b). The values of β_R reveal a familiar pattern of high participant heterogeneity: the average sensitivity to the R_M term is close to zero, as participants exhibit a wide range of qualitatively different responses, with some showing $\beta_R < 0$ and others showing $\beta_R > 0$ (Fig. 3.5a). One possibility to explain these values is that, indeed, there is a wide range of qualitatively different strategies. Some participants penalize models with a high R_M , i.e., models whose parameters are strongly determined by the data: when R_M is big, it usually is a sign of overfitting, as perturbing the parameters leads to big changes in the likelihood. Some other participants, instead, favor models with high R_M , i.e., models that allow them to more precisely estimate the latent causes, suggesting a preference for models where the data strongly dictates the parameter values. Finally, some participants demonstrate minimal or no reliance on this specific property of the models in their decision-making.

Even if this span of strategies could be reasonable, there is a different interpretation for the values of β_R : as it was the case with S_{naive} , the term we used might be misspecified or incorrectly defined, and thus leading to noisy results as it tries to make up for missing sources of variability, or for a different definition of the penalty. The reason to believe so is once again to be found in the nuances² of the Gaussian Mixture Models. As we said, the Laplace approximation only works when the function has one isolated peak that grows as $\exp(N\psi(\theta))$, since it can be approximated as a Gaussian. While this is never the case with GMMs with $K_M \geq 2$, since they have more than one peak, we saw that the problem can be solved in most cases (i.e., when the maxima of the likelihood are distant from each other) by multiplying by the appropriate factor. In these cases, the R_M term is well defined and properly describes the likelihood integral. However, when the peaks are close (i.e., when some μ parameters have similar values), they start to interact with each other, losing more and more of the Gaussian-like shape the closer they get (Fig 3.6b). More in detail, when two peaks get

²or rather nuisances

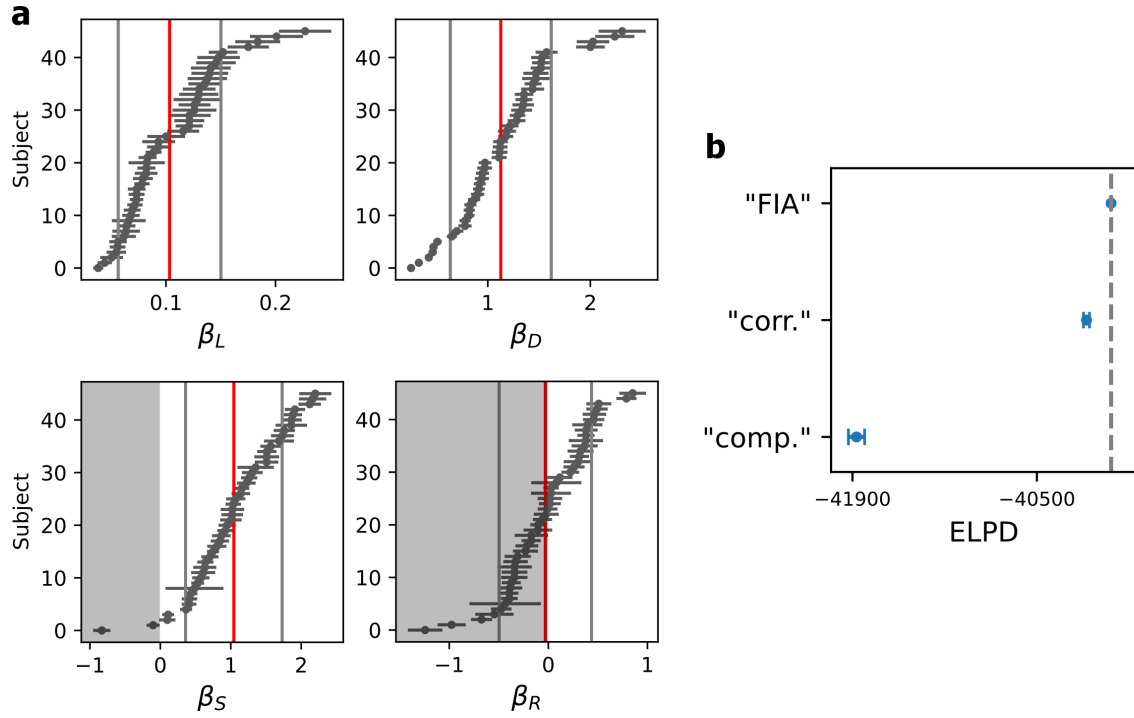


Fig. 3.5 Parameters and comparison of the “FIA” description. **a.** Inferred sensitivities to likelihood (β_L), dimensionality (β_D), symmetry (β_S) and remainder term (β_R) for the “FIA” description. For each plot, the red vertical line represents the posterior mean of the population average μ_β , while the two gray lines represent the population standard deviation σ_β . The dots represent the posterior mean of the values of each participant, and the errorbars are the s.e.m.. The values of all β are sorted. The shaded gray area indicates where the inferred values are negative. **b.** Result of the LOO CV comparison between the BIC “computation” description, the “correction” description and the “FIA” description. The x-axis shows the Expected Log Predictive Density (ELPD) of each description, while the errorbars represent the standard error of the ELPD differences with the best model, $ELPD_{\text{best}} - ELPD_M$.

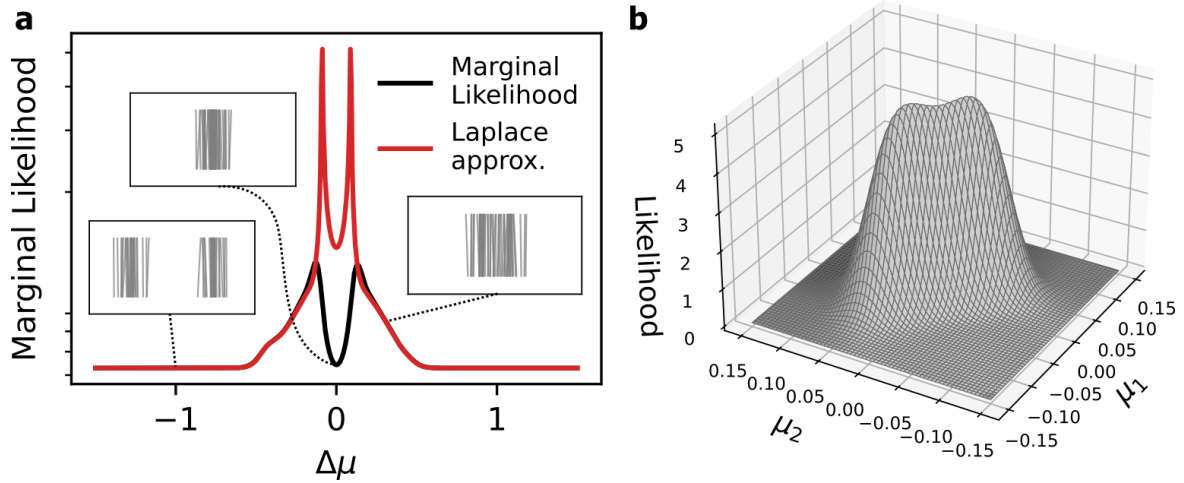


Fig. 3.6 Failure of Laplace approximation for close components. **a.** Comparison between the marginal likelihood and the Laplace approximation for a model with $K_M = 2$ on data from two clusters at varying distances. On the x-axis, the distance between the means of the two clusters. The inserts show some of the resulting compositions. In black, the marginal likelihood computed numerically. In red, the corresponding Laplace approximation up to the $O(1)$ term, multiplied by 2 (to account for the presence of two peaks in the likelihood landscape). The y-axis is in logarithmic scale. Notice that, since we simply multiply by 2, here we are using the S_{naive_M} term. The approximation fails when the clusters become too close. Interestingly, the approximation starts working better again in the proximity of $\Delta\mu = 0$, particularly if we remove the $\times 2$ factor, thus applying the S_M term. **b** Likelihood landscape for the clusters in **a** at $\Delta\mu = 0.1$, i.e., where the Laplace approximation explodes. The two peaks generated by the symmetry are merging, thus creating a ridge. The y-axis scale factor is 10^{18} . Note that only the landscape around the peak is shown.

very close, they tend to form a ridge that connects them: this ridge, being approximately flat, has an extremely small second derivative, leading to very large eigenvalues of the covariance matrix. These eigenvalues, and therefore the determinant of the Hessian, are usually much bigger than the actual widths of the bell (or rather, of the mixture of bells), thus breaking the approximation. In other words, we can't trust the value of R_M when two means are getting close (Fig. 3.6a). As we discussed, this is a very common scenario for our models, making the R_M unreliable for many trials, for at least some of the models.

It is worth noticing, however, that this is only true when the means are approaching, and the approximation improves again when the means are overlapping (Fig 3.7). In these areas, accounting for the number of peaks gives a value close to the true marginal likelihood, confirming the necessity for the S_M term.

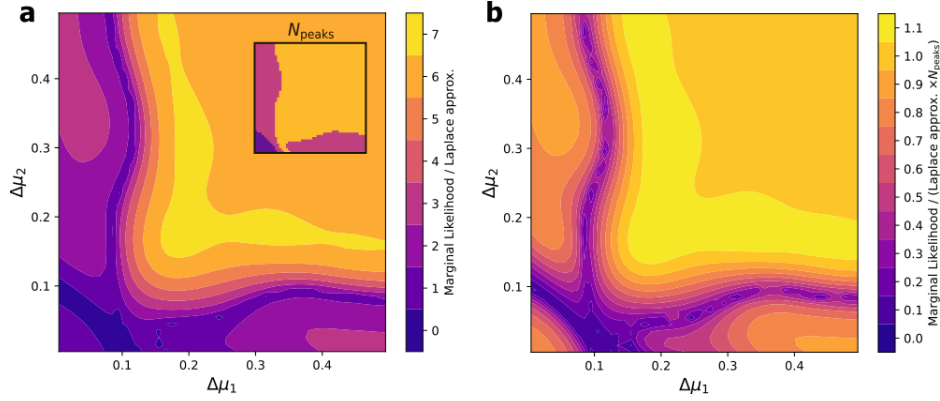


Fig. 3.7 Comparison between the marginal likelihood and Laplace approximation. a. Ratio between the marginal likelihood, integrated numerically, and the Laplace approximation for a model with $K_M = 3$ on data from three clusters at varying distances. The distances are always computed relatively to a central cluster. As in Fig 3.6a, the three clusters are always the same and only their relative position changes. On the x-axis, the distance between the central cluster and the cluster on the left. On the y-axis, the distance between the central cluster and the cluster on the right. The distances are shown up to a maximum of $\Delta\mu = 0.5$, even if the values could go as far as $\Delta\mu = 2$, to focus on the cases of close clusters. The inset shows the value of $\mathcal{N}_{\text{peaks}}$ for each value of $\Delta\mu_1$ and $\Delta\mu_2$. The colors follow the same scale as the main figure, indicating six (yellow), three (purple), or one (dark purple) peaks. **b.** Same as **a**, but the Laplace approximation is multiplied by $\mathcal{N}_{\text{peaks}}$.

3.6 Discussion

This study investigated the cognitive mechanisms underlying the human simplicity bias, focusing on the shape and composition of the complexity penalty used in intuitive model selection. Building on the finding that dimensionality strongly influences human choices (Piasini et al., 2025), we investigated potential biases in its perception. We found, on average, that the number of parameters appears to be perceived sublinearly by human participants. Our central finding is that this sublinearity is not a simple perceptual artifact but an emergent property of participants' sensitivity to the internal symmetries inherent in the models.

Our analysis began by swapping the dimensionality penalty for a free function $f(K_M)$ inferred from the data. This function showed an average sublinear trend across the population. However, this observation presented two challenges. First, contrary to our findings in the previous chapter, it could not be readily explained by known perceptual biases of numerosity, as the model dimensionalities (1–5) fall largely within the subitizing range, where perception is typically accurate and linear (Kaufman et al., 1949). Second, the high degree of inter-subject variability, with some participants exhibiting superlinear or even non-monotonic trends, suggested the lack of a single, uniform cognitive strategy. However, this heterogeneity

might have also been a sign that the “free penalty” description, while predictively powerful, is capturing the effects of a missing, more fundamental factor.

By accounting for the permutation symmetries of the Gaussian Mixture Model components, we were able to introduce a new term that solved both issues. This theoretically-motivated symmetry correction, S_M , improved the model predictive accuracy on unseen data (compared to a description without the term) while also resulting in a homogeneous description of participants’ strategies. The ambiguous superlinear or non-monotonic behaviors were not evidence of fundamentally different cognitive approaches, but rather artifacts of a misspecified description that mistook participants’ sensitivity to symmetry for a distorted dimensionality penalty. With the inclusion of the S_M term, nearly all participants converged on a cohesive strategy: a clear penalty for dimensionality ($\beta_D > 0$) combined with a consistent accounting for model symmetry ($\beta_S > 0$). Finally, the addition of the log-likelihood curvature (or *remainder*) term, R_M , provided a further, albeit modest, improvement in predictive performance. However, its inclusion reintroduced the issue of high inter-subject variability, with sensitivities (β_R) distributed roughly symmetrically around zero. We interpret this not as evidence for a wide range of participants’ strategies regarding model curvature, but as an indication of a mismatch between the formal term and the cognitive heuristic. Given the mathematical instability of FIA for GMMs with overlapping components, the R_M term is often unreliable, and its variable effect on participants’ descriptions likely reflects fitting noise rather than a consistent cognitive process.

Taken together, these findings have multiple implications for our understanding of intuitive model selection. First, our work strongly supports the conclusion that human simplicity biases are adaptive and model-specific. In concert with the “Boundary term” identified in [Piasini et al. \(2025\)](#), our “Symmetry term” suggests that the cognitive computation of complexity is not a fixed, task-agnostic function like those found in AIC or BIC. Instead, the brain appears to tune its complexity penalties to the structural features of the problem at hand. While our previous work on the effect of sample size indicated that participants do not perform a full Bayesian integration, the current results show that their strategies are nevertheless sensitive to key properties of the underlying likelihood landscape, pointing towards the use of sophisticated heuristics that approximate features of a Bayesian calculation without performing it explicitly.

Despite the clarity of these results, several limitations point toward the need for further research. The most critical next step is to test the symmetry hypothesis directly. A future study should employ models with a similar range of dimensionalities but without permutation symmetries; in such a task, we would predict the sublinearity in dimensionality and the explanatory power of the S_M term to disappear. Furthermore, definitive conclusions about

human sensitivity to likelihood curvature cannot be drawn from this task. Future work could use models with more stable likelihood geometries to probe this specific aspect of complexity. Finally, while fixing the sample size reduced the influence of numerosity perception, it did not eliminate all perceptual factors. The interplay between the number of points in a cluster and its perceived width, the perception of relative sizes of clusters, the way points are intuitively assigned to close clusters, as well as fluctuations in the internal value of the expected width σ are all, among others, factors to be considered in future investigation.

In conclusion, this study reframes our understanding of the dimensionality penalty in human model selection. We have shown that an apparent and ambiguous sublinear bias is, in fact, a rational cognitive adaptation to the symmetries of the models under consideration. The human simplicity bias is not a blunt, inflexible rule but a versatile and adaptive cognitive tool, sensitive to the structure of the problem at hand.

3.7 Supplementary Material

3.7.1 Experimental Data Analysis

The Bayesian regression details of the descriptions used in this chapter are very similar to the ones of Chapter 3. Since the distribution and structure have already been presented there, what follows is only a brief recap.

The **BIC “computation” description** is exactly the same as the one in Chapter 3. The fixed value $N = 50$ doesn’t affect the structure of the description. The probability for a participant to choose a model M for a certain trial will be given by:

$$P(\text{choice} = M \mid X) = \frac{\varepsilon}{5} + (1 - \varepsilon) \frac{\exp(\beta_L L_M - \beta_D D_M)}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m - \beta_D D_m)} \quad (3.14)$$

where ε is the lapse rate, β_{term} the sensitivity to the relative term, and $\mathcal{M}$ the set of five models the participants could choose from.

The **“free penalty” description** is the same as the “computation” description for what concerns sensitivity to the goodness-of-fit β_L and the relative population parameters μ_{β_L} and σ_{β_L} , and the lapse ε and the relative population parameters $\mu_{\logit \varepsilon}$, $\sigma_{\logit \varepsilon}$ and δ . The sensitivity parameter β_D and the relative population parameters μ_{β_D} and σ_{β_D} were treated differently, and are related to the definition of $\hat{K}_M$.

As mentioned in the main text, the dimensionality D_M is substituted here with a free dimensionality $\hat{D}_M = \hat{K}_M \frac{\log 50}{2}$, where $\hat{K}_M$ is the perceived number of parameters (mixture components). Using a free $\hat{K}_M$ in the definition of the description would result in non-

identifiability problems. This becomes evident when considering the probability for a participant to choose a model M for a certain trial in this description:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L L_M - \beta_D \hat{K}_M \frac{\log(50)}{2})}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m - \beta_D \hat{K}_m \frac{\log(50)}{2})} \quad (3.15)$$

Adding the same shift c to all the models would not change the probability:

$$P(\text{choice} = M \mid X) = \frac{\exp(\beta_L L_M - \beta_D \hat{K}_M \frac{\log(50)}{2} + c)}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m - \beta_D \hat{K}_m \frac{\log(50)}{2} + c)} \quad (3.16)$$

$$= \frac{\exp(c) \exp(\beta_L L_M - \beta_D \hat{K}_M \frac{\log(50)}{2})}{\exp(c) \sum_{m \in \mathcal{M}} \exp(\beta_L L_m - \beta_D \hat{K}_m \frac{\log(50)}{2})} \quad (3.17)$$

$$= \frac{\exp(\beta_L L_M - \beta_D \hat{K}_M \frac{\log(50)}{2})}{\sum_{m \in \mathcal{M}} \exp(\beta_L L_m - \beta_D \hat{K}_m \frac{\log(50)}{2})} \quad (3.18)$$

Therefore, for a certain value of the terms $\hat{K}_M$, the same probability will be given by $\hat{K}_M + \tilde{c}$, as long as there is the same difference between all the $\hat{K}_M$. To overcome this non-identifiability issue, we fixed $\hat{K}_1 = 1$.

A similar problem arises when considering the product between $\hat{K}_M$ and β_D . For a certain choice of β_D and $\hat{K}_M$, the same probability is obtained by multiplying β_D by c and dividing $\hat{K}_M$ by the same value. To solve the issue, instead of fixing another value of $\hat{K}_M$, we considered $\beta_D = \hat{K}_2 - \hat{K}_1$.

The priors of the new parameters are set to:

$$\mu_{\hat{K}_M} \sim \mathcal{N}(\mu = K_M, \sigma = 1) \quad \forall M : K_M > 1 \quad (3.19)$$

$$\sigma_{\hat{K}_M} \sim \text{Exponential}(\text{scale} = 1) \quad \forall M : K_M > 1 \quad (3.20)$$

$$\hat{K}_M^s \sim \mathcal{N}(\mu = \mu_{\hat{K}_M}, \sigma = \sigma_{\hat{K}_M}) \quad \forall s, M : K_M > 1 \quad (3.21)$$

$$\mu_{\beta_D} = \mu_{\hat{K}_2} - 1 \quad (3.22)$$

$$\sigma_{\beta_D} = \sigma_{\hat{K}_2} \quad (3.23)$$

$$\beta_D^s = \hat{K}_2^s - 1 \quad (3.24)$$

Where s indicates the participant.

The “**naive correction**”, “**correction**”, and “**FIA**” descriptions are structurally the same as the “**computation**” description. Similarly to the other terms, the new S_{naive_M} , S_M , and R_M terms have been standardized to a mean of zero and a standard deviation of one before fitting the description.

The priors for the sensitivity of the standardized terms are:

$$\hat{\mu}_{\beta_{\text{term}}} \sim \mathcal{N}(\mu = 0, \sigma = 10) \quad \forall \text{ terms} \quad (3.25)$$

$$\hat{\sigma}_{\beta_{\text{term}}} \sim \text{Exponential}(\text{scale} = 1) \quad \forall \text{ terms} \quad (3.26)$$

$$\hat{\beta}_{\text{term}}^s \sim \mathcal{N}(\hat{\mu}_{\beta_{\text{term}}}, \hat{\sigma}_{\beta_{\text{term}}}) \quad \forall \text{ terms}, s \quad (3.27)$$

.

3.7.2 Multiplicity threshold of the Symmetry term

As we discussed, if two clusters of data are far enough, switching two means in the likelihood landscape generates two symmetric peaks. We want to define a threshold that indicates when these two peaks are close enough to be considered as merging in a single peak.

The peaks are symmetric around the diagonal $\mu_i = \mu_j$. Their distance will therefore be $d_{ij} = \sqrt{2(\mu_i - \mu_j)^2}$. To consider them close, we want this distance to be smaller than a proportion α of their width along that direction. If their summed width along said direction is bigger than the distance, we can consider the peaks to be merging. Assuming that far enough peaks are gaussian-like, along each direction i corresponding to a mean μ_i , the width of the peak will be approximately $\sigma / \sqrt{N_i}$. Therefore, the half-width of the peak along the direction perpendicular to $\mu_i = \mu_j$ is $\sqrt{(\sigma^2/N_i + \sigma^2/N_j)/2}$. For simplicity, we ignore the deformation of the peaks that we get from close clusters: one could try to account for those by using the inverse of $\hat{F}$, but as we discussed, the Gaussian approximation obtained with it is not reliable, particularly when the peaks are close.

To check if two peaks are far enough to be considered separate, therefore, we check if $\sqrt{2(\mu_i - \mu_j)^2} \geq 2\alpha\sigma\sqrt{(1/N_i + 1/N_j)/2}$, which simplifies to:

$$|\mu_i - \mu_j| \geq \alpha\sigma\sqrt{\frac{1}{N_i} + \frac{1}{N_j}}. \quad (3.28)$$

Where we set $\alpha = 0.7$. The values of N_i where obtained as $N_i = \sum_j \gamma_{ji}$ where γ_{ji} is the responsibility of component i for the datapoint x_j . The responsibility is obtained from the Expectation Maximization algorithm ([Dempster et al., 1977](#)).

Chapter 4

Seeing what you hear: compression of rat visual perceptual space by task-irrelevant sounds

This chapter is a reproduction of the paper:

Zanzi M.*, **Rinaldi F. G.***, Fornasaro S., Piasini E., Zoccolan D. (2025). “Seeing what you hear: Compression of rat visual perceptual space by task-irrelevant sounds.” PLoS Comput Biol, 21(10). doi:[10.1371/journal.pcbi.1013608](https://doi.org/10.1371/journal.pcbi.1013608). (* denotes equal contribution)

Statement of Contribution:

Mattia Zanzi and I contributed equally to this work as co-first authors. My primary role was the conceptualization of the ideal observer models, the execution of formal Bayesian analyses, and the estimation of sound intensities. Mattia Zanzi’s contribution included designing the behavioral task, collecting the data from animal subjects, and performing the statistical analysis presented in Figure 4.3 and Tables 4.1 and 4.2. We wrote the manuscript, developed the figures, and responded to reviewer comments jointly, together with the other authors and under the supervision of Davide Zoccolan and Eugenio Piasini.

4.1 Introduction

To provide a coherent and reliable representation of the sensory environment, the information collected by our senses needs to be combined in multimodal percepts. The traditional view is that the process of multimodal integration takes place in high-order association cortices, after sensory signals have been independently and extensively processed by unimodal cortical regions. This view is supported by many psychophysical and neurophysiological studies,

showing that multimodal representations in such regions often combine unimodal inputs in a statistically optimal way, increasing the accuracy of perceptual judgements (Ernst and Banks, 2002; Ernst and Bühlhoff, 2004; Fetsch et al., 2009, 2013; Gleiss and Kayser, 2012; Morgan et al., 2008; Nikbakht et al., 2018; Noppeney, 2021; Raposo et al., 2012, 2014; Seilheimer et al., 2014; Shams and Beierholm, 2010).

More recent work, however, has shown that cross-modal interactions already occur at the level of primary sensory cortices (Bizley et al., 2016). Imaging studies in humans, as well as single-unit recordings in monkeys have shown that neuronal responses in auditory cortex are modulated by both visual and somatosensory inputs (Kayser and Logothetis, 2007; Kayser et al., 2005, 2008, 2010; Schroeder and Foxe, 2005). Conversely, sound-driven responses and facilitation of peak latencies under audiovisual stimulation have been reported in human primary visual cortex (V1) (Martuzzi et al., 2007) and auditory-mediated enhancement of BOLD responses in V1 appears to partially account for the sound-induced flash illusion (Watkins et al., 2006) — the illusory perception of multiple light flashes when a single flash is paired to multiple, closely spaced beep sounds (Hirst et al., 2020; Keil, 2020; Shams et al., 2000).

These early cross-modal interactions could be mediated by the direct projections from auditory cortices to V1 that have been reported in human and monkey neuroanatomy studies (Clavagnier et al., 2004; Eckert et al., 2008; Falchier et al., 2002; Rockland and Ojima, 2003). In primates, however, these projections are sparse, originate mainly from parabelt areas of auditory cortex and target mainly peripheral V1. Denser cortico-cortical connections from primary auditory cortex (A1) to V1 have been instead found in rodents (Deneux et al., 2019; Ibrahim et al., 2016; Iurilli et al., 2012; Oude Lohuis et al., 2024; Stehberg et al., 2014). Given the accessibility of these model systems to the genetic dissection of neural circuits (Luo et al., 2008; Panzeri et al., 2017), several authors have investigated the impact of A1 inputs to V1 on visual cortical dynamics and encoding of visual information. Surprisingly, the conclusions of these studies are highly heterogeneous and often contrasting.

Sound was originally reported to hyperpolarize supragranular pyramidal cells in mouse V1, by recruiting a translaminar GABAergic network activated by A1 projections to infragranular V1 neurons (Iurilli et al., 2012). Later studies, however, found that most A1 projections terminate in superficial layers of V1 and engage local inhibitory and disinhibitory circuits underlying a much richer variety of auditory influences on V1: enhancement and suppression of visually evoked responses have both been documented, as well as sharpening of orientation tuning and improvement in the coding of visual information (Deneux et al., 2019; Ibrahim et al., 2016; McClure Jr and Polack, 2019; Meijer et al., 2017; Williams et al., 2023). This heterogeneous assortment of sound-mediated modulations in V1 seems to depend

on factors such as the contrast of the visual stimuli, the luminosity of the environment, the spectral (e.g., pure tones vs. noise bursts) and envelope (e.g., loud vs. quiet onset) properties of the sounds and the temporal congruency between visual and auditory stimuli. Finally, it remains debated the extent to which, in awake mice, sound-driven responses in V1 reflect auditory inputs or, rather, behavioral modulation via sound-evoked orofacial movements (Bimbard et al., 2023; Oude Lohuis et al., 2024; Williams et al., 2023).

Amid such contrasting findings at the cortical circuitry level, the functional impact of auditory signals on rodent visual perception has been poorly explored. Does sound improve acuity in visual discrimination tasks? Can any signature of sound-induced suppression (or enhancement) of visual representations be found at the behavioral level? Is the temporal consistency between sounds and visual stimuli important in this regard? The goal of our study was to address these questions, by measuring the extent to which task-irrelevant sounds can affect the perceptual choices of rats engaged in a visual classification task.

4.2 Results

In our study, we trained 12 rats in a visual temporal frequency (TF) classification task, where the visual (V) stimuli to categorize were paired with simultaneous but task-irrelevant auditory (A) stimuli (Fig 4.1). The visual stimuli were circular sinusoidal gratings (see examples in Fig 4.1A,B) moving outward with 9 possible, randomly selected TFs (Fig 4.1C; x axis). Rats categorized these stimuli as “Low TF” or “High TF”, relative to an ambiguous boundary value (2.12 Hz) in the middle of the tested TF range. The auditory stimuli consisted of either fixed amplitude or amplitude modulated (AM) white noise bursts (Fig 4.1D; see also A.2 Fig for details). The sounds were uninformative about the correct classification of the visual stimuli, to avoid high-level multisensory integration effects and focus instead on the direct impact of auditory signals on the representation and perception of visual stimuli. Following an approach we have established in previous studies (Caramellino et al., 2021; Matteucci et al., 2021), the rats had to lick the central response port in an array of three to trigger stimulus presentation and then lick either the left or the right port to choose the “Low TF” or the “High TF” category (with the stimulus-response association being counterbalanced across animals), so as to collect liquid reward in case of correct classification (Fig 4.1B; see the Materials and Methods for details).

During the training phase, the gratings were paired with a constant amplitude sound - i.e., a white noise burst sampled at 44.1 kHz. (Fig 4.1D, green trace). As explained in the Materials and Methods, rats were gradually introduced to the visual TFs they had to classify, starting from the most discriminable ones, i.e., 0.25 vs. 4 Hz (Fig 4.2; Training Phase 2).

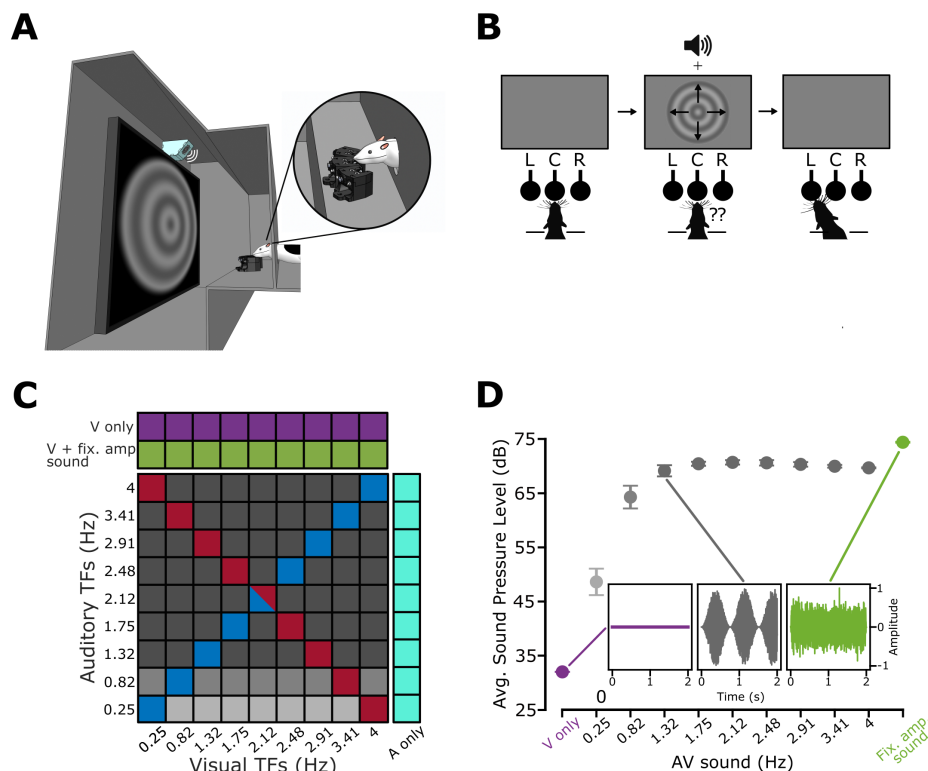


Fig. 4.1 Experimental setup and audio-visual stimulus conditions. **A.** Rendering of a rat extending its head through a viewing hole in the wall of a sound-isolated operant box, equipped with a display and a speaker for presentation of the audio-visual stimuli. The hole was placed in front of the stimulus display and the speaker (in cyan). The hole also allowed access to three nose-poke ports used to trigger stimulus presentation and report the temporal frequency of the visual stimuli. **B.** Schematic representation of a behavioral trial. As in previous studies (Caramellino et al., 2021; Matteucci et al., 2021), the rats learned to protrude their head through the viewing hole and trigger stimulus presentation by touching the central nose-poke port. The visual stimuli were circular, outward-moving sinusoidal gratings that rats classified according to their temporal frequency (TF). Gratings with TF < 2.12 HZ had to be classified as “Low TF”; those with TF > 2.12 Hz as “High TF”. Each class was associated with one of the two lateral nose-poke ports. Auditory stimuli did not provide any information about the correct classification of the visual stimuli. **C.** Stimulus conditions presented to the rats during the test phase of the experiment. These included: 1) visual gratings paired with fixed amplitude sounds (green); 2) visual gratings paired with AM sounds (cells of the square matrix: blue, gray and red refer, respectively, to congruent, incongruent and anti-congruent audiovisual conditions; the shades of gray, from light to dark, indicate AM sounds with progressively larger intensity; see Methods and A.2 Fig for details); 3) unimodal purely visual gratings (purple); and 4) unimodal purely auditory AM sounds (cyan). **D.** Details on the auditory stimuli. Green dot: white noise burst with a fixed maximal amplitude. Gray dots: amplitude-modulated white noise bursts, each with a different temporal frequency of the sinusoidal envelope. Purple dot: absence of an auditory stimulus. See Methods and A.2 Fig for details on estimation of average perceived sound intensity. Error bars: s.e.m. across rats. For the modulated noise bursts, different shades of gray are used to group stimuli based on their average sound intensity (same as in C). Insets: sound amplitude control signal for white noise burst (green), the absence of auditory stimulus (purple) and one example of modulated noise burst (gray).

Once they had achieved a performance higher than 70% correct in at least two consecutive sessions, gratings with additional TFs were progressively added through Training Phases 3-5, until, in Training Phase 5, the animals had to classify the full range of visual TFs. Responses to the 2.12 Hz boundary stimulus were randomly rewarded. Two of the rats failed to achieve the performance criteria in Training Phase 2, after a total of 28 sessions in Phase 1 and 2, and were therefore excluded from further testing. The contrast of the gratings was initially set at 100%, but it was progressively reduced to 25% (Fig 4.2; gray patches), given that the impact of sound on V1 responses has been found to be stronger when visual stimuli are shown at low/intermediate contrast (Ibrahim et al., 2016; Williams et al., 2023). As illustrated in Fig 4.2 for two example rats, the contrast reduction procedure was applied several times, to ensure that the animals would successfully classify the gratings' TFs at both the lowest and full contrast. More specifically, our procedure included: 1) an early exploratory stage, where contrast was quickly reduced, within 6-10 sessions, from 100% to values as low as 15% (e.g., see sessions 39-40 in Fig 4.2A, top); 2) a re-exposure to full contrast in case the performance had dropped to chance following a step of contrast reduction (see sessions 40-46, following the accuracy drop in session 39, in Fig 4.2B, top); and 3) one last re-exposure to full contrast before the final, stable reduction to 25%. Rats moved quickly through the training phases, occasionally displaying some drop of discrimination accuracy after transitioning from one phase to the next, which, in some cases, required moving them back to the previous phase (e.g., example rat 2 went back to Phase 4 for one session after a substantial performance drop following 5 sessions in Phase 5; see Fig 4.2B). These temporary drops, lasting a few sessions, are typical for rodents that are consolidating the acquisition of a perceptual task (Zoccolan and Di Filippo, 2018) and are likely due to fluctuations in the level of motivation, deployment of exploratory strategies (Pisupati et al., 2021), or temporary failures of the behavioral rig (e.g., uneven delivery of liquid from the reward ports), as suggested by the often concomitant decrease in the number of behavioral trials completed by the animals (see Fig 4.2B, bottom). Nevertheless, all the animals eventually achieved the criterion performance of 70% correct choices with 25% contrast gratings that was necessary to advance them to the test phase.

4.2.1 Task-irrelevant sounds affect rat perception of drifting gratings in a way that is independent of the temporal congruency between the auditory and visual stimuli

Once the animals had become proficient in the TF classification task with 25% contrast gratings, they were moved to the testing phase, where they faced a rich variety of audiovisual stimuli (Fig 4.1C). The gratings were paired not only with the fixed amplitude burst used

during training (Fig 4.1C, green cells), but also with AM white noise stimuli. These sounds were obtained by modulating the amplitude of a white noise burst with sinusoids having 9 possible TFs (the same of the visual gratings). The rats were presented with all possible pairwise combinations of visual and auditory TFs (cells in the squared matrix of Fig 4.1C). This yielded trials where the auditory and visual stimuli had either the same TF (blue cells) or different TFs (gray and red cells). The latter included a special pool of anti-congruent conditions (red cells), where progressively larger visual TFs were paired with progressively smaller auditory TFs. Additionally, we also tested unimodal, purely visual trials (purple cells) and unimodal, purely auditory trials with AM sounds (cyan cells). Since the rats were never trained to classify the TFs of the auditory stimuli, with the latter trials we simply measured the spontaneous responses of the animals to the AM sounds, without providing any feedback about their choices (see Material and Methods). The rats did not spontaneously classify the AM bursts according to their TF in the auditory only conditions, thus confirming the task-irrelevance of the sounds (see A.1 Fig).

Overall, this design allowed probing the impact of task-irrelevant auditory stimuli on visual perception along two distinct dimensions: 1) whether sound alters the sensitivity of visual discrimination or, rather, introduces a bias in visual classification; and 2) whether the impact of sound depends on the temporal consistency of audio-visual stimuli or, rather, on the intensity of the auditory stimuli.

When the gratings were paired with the fixed amplitude noise, the group average psychometric curve reporting the proportion of “high TF” choices as a function of the grating frequency was sharp and symmetrical, with the point of subjective equality sitting squarely on the ambiguous TF (Fig 4.3A, green). Interestingly, pairing the gratings with either the congruent (blue) or anti-congruent (red) AM sounds produced an identical increase in the proportion of “high TF” choices, which was particularly prominent for the highest TFs. This observation was confirmed by a 2-way ANOVA having as factors the audiovisual experimental condition and the grating TF (see the Materials and Methods). Both factors significantly modulated rat choices and so did their interaction (Table 4.1), thus confirming that the increase of “high TF” choices yielded by the modulated sounds was not homogeneous along the TF axis. A post-hoc Tukey test confirmed that the curves observed for the two conditions featuring AM sounds were statistically indistinguishable. This indicated that the congruency between the TF of the AM sounds and the TF of the gratings did not affect the impact of the auditory stimuli on the visual discrimination task.

This observation led us to group together all the trials where the visual stimuli were paired with an AM sound, regardless of the congruency of their TFs. The resulting psychometric curve (Fig 4.3B, black) featured an asymmetrical vertical shift, with respect to the reference

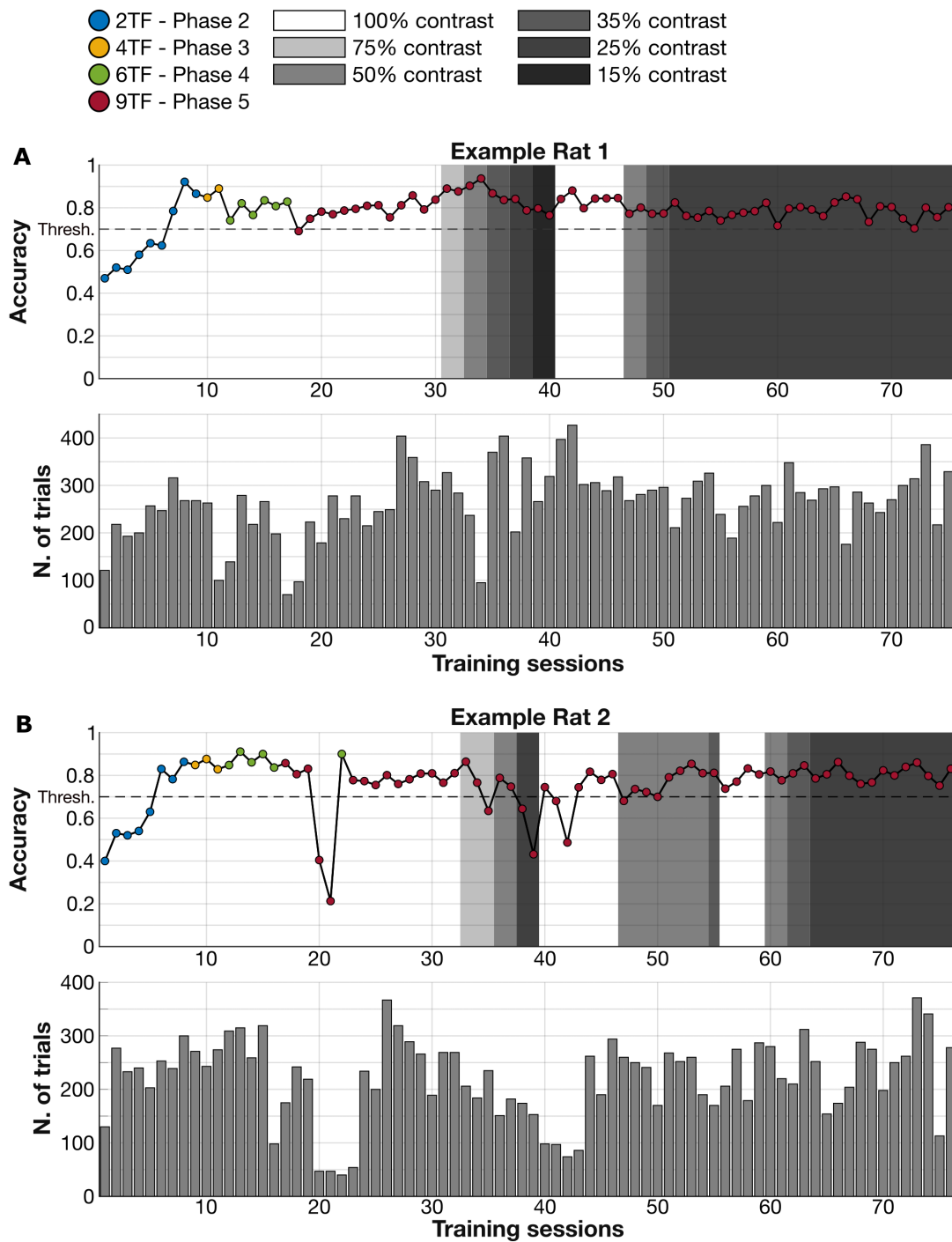


Fig. 4.2 Discrimination accuracy of two example rats across the different phases of the training procedure. **A** The line plot (top) shows the percentage of correct choices of example rat 1 in classifying the temporal frequency of the drifting gratings across more than 70 daily training sessions. The color of the dots indicates the training phase (from 2 to 5; see caption), while the intensity of the gray patches denotes the various stages of contrast reduction that were performed during the training (see legend and the Materials and Methods). The dashed horizontal line represents the criterion accuracy (70%) required to progress to the next training phase. To advance, rats had to achieve performance above criterion for at least two consecutive days within the same training phase. The bar plot (bottom) shows the number of behavioral trials that were collected for the rat in each daily session. **B.** Same as **A**, but for example rat 2.

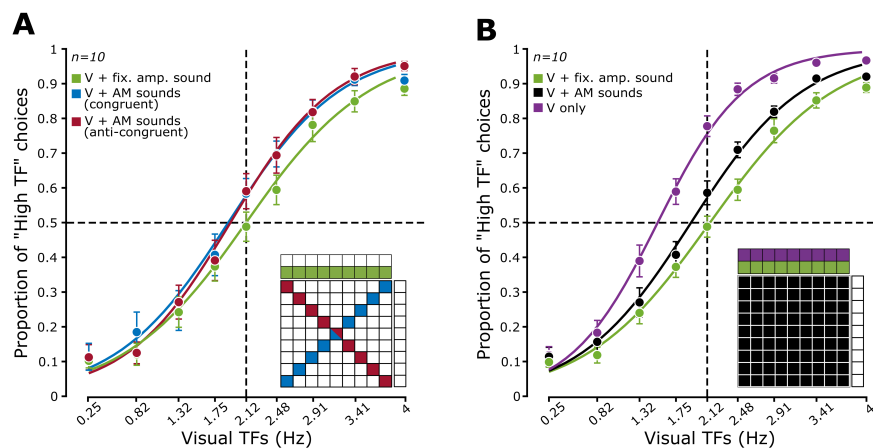


Fig. 4.3 Task-irrelevant sounds strongly modulate rat classification of the temporal frequency of drifting gratings. **A.** Group average proportion of “high TF” choices ($n=10$ rats) as a function of the TF of the visual gratings, when the latter were paired with: 1) fixed amplitude sounds (green); 2) congruent AM sounds (blue); and 3) anti-congruent AM sounds (red). The error bars denote s.e.m. over the group of 10 rats. The curves are the result of logistic regressions to the data points. The inset illustrates which of the stimulus conditions shown in Fig 4.1C contributed to the data points/curves. See Table 4.1 for a statistical comparison among the three curves. **B.** Same as in **A**, but for stimulus conditions where the gratings were paired with: 1) fixed amplitude sounds (green; same data as in **A**); 2) AM sounds (black); and 3) no sounds (purple). See Table 4.2 for a statistical comparison among the three curves.

Table 4.1 Two-way ANOVA table (A) and Tukey post-hoc tests (B) for the comparison among rat psychometric curves shown in Fig 4.3A. A repeated-measures two-way ANOVA was run on the proportion of “High TF” choices reported in Fig 4.3A, having as factors the experimental condition (“V + fixed amplitude sound”, “V + AM congruent” and “V + AM anti-congruent”) and the grating TF.

(A) Two-way ANOVA with repeated measures

Effect	df	SS	MS	F	p
Visual TF	8	33.483	4.185	289.291	$0.001 \cdot 10^{-48}$
Experimental condition	2	0.212	0.106	4.897	0.02
Visual TFs:Experimental condition	16	0.191	0.012	2.533	0.0018

(B) Tukey post-hoc tests

Condition 1	Condition 2	Difference	StdErr	p	Lower	Upper
V + AM sounds (congruent)	V + AM sounds (anti-congruent)	-0.0069	0.0118	0.8297	-0.0397	0.0259
V + AM sounds (anti-congruent)	V + fixed-amplitude sound	0.0625	0.0254	0.0839	-0.0084	0.1336
V + fixed-amplitude sound	V + AM sounds (congruent)	-0.0556	0.0256	0.1298	-0.1271	0.0159

curve measured with the fixed amplitude sounds (green dots/line), that was equivalent to the one previously observed with the congruent and anti-congruent AM sounds (compare to the blue and red curves in Fig 4.3A). Strikingly, a similar but even more prominent shift of the psychometric curve was observed for the unimodal visual conditions, when the gratings were presented without any concomitant sound (purple dots/line). Again, a two-way ANOVA showed a significant main effect of experimental condition, TF of the gratings and their interaction, and a post-hoc Tukey test revealed a significant difference between the curve obtained with the unimodal visual stimuli and the curves measured with the paired sounds, either AM or fixed amplitude (Table 4.2).

These results showed that rat perception of visual temporal frequencies was systematically shifted towards reporting the “high-TF” category if the noise bursts became amplitude modulated and, even more so, when the sounds disappeared altogether. Wondering about the possible cause of these progressively larger shifts, we realized that one key attribute of the sound did decrease from the fixed amplitude to the AM bursts and then further to the purely visual conditions: its intensity. In fact, because of the sinusoidal modulation, there were moments during the AM bursts where the instantaneous sound intensity dropped to zero (A.2B Fig). As a result, the average sound intensity (as computed over the inferred reaction time of the animals) was lower for the AM bursts than for the fixed amplitude burst (see

Table 4.2 **Two-way ANOVA table (A) and Tukey post-hoc tests (B) for the comparison among rat psychometric curves shown in Fig 4.3B.** A repeated-measures two-way ANOVA was run on the proportion of “High TF” choices reported in Fig 4.3B, having as factors the experimental condition (“V + fixed amplitude sound”, “V + AM sound” and “V only”) and the grating TF.

(A) Two-way ANOVA with repeated measures

Effect	df	SS	MS	F	p
Visual TF	8	33.809	4.226	314.013	$0.002 \cdot 10^{-53}$
Experimental condition	2	1.766	0.883	32.593	$0.001 \cdot 10^{-5}$
Visual TFs:Experimental condition	16	0.461	0.29	8.763	$0.001 \cdot 10^{-11}$

(B) Tukey post-hoc tests

Condition 1	Condition 2	Difference	StdErr	p	Lower	Upper
V + fixed-amplitude	V + AM sounds	-0.0614	0.0244	0.0769	-0.130	0.0067
V + AM sounds	V only	-0.1324	0.1776	0.0001	-0.182	-0.0829
V only	V + fixed-amplitude	0.1938	0.0299	0.0003	0.1102	0.2773

Figs 4.1D and A.2C), besides being minimal for the purely visual stimuli (corresponding, in this case, to the environmental background noise). In a previous study (Iurilli et al., 2012), noise bursts were reported to hyperpolarize V1 neurons, with the magnitude of this inhibition increasing monotonically with their intensity. Given this observation, we hypothesized that the perceptual effect of sound is to compress visual cortical representations in a way that is proportional to its intensity.

4.2.2 A computational model accounts for the behavioral impact of task-irrelevant sounds as the result of visual perceptual space compression

To test the hypothesis that response suppression is the key mediator of auditory influences on visual representations, we developed a Bayesian ideal observer model of rat perceptual choices. The model rests on three assumptions, which are illustrated in the cartoons of Fig 4.4 (the full mathematical derivation of the model can be found under the “[Ideal observer model](#)” section in the Materials and Methods). The first is that V1 neurons respond to progressively larger TFs with an approximately linear increase of firing rates, at least within the TF range tested in our study (Fig 4.4A). Although no systematic studies exist about the tuning of rat visual cortical neurons for temporal frequency, this hypothesis is grounded on the periodic activity bursts evoked in rat V1 simple cells by the succession of luminance waves in

drifting gratings (Matteucci and Zoccolan, 2020; Matteucci et al., 2019, 2023). It is therefore reasonable to assume that gratings with a larger number of cycles (i.e., higher TF) will tend to drive V1 simple cells more vigorously. The second assumption is that sound dampens neuronal responses, thus lowering the slope of the linear dependency of neuronal firing from the TF of the gratings, with this decrease being stronger for higher intensity sounds (Fig 4.4A; compare the purple line, corresponding to the unimodal, purely visual condition, to the gray and green lines, corresponding to conditions where the visual stimuli are paired with sounds with progressively larger intensity). This hypothesis is consistent with the sound-induced (and intensity-dependent) hyperpolarization of V1 pyramidal neurons reported by (Iurilli et al., 2012), as well as with the inhibitory impact of sounds on V1 responses reported by several authors under certain stimulus conditions (Deneux et al., 2019; Meijer et al., 2017; Williams et al., 2023) (see the Discussion for details). In the representational space defined by the activity of a population of V1 neurons, such response suppression results in a compression of the population vectors encoding the visual TFs towards the origin of the space (Fig 4.4B; compare the green to the purple distributions). Finally, the third hypothesis underlying our model is that downstream decision neurons learn an optimal discrimination boundary in the representational space (Fig 4.4B; dashed line) under the sensory conditions they were repeatedly exposed to during training (i.e., with the visual gratings being paired with the fixed amplitude sound; green distributions in Fig 4.4B). This optimality is what gives our model its character as an ideal observer (Geisler, 2011; Ma et al., 2023a). We also assume that the boundary remains unchanged when new conditions (e.g., the purely visual gratings; purple distributions in Fig 4.4B) are interleaved with those used for training.

In our model, on each trial, the combination of a visual and an auditory stimulus generates an internal representation sampled from a Normal distribution with fixed standard deviation σ , centered on some average representation of the TF value of the visual stimulus s . Following from the assumptions above, the internal representation is then scaled by a factor γ (see the [Ideal observer model](#) section in the Materials and Methods), and the decision rule is the one that would be Bayes-optimal in the training condition (i.e., with the gratings paired with the fixed amplitude sound). We can thus derive the probability for a rat of classifying a given stimulus s as belonging to the H (i.e., “High TFs”) category as

$$p(\text{report } H \mid s) = \Phi \left[\frac{\gamma s - s_0}{\sigma} \right], \quad (4.1)$$

where Φ is the standard Normal cumulative function and s_0 is the boundary 2.12 Hz visual stimulus that separates the “Low TFs” class from the “High TFs” class (see the [Ideal observer model](#) section in the Materials and Methods). In the model, the σ parameter controls the

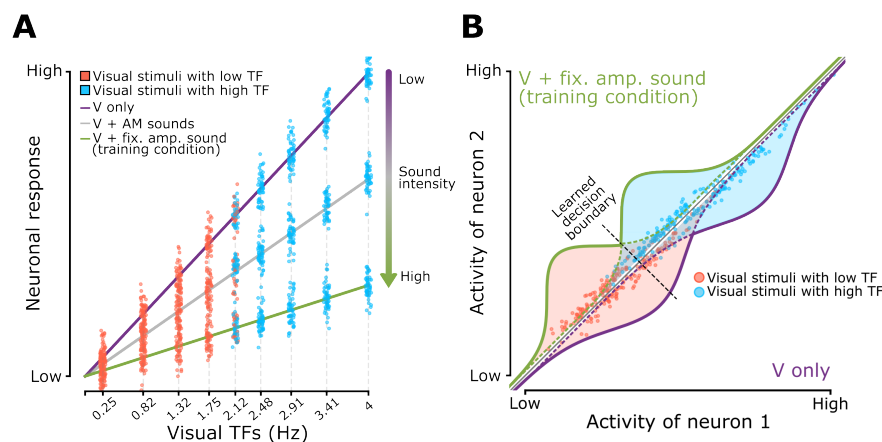


Fig. 4.4 Schematic of an ideal observer model where neuronal populations encode linearly visual temporal frequencies and are suppressed by concomitant sounds in a way that depends on their intensity. **A.** Hypothesized tuning of a visual cortical neuron in the ideal observer model. The neuron responds to the visual gratings with an average firing rate that grows linearly as a function of the grating TF. Due to neuronal noise, the responses to repeated presentations of the same grating are distributed according to a Gaussian function centered on the average firing rate elicited by the grating. More specifically, the red and cyan dots refer to those responses that are elicited by the gratings that, in our task, had to be classified as belonging to the “low TF” and “high TF” classes, respectively. When only the gratings are shown (purely visual condition), the neuronal response evoked by any given TF reaches its maximal possible value (purple line). When the gratings are paired with concomitant sounds, the responses to all the TFs are inhibited by a factor that depends on sound intensity, resulting in a reduction of the slope of the linear dependency between firing rate and TF. Since the AM sounds have a lower average intensity than the fixed amplitude sound (see Fig 4.1D, Materials and Methods, and A.2 Fig), the resulting compression of the firing rate is larger for the latter (green line) than for the former (gray line). **B.** Illustration of the ideal observer model when a population of multiple neurons is considered. The drawing shows a hypothetical neuronal representational space made of two units. Each neuron behaves as shown in A. In the example, two TFs yields two distributions of population vectors (red and cyan dots). When the gratings are paired with a sound with strong intensity (e.g., the fixed amplitude noise burst), all responses are inhibited and, as a result, the distributions are compressed toward the origin of the representational space (green curves). Since rats are trained to discriminate the gratings under these conditions, they learn a decision boundary (dashed line) that optimally separates such compressed distributions. When the animals are presented with unimodal, purely visual gratings, the sound-induced inhibition is released and the response distributions shift towards higher firing rate values (purple curves). The animals, however, still rely on the previously learned decision boundary to discriminate these stimulus conditions. As a result, the overall proportion of “high TF” choices increases (i.e., more red dots fall on the “false alarm” side of the decision boundary).

slope of the psychometric function and measures the internal noise of the representation (i.e., rat sensitivity over the visual TF axis), while the γ parameter controls the gain of the internal perceptual representation.

This model allowed testing alternative hypotheses about the impact of sounds on visual representations, depending on the number of distinct values that the σ and γ parameters were free to take. At one extreme, allowing a different value of σ for each sound condition would represent the hypothesis that sound affects the sensitivity of the visual representation (i.e., the sharpness of the psychometric curves). At the other extreme, keeping a constant σ but allowing a different value of γ for each level of sound intensity would represent the hypothesis that sound compresses the visual representation according to its intensity, but leaves visual perceptual sensitivity unaltered.

Using the model to test for the impact of sound intensity on visual perception required estimating the sound intensity experienced by the rats in the time interval between stimulus onset and each animal's estimated reaction time (see the Materials and Methods and A.2B Fig, second to last row). Analyzing the resulting, trial-averaged intensity levels revealed a finer-grain grouping of stimulus conditions (see Figs 4.1D and A.2C) compared to the one shown in Fig 4.3B. In addition to the conditions with maximal and minimal intensity (i.e., 74.8 dB for the fixed amplitude sounds and 32 dB for the purely visual stimuli, respectively), we found that the AM conditions could be grouped into three main clusters with increasingly larger intensity, based on the TF at which the noise bursts were modulated: TF = 0.25 Hz (48.5 ± 2.4 dB; mean $\pm$ s.e.m. across rats); TF = 0.82 Hz (64.2 ± 2.3 dB); and $1.32 \text{ Hz} \leq \text{TF} \leq 4 \text{ Hz}$ (70.7 ± 0.5 dB; these conditions are shown with progressively darker shades of gray in the stimulus matrix of Figs 4.1C, 4.1D and A.2).

We inferred the parameters of the ideal observer using a hierarchical Bayesian approach (see the section [Ideal observer model](#) of the Materials and Methods, A.3 and A.4 Figs for details on the model and our inference approach), which yielded a posterior probability distribution over the values σ and γ that characterized each specific rat (referred to as rat-level parameters), as well as the values that characterized all rats taken together (referred to as group-level parameters). The group-level estimates obtained for a model with a single σ and 5 different γ per rat (i.e., one γ for each of the 5 levels of sound intensity in our stimulus set) are summarized by the psychometric curves shown in Fig 4.5A.

The curves fitted tightly the measured classification accuracies and captured well the way the proportion of “high TF” choices progressively increased, while sound intensity dropped from maximal (green curve; fixed amplitude sound) to minimal (purple curve; no sound), passing through the three intermediate levels in which the AM sounds were clustered (gray curves) (Fig 4.5; see Materials and Methods and A.5 Fig for a posterior predictive analysis

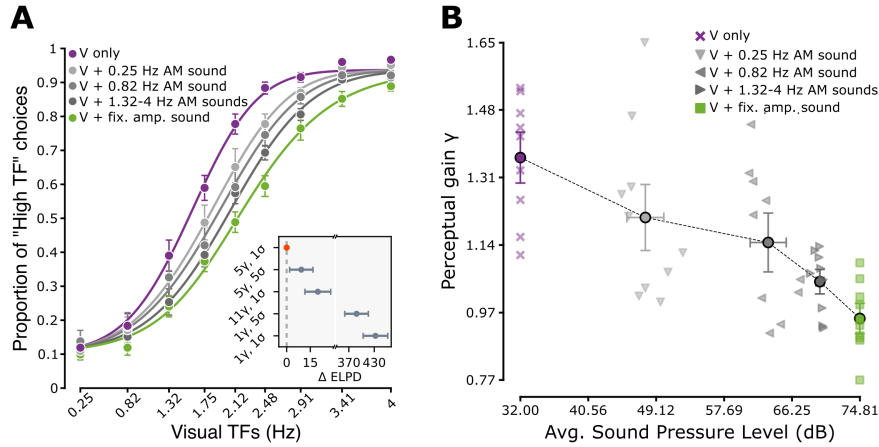


Fig. 4.5 An ideal observer model where task-irrelevant sounds compress rat visual perceptual space according to their average intensity accounts for rat perceptual choices.

A. Group average proportion of “high TF” choices ($n=10$ rats) as a function of the TF of the visual gratings, when the latter were paired with: 1) fixed amplitude sounds (green); 2) AM sounds with increasingly larger TFs (progressively darker shades of gray, matching those in the stimulus matrix shown in Fig 4.1C); and 3) no sounds (purple). The error bars denote s.e.m. over the group of 10 rats. The curves are group-level predictions of the proportions of “high TF” choices according to the Bayesian ideal observer model defined in Eq (4.1), with a single value of the sensitivity parameter σ and 5 different values of the scaling factor γ , one for each level of sound intensity (more precisely, each curve is the mean of the distribution of psychometric curves induced by the posterior distribution of the σ and γ parameters obtained for a given sound intensity level). For this model and the data as presented in the plot, $R^2 = 0.9973 \pm 0.0002$ (posterior mean $\pm$ st. dev.). The inset shows the difference between the expected log predictive density (ELPD) of the reference model with 1 σ and 5 γ values (i.e., the one yielding the psychometric curves) and variants of the model having a given number of free σ and γ parameters (as indicated on the ordinate axis). In all comparisons, the reference model was the one with lowest ELPD, i.e., with the higher predictive power. **B.** Relationship between the magnitude of the scaling factor γ in the ideal observer model and the intensity of the sounds that were paired with the visual gratings. Triangle, square and cross symbols refer to individual rats, while circles are group averages. Sound conditions are labeled according to the same color code as in A and Fig 4.1C. Vertical error bars denote the std. dev. of the posterior distribution of the group-level γ parameters (see Methods). The horizontal error bars denote s.e.m. over the group of 10 rats.

illustrating the tightness of the model fit to the data, and Table A.15 for details on fitted model parameters and inference diagnostics). This means that an ideal observer with constant sensitivity (i.e., a single σ) could describe well the impact of sound on rat visual perceptual choices, provided that the internal representation of visual TFs was appropriately scaled by a gain factor (γ). Importantly, the magnitude of this gain decreased monotonically across the 5 levels of increasing sound intensity in which our stimulus conditions could be grouped (Fig 4.5B).

To verify that this was the best description of the data, we compared the predictive power of this “reference” model to that of other variants, with a different number of values that σ and γ were allowed to take. The predictive power of each model was measured using the expected log predictive density (ELPD), that is, the average, expected log-likelihood of the model on new data. This metric is estimated by leave-one-out cross-validation using a standard approach and automatically penalizes overly complex models that tend to overfit (see section Model comparison in the Materials and Methods). In our comparison, we computed the difference in ELPD between the reference model with 1 σ and 5 γ and each other model. We refer to this metric as Δ ELPD.

The reference model with a fixed σ and 5 γ (one for each of the 5 levels of sound intensity) was the best (i.e., the one with the lowest ELPD) of a broad range of models we tested (Fig 4.5A, inset). In particular, allowing a separate γ for each of the 9 AM conditions, thus obtaining a total of 11 possible γ values (while still keeping σ fixed to a single value), yielded a worse model (an indication of overfitting). This confirmed that noise bursts modulated within the [1.32 Hz, 4 Hz] range were all equivalent in terms of their impact on visual perceptual choices, as expected given their very similar intensity (see Figs 4.1D and A.2). A model with 5 σ and 1 γ also performed much worse than the reference model, thus confirming that the different sound stimuli did not impact rat sensitivity to the visual TFs, but, rather, the dynamic range of the internal representation of the TFs. This conclusion was further reinforced by noticing that not even a model with 5 σ and 5 γ outperformed the reference model with only 1 σ and 5 γ . Finally, in order to describe more quantitatively the relationship between perceptual gain and sound intensity, we also compared the models just described with one that puts a linear constraint on the dependence between these two quantities (while keeping σ fixed as in the best-performing model seen so far). This model performed just as well as the (5 γ , 1 σ) model, confirming quantitatively that γ decreases for larger values of sound intensity (slope: $(-8.6 \pm 1.4) \times 10^{-3}$, intercept: 1.636 ± 0.083 , posterior mean $\pm$ standard deviation; A.6 Fig).

In summary, our Bayesian ideal observer analysis supports the hypothesis that the effect of sound on the representation of temporally modulated visual gratings is to compress the visual

perceptual space rather than increase or decrease perceptual uncertainty. Mechanistically, this finding suggests that sound acts as a powerful modulator of the responses of visual cortical neurons (see Fig 4.4) but leaves their tuning for temporal frequency unaltered.

4.3 Discussion

The goal of our study was to understand the functional impact of task-irrelevant sounds on rodent visual perception. In designing our experiments, we took inspiration from the influences that sounds or direct activation of A1 neurons (e.g., via optogenetics) have been reported to exert on visual cortical activity in rodents. As already mentioned, these influences are extremely heterogeneous.

Iurilli and colleagues (Iurilli et al., 2012) found a sound-induced hyperpolarization of V1 L2/L3 pyramidal neurons, suggesting that the functional impact of A1 projections to V1 is mainly the inhibition of visual cortical activity. The authors also showed that pairing a sound with the light flash used to trigger a visually-driven fear response significantly reduced its magnitude, suggesting a degraded processing of the visual stimulus. These findings are consistent with a view where primary sensory cortices engage in mutual interareal inhibition to compete for the delivery of the most behaviorally salient information to downstream decision centers, in agreement with the predominantly suppressive effect of hetero-modal stimuli that has been reported in auditory and somatosensory cortices of monkeys, cats and ferrets (Bizley et al., 2007; Dehner et al., 2004; Kayser and Logothetis, 2007; Kayser et al., 2008; Meredith et al., 2006). This does not exclude that sound-induced inhibition could differently affect distinct neuronal subpopulations. For instance, it could impact more on the responses of pyramidal cells that are only weakly activated by visual stimuli, thus potentially sharpening the selectivity of V1 neurons. This hypothesis is consistent with a later mouse study (Ibrahim et al., 2016), showing that orientation tuning of V1 L2/L3 pyramidal neurons was sharpened in the presence of sound, thanks to a disinhibitory circuit activated by A1 projections to V1 interneurons in L1 and resulting in an increase of firing at the preferred orientation of the pyramidal cells. Such a sharpening of stimulus selectivity was confirmed by (McClure Jr and Polack, 2019), who showed how pure tones enhanced the responses of V1 neurons tuned for the orientation/direction of the test grating, while suppressing the activity of neurons with orthogonal/opposite preference. The gating of A1 inputs to V1 by L1 interneurons was confirmed by (Deneux et al., 2019), who, however, found the resulting modulation of L2/L3 neurons to be strongly context-dependent – inhibition was dominant in the dark, while excitation prevailed in the light. In addition, this study found that loud sound onsets increased the response of L2/L3 neurons to temporally coincident visual stimuli.

This suggests a functional role of A1 projections in boosting the saliency of visual events that co-occur with abrupt sounds, in agreement with the observed enhancement of BOLD responses in human V1 during the sound-induced flash illusion ([Watkins et al., 2006](#)).

Yet another study ([Meijer et al., 2017](#)) stressed the dependence of auditory modulation of V1 responses from the features that define the visual and auditory stimuli: visual contrast, sound envelope, and temporal congruency. Again, both sound-mediated enhancement and suppression of visually evoked responses were observed in V1, with suppression becoming dominant when visual drifting gratings were paired with frequency modulated sounds having an incongruent TF - an indication that V1 neurons may encode the temporal congruency between audio-visual stimulus features. More recently, ([Williams et al., 2023](#)) found that V1 responses to drifting gratings were enhanced by the concomitant presentation of auditory white noise, resulting in an increased discriminability of grating orientation and direction from neuronal signals. Interestingly, these auditory modulations were not accounted for by sound-evoked orofacial movements and locomotory behavior, as previously suggested by ([Bimbard et al., 2023](#)). The dissociation of early auditory-related and a later motor-related modulations of V1 activity was confirmed by ([Oude Lohuis et al., 2024](#)), which, however, found that the encoding of visual features was poorly affected by both kinds of influences.

In our study, we designed a discrimination task and a set of audio-visual stimuli that allowed looking for the perceptual manifestation of several of the (often contrasting) neurobiological processes reviewed above. Consistently with ([Meijer et al., 2017](#)), our visual stimuli were circular drifting gratings spanning a range of TFs (Fig 4.1A and B), i.e., rich spatiotemporal patterns that are known to strongly activate rat visual neurons ([Matteucci and Zoccolan, 2020](#); [Matteucci et al., 2019, 2020, 2023](#)). Also as in ([Meijer et al., 2017](#)), some of the sounds that we paired with the gratings were modulated at the same TFs of the gratings in such a way to yield consistent, inconsistent or anti-consistent matches (Fig 4.1C). However, to maximize the possible enhancement of visual perceptual events by the concomitant sounds, the auditory stimuli were not modulated in frequency but in amplitude (A.2A Fig). The rationale was that the periodic bursts of visual cortical activity evoked by the succession of luminance waves in the gratings (as expected, for instance, for V1 simple cells; see ([Matteucci and Zoccolan, 2020](#); [Matteucci et al., 2019, 2023](#))) could be boosted by the close succession of up-ramps in the sinusoidal envelopes of the AM sounds ([Deneux et al., 2019](#)), thus potentially improving the discriminability of the gratings in the case of congruent TFs. At the same time, the amplitude modulation at different TFs yielded sounds with variable average intensity over the response window of the animals (see Figs 4.1D and A.2C). This allowed testing whether the impact of sound on visual perception depended on sound intensity, as expected according to ([Iurilli et al., 2012](#)).

Overall, our experiments provided a clear answer to the question of whether task-irrelevant sounds have an impact on visual perception, raising specific, testable hypotheses about the nature of the underlying audio-visual interactions.

First, we found that sounds have an impact on rat visual perceptual decisions (compare the purple to the green curve in Fig 4.3B) as large as that typically reported in multimodal integration studies (Nikbakht et al., 2018; Raposo et al., 2012, 2014). This means that the direct, automatic influence of auditory inputs on visual representations is as effective on rat perceptual choices as the process of learning how to combine different sensory cues in supra-modal, Bayes-optimal representations.

Second, both the magnitude and direction of auditory influences on rat visual perception were independent of the temporal consistency between the AM sounds and the gratings (compare the red to the blue curve in Fig 4.3A). This indicates that, differently from what observed by Meijer and colleagues (Meijer et al., 2017) with frequency modulated sounds, no encoding of the congruency between the temporal features of the audio-visual stimuli took place under our experimental conditions.

Third, no sharpening of the psychometric curves was observed – rather, we found an asymmetrical (i.e., TF-dependent) bias towards reporting more often the “High TF” class (Fig 4.3A). This bias points to a sound-mediated increase of the “High TF” evidence encoded by the visual cortical representation when the AM sounds, rather than the fixed amplitude burst, were paired with the gratings. Assuming that visual cortical neurons encode progressively higher visual TFs with increasingly larger firing rates (see Fig 4.4A), such enhancement of “High TF” evidence could be accounted for by two alternative mechanisms: excitation or disinhibition of visually driven responses. The first hypothesis would be consistent with the conclusion of Deneux and colleagues (Deneux et al., 2019) that sound onsets are especially effective at boosting visual cortical responses. As such, the succession of up-ramps in the sinusoidal envelopes of the AM sounds would have been more effective in enhancing V1 responses, as compared to the stationary, fixed amplitude noise burst. Hence, the increase of “High TF” evidence carried by the neuronal population and the resulting increase of “High TF” choices. Such a direct excitation of visual cortical neurons would also be consistent with the early cross-modal integration mechanisms that are thought to contribute to the sound induced flash illusion (Hirst et al., 2020; Keil, 2020; Shams et al., 2000; Watkins et al., 2006).

This explanation, however, is not supported by the observation that the proportion of “High TF” choices further increased when sounds were removed altogether (Fig 4.3B; purple curve). Hence the intuition that sounds must have inhibited, rather than excited, the firing of the visual neurons encoding the TF of the gratings, in a way that was dependent on their

intensity (Fig 4.4A), in agreement with the findings of Iurilli and colleagues (Iurilli et al., 2012). According to this hypothesis, the observed shifts in the psychometric curves were the result of an increasingly large disinhibitory effect when the sounds paired with the gratings progressively reduced their intensity, passing from the fixed amplitude burst to the AM sounds and, finally, to the purely visual conditions. Moreover, according to this hypothesis, AM sounds with lower TF, given their lower average intensity, should have yielded a larger disinhibitory effect and, therefore, a larger shift of the psychometric curve, as compared to AM sounds with higher TF. This trend was indeed observed in our data (Fig 4.5).

Importantly, the above-mentioned hypothesis was implemented in a Bayesian ideal observer model of rat perceptual choices, combined with a simple neural coding scheme. The scheme assumes the existence of a population of visual neurons that encode the TF of the gratings with linearly increasing firing rates and that are inhibited by sounds by a measure that depends on sound intensity (Fig 4.4). These assumptions, along with the further assumption of a fixed decision boundary learned by the rats during training, led to a model (Eq. 4.1) that was able to capture very precisely the full spectrum of psychometric curves we measured, using just one parameter for rat sensitivity over the TF axis and five levels of sound-induced suppression (Fig 4.5A, inset), related to sound intensity (Fig 4.5B).

Overall, these findings imply that, among the variety of auditory influences that have been reported in rodent visual cortex (Deneux et al., 2019; Ibrahim et al., 2016; Iurilli et al., 2012; Oude Lohuis et al., 2024), inhibition plays the dominant role at the functional level, at least under the experimental conditions of our study. This conclusion is in agreement with the findings of Iurilli and colleagues (Iurilli et al., 2012). It also resonates with the experience-dependent suppression of V1 activity by auditory stimuli reported by Garner and Keller (Garner and Keller, 2022). This study, however, is not directly comparable to ours because: (1) audio and visual stimuli were delivered sequentially and not concomitantly; and (2) auditory stimuli were always informative about the upcoming delivery of a reward, unlike in our task where they were completely devoid of information.

At the same time, our results appear inconsistent with the sound-induced enhancement of visual cortical responses and the sharpening of feature selectivity reported by some authors (Ibrahim et al., 2016; McClure Jr and Polack, 2019; Williams et al., 2023). One possibility to reconcile these findings is that auditory signals may differently affect distinct kinds of perceptual tasks. In tasks requiring the discrimination of visual features that are related to the energy of the stimulus (such as its TF) and that are likely encoded by locally linear tuning curves (as postulated in our model; see Fig 4.4A), sound-mediated engagement of inhibitory circuits may lead to a dominant suppression of visually-evoked responses. In tasks where spatiotemporal features (e.g., orientation and direction) are encoded by neuronal

populations with bell-shaped, unimodal tuning curves (e.g., over the orientation axis), the interplay between local inhibitory and disinhibitory circuits activated by sound could instead lead to a decrease of perceptual uncertainty and an improvement of decoding accuracy.

Another possible explanation at the root of the discrepancy between our findings and those of some previous studies is that sound may enhance the early portion of a visually evoked response, while suppressing its sustained epoch, as reported by (Williams et al., 2023). Given the nature of our task, which required perceptual integration of TF evidence over a prolonged time window, it is possible that inhibition played a dominant role in modulating visual perception. In other tasks, where evidence about the relevant stimulus feature (e.g., orientation) is available within a few tens of milliseconds after stimulus onset (Matteucci and Zoccolan, 2020), enhancement of early visual cortical responses by concomitant sounds may lead to an increased neuronal (and, possibly, perceptual) discriminability of the visual stimuli.

In conclusion, our experiments demonstrate that task-irrelevant auditory signals have a strong influence on the way task-relevant visual information is processed in the rodent brain. By combining psychophysics and computational modeling, our study points to very specific processes underlying these hetero-modal influences. As graphically illustrated in the cartoon of Fig 4.6, our prediction is that sound bursts suppress visual cortical responses to drifting gratings via cortico-cortical projections from A1 to V1, lowering the gain of the tuning of V1 neurons for the gratings' temporal frequency. This provides clear mechanistic hypotheses that future neurophysiological and neuroanatomical work on visual cortical codes will have to test, in order to fully understand the nature of this modulation and its interplay with task requirements.

4.4 Materials and methods

4.4.1 Subjects

All animal procedures followed international and institutional standards for animal care and use in research and were approved by the Italian Ministry of Health (project approval 940/2015-PR on September 4, 2015; project approval 1183/2020-PR on December 4, 2020).

We trained 12 adult Long Evans rats (Charles River Laboratories) in a visual temporal frequency (TF) classification task. The animals, seven weeks old on arrival, weighed approximately 240 g each. They were housed three per cage in a climate-controlled chamber ($23 \pm 1^\circ\text{C}$, 40-50% humidity). Training began at 10 weeks, following two weeks of handling to accustom them to human interaction. Rats always had free access to food (20 g/day) but

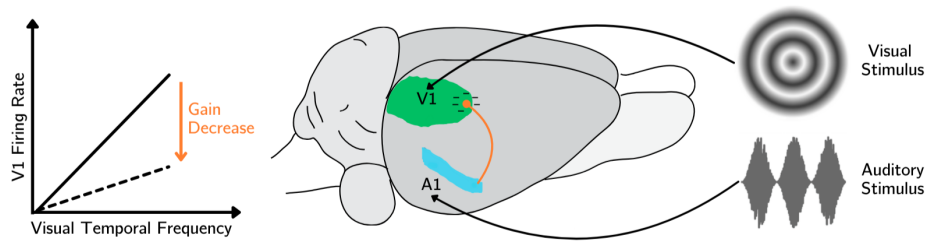


Fig. 4.6 Outline of the hypothesized interaction between auditory and visual cortex underlying the sound-driven compression of rat visual perceptual space. Activation of A1 (in cyan) by sound bursts leads to an inhibition of the responses evoked by circular drifting gratings in V1 (in green) via cortico-cortical projections (orange line). This produces a decrease in the gain of the tuning curves of V1 neurons over the temporal frequencies of the gratings.

their access to water was restricted in the days of the behavioral training/test (5 days a week). Water was provided as a reward for correct responses in the discrimination task and for about 30 minutes after each session, ensuring at least the recommended 50 ml kg^{-1} daily intake. During the course of the study, their body weight increased to about 500 g. The duration of the training phase (see below) varied depending on the animal. Five rats completed it in about 3 months and proceeded to the test phase (see below), while for the remaining rats it took about 6 months. The test phase lasted approximately 12 months for all the rats.

4.4.2 Behavioral apparatus

Rats were trained in a rig made of two identical racks, each with three vertically stacked slots, allowing six animals to be trained simultaneously. Each slot contained a custom-built black operant chamber ($20 \times 20 \times 30 \text{ cm}$) with a 4 cm-diameter viewing hole for the rats to extend their heads and face the stimulus display (Fig 4.1A). The visual stimuli were presented on a 21.5" LCD monitor (ASUS VE228) placed at 30 cm from the viewing hole, while the auditory stimuli were presented from a full-range speaker (Visaton BF37, 4 Ohm, 1.5" diameter, frequency response 100–20000 Hz) positioned above the monitor, tilted 45° towards the viewing hole, and located 35 cm from it. Each speaker received a pre-amplified signal (Q-BAIHE TDA7377PRO). A 3D-printed block with three equidistant response ports was positioned at 3 cm from the viewing hole. Each port had a stainless-steel feeding needle and a LED-photodiode pair that worked as a licking sensor to detect interactions of the rats with

the port. Signals from the photodiodes were digitized with a microcontroller (Phidgets 1203 input/output device) and sent to a computer running the open-source MWORKS software (<https://mworks.github.io/>). This system automatically presented the visual and auditory stimuli when the animal's nose was detected in the central port, while the lateral ports were used to record the responses of the animals to the visual stimuli and decide whether to deliver the reward, along with reinforcement sounds (Fig 4.1B). Correct responses were rewarded with a 4% stevia-water solution from the lateral feeding needles connected to computer-controlled syringe pumps (New Era Pump Systems; NE-500). Each animal was individually monitored with a camera (AXIS M1033-W Network Camera) positioned 14 cm above the chamber viewing hole. The operant chambers were acoustically insulated from each other and the environment with sound-proof foam panels.

4.4.3 Visual and auditory stimuli

Rats were presented with grayscale outward-moving sinusoidal gratings (30 fps) with a spatial frequency of $0.05 \text{ cycles deg}^{-1}$ and temporal frequencies of 0.25, 0.82, 1.32, 1.75, 2.12, 2.48, 2.91, 3.41, and 4 Hz (the temporal frequency of the gratings can be defined as the number of grating cycles moving through a point of the screen in 1 second). These frequencies were symmetrically distributed around 2.12 Hz on a logarithmic scale (Fig 4.1C, x axis). The gratings, displayed on a middle-gray background, were enveloped with a Gaussian circular aperture with diameter corresponding to the monitor height (see examples in Fig 4.1A and B).

Two classes of auditory stimuli were used in our study: a fixed amplitude sound (Fig 4.1D, green curve) and amplitude modulated (AM) sounds (see Fig 4.1D, gray curve, for an example; see A.2A Fig, gray curves, for the complete set of stimuli). In both cases, the stimuli were generated in MATLAB (<https://www.mathworks.com/>) from a full-range white noise sampled at 44.1 kHz. In the case of the AM sounds, the white noise time series was used as a carrier and multiplied with positive sinusoid envelopes (the modulators), having the same temporal frequencies of the visual stimuli (Fig 4.1C, y axis). More in detail, sound stimuli were generated as digital control signals, sampled at 44.1 KHz in arbitrary units, which were then converted in WAV format (16 bits per sample). The WAV files were then passed through our custom-built amplifier and loudspeaker (see "Behavioral Apparatus" and Fig 4.1A).

The control signal was generated starting from a white noise sequence. We used two distinct white noise sequences, one for the fixed amplitude stimulus and one for the AM stimuli. In the fixed amplitude condition, the white noise sequence (denoted as $a_{\text{fix}}(t)$) was composed of independent and identically distributed random variables sampled from a

normal distribution with a sampling rate matching that of the control signal ($v_{WN} = 44.1$ kHz). The distribution had parameters $\mu = 0$ and $\sigma = 1$. Before conversion to WAV format the sequence was normalized to have values in the interval $[-1, 1]$, hence giving the distribution an effective standard deviation $\sigma_{eff} = 0.283$. The same sequence was used in all the “V + fixed amplitude sound” trials.

The white noise carrier of the amplitude modulated sounds, denoted as $a_{AM}(t)$, was composed of independent and identically distributed random variables sampled from a uniform distribution in an interval $[-1, 1]$, again with a sampling rate of $v_{WN} = 44.1$ kHz. The same sequence was used in all the trials of the modulated conditions. The sequence $a_{AM}(t)$ was multiplied by a modulation function $m(t) = (1 - \cos(\omega_n t))/2$, where $\omega_n = 2\pi v_n$ is the angular frequency of modulation and v_n is the modulation frequency of condition n (A.2A Fig, gray curves). The modulation was non-negative and bounded in $[0, 1]$, ensuring that the modulated signal remained within $[-1, 1]$.

4.4.4 Estimation of sound intensity

We measured the intensity of the sound stimuli with a phonometer (Tadeto SL720). At any point in time, the phonometer provides an instantaneous estimate of the sound intensity based on a characteristic sampling window (0.125 s). To avoid any interactions between the temporal modulation of our sound stimuli and the finite sampling rate of the phonometer, we devised and calibrated a transformation that estimated the instantaneous sound intensity as a function of the nominal control signal. This transformation takes into account that the control signal $a(t)$ is converted into a sound pressure wave $p(t)$ by the amplifier/speaker, assuming minimal distortion (i.e., that there is a linear relationship between a and p), and that the measurements of sound intensities $L(t)$ returned by the phonometer are in dB - i.e., they are a nonlinear function of $p(t)$. We note here that our dB measurement are done using “A” frequency weighting, as in previous studies (Turilli et al., 2012).

To carry out this computation, we first verified that the relationship between the nominal control signal $a(t)$ and $p(t)$ as measured by the phonometer was linear. To do so, we measured with the phonometer the intensity L of a uniform white noise sequence lasting 10 seconds. The long duration ensured that we got a precise estimate of the intensity in dB. We repeated the measurements 10 times, each time dimming the control signal a by a factor of 0.8. All these measures were taken by placing the phonometer inside the sound-proof operant chambers, in the same position where the head of the rats was located during the task. Under the same settings, we also measured the intensity L_a of the ambient noise inside the operant chambers (32 dB). The relationship between the sound intensity L (in decibels) of a white

noise sequence and its sound pressure level p (in Pascals) is given by:

$$L = 10 \times \log_{10} \left(\frac{\bar{p}^2 + p_a^2}{p_0^2} \right) \quad (4.2)$$

where $\bar{p}$ is the root-mean-square (RMS) of $p(t)$ over a time window τ , $p_0 = 20 \mu\text{P}$ is the reference pressure, and p_a is the pressure of the ambient noise. Inverting the equation, we can compute the sound pressure from the intensity measurements. The ambient sound pressure is given by:

$$p_a = p_0 \sqrt{10^{\frac{L_a}{10}}} \quad (4.3)$$

The sound pressure of the white noise sequences is, therefore:

$$p = p_0 \sqrt{10^{\frac{L}{10}} - \left(\frac{p_a}{p_0} \right)^2} \quad (4.4)$$

We could thus compute p from the measured values of L , finding that, as expected under the assumption of a linear mapping from a to p , the average pressure for each sequence was 0.8 ± 0.01 times as intense as the previous one. This confirmed that $p(t) = f \times a(t)$ for some value of f and allowed us to estimate the factor f of the linear relationship as $f = 0.39 \text{ P}$. We repeated the same procedure with gaussian white noise sequences, confirming the result and obtaining a compatible estimate for f . We could then use the estimated value of f to calculate the sound intensity (in dB) of all sound stimuli used in the experiment. For the fixed amplitude, white noise sound $a_{\text{fix}}(t)$, this results in $L_{\text{fix}} = 74.8 \text{ dB}$. For the amplitude modulated stimuli, their sound intensity profile $L_{\text{AM}}(t)$ is in principle influenced by the duration τ of the time window over which the RMS of $p_{\text{AM}}(t)$ is computed. For short values of τ , close to the white-noise-sampling period ($T_{\text{WN}} = \frac{1}{v_{\text{WN}}}$), noise fluctuations dominate the temporal profile. By contrast, long τ s, comparable to the period of the modulation, will effectively smooth the modulation, resulting in L_{AM} values that do not accurately represent the dynamic characteristics of the signal. For a wide range of intermediate time windows, however, the computed sound intensity is independent of the details of the fluctuation of the carrier and its temporal profile correctly tracks the temporal modulation.

More specifically, the RMS of $p_{\text{AM}}(t)$ over a time window τ is defined as

$$\bar{p}_{\text{AM}}(t)_\tau = \sqrt{\langle f^2 \cdot a_{\text{AM}}(t)^2 \cdot m(t)^2 \rangle_\tau} \quad (4.5)$$

If τ is smaller than the temporal scale of the modulation, then $m(t)$ is approximately constant inside the integration window and

$$\langle f^2 \cdot a_{\text{AM}}(t)^2 \cdot m(t)^2 \rangle_\tau \simeq f^2 m(t)^2 \langle a_{\text{AM}}(t)^2 \rangle_\tau \quad (4.6)$$

On the other hand, if τ is larger than the sampling rate of the white noise ($\tau \gg T_{\text{WN}}$) there will be enough samples in the window for the RMS of $a_{\text{AM}}(t)$ to be equal to the uniform distribution average:

$$\bar{a}_{\text{AM}}^2 = \langle a_{\text{AM}}(t)^2 \rangle_\tau \simeq \mathbb{E}[a_{\text{AM}}(t)^2]_{p(a_{\text{AM}})} = \int_{-1}^1 \frac{1}{2} a^2 da = \frac{1}{3} \quad (4.7)$$

Therefore, the resulting temporal intensity profile of the AM sounds (in dB), when using one of the intermediate values of τ can be computed as:

$$L_{\text{AM}}(t) = 10 \log_{10} \left(\left(\frac{f \cdot \bar{a}_{\text{AM}} \left(\frac{1 - \cos(\omega t)}{2} \right)}{p_0} \right)^2 + \left(\frac{p_a}{p_0} \right)^2 \right) \quad (4.8)$$

The sound intensity profiles $L_{\text{AM}}(t)$ for each modulation frequency (computed with $\tau = 5 \text{ ms} \simeq 220 T_{\text{WN}}$), as well as the profiles of L_{fix} and L_a (with the latter corresponding to the “visual only” condition) are shown in A.2B Fig.

To estimate the sound intensity experienced by a rat during each trial, we computed the average intensity $\bar{L}_n(T)$ of the trial condition n throughout a specific time window with duration T . This duration was equal to the *reaction time* (RcT) of the animal in that given trial, i.e., to the period during which the rat was exposed to the sound stimulus before initiating a motor response. The reaction time is defined as:

$$RcT = RT - MoRsT \quad (4.9)$$

where RT is the *response time* (i.e., the time from the trial onset to the moment a response port is reached), and $MoRsT$ is the *motor response time* (i.e., the time taken by the animal to reach the selected response port, after leaving the central port for the first time). In the recorded data, only the response time was directly measured for each trial. To estimate the reaction time, we assigned to $MoRsT$ a fixed value across all trials, as determined in one of our previous studies, which carefully mapped the trajectories and durations of rat head movements in a similar experimental setup (Vanzella et al., 2019). Specifically, we used $MoRsT = 0.3 \text{ s}$, which corresponds to the duration of a direct ballistic movement from the central port to either response port. This type of movement was the most frequently observed

in that study (1149 out of 1359 recordings). Hence, for each trial i , we computed the average perceived intensity as $\bar{L}_n(RcT_i) = \bar{L}_n(RT_i - MoRsT)$. An example of the computation is shown on the “0.25 Hz AM sound” row of A.2B Fig.

The resulting values for each trial are shown as scatter plots in A.2C Fig, along with violin plots to illustrate their distribution in each sound condition for every individual rat (the dashed lines report the rat group average intensities). The averages of these sound intensities obtained for each rat are those reported on the x axis of Fig 4.5B (squares, triangles, and crosses). Note, however, that, in that figure, the AM conditions with modulation frequencies $1.32 \text{ Hz} \leq \text{TF} \leq 4 \text{ Hz}$ have been grouped together, given that their average intensity was virtually identical (as shown in Fig 4.1D).

4.4.5 Behavioral task

The rats learned to initiate a behavioral trial by nose-poking the central port, and to approach the lateral ports to classify the visual stimuli based on their temporal frequency (Fig 4.1B). The “low temporal frequencies” class (0.25-1.75 Hz) and the “high temporal frequency” class (2.48-4 Hz) had to be reported by half of the rats by poking respectively the left and right port, while the other half was trained with the opposite association. To initiate a trial, rats had to keep their nose in the central port for 300 ms, followed by a 100 ms tone (500 Hz pure tone) signaling the stimulus presentation, occurring 200 ms later. After stimulus onset, rats were given 2 seconds to make a response; responses within 300 ms were aborted to prevent impulsivity. Correct responses were rewarded with 4% stevia-water solution and a concomitant reinforcement sound, while incorrect responses were followed by an unpleasant sound (400 ms, 300 Hz square-wave) and a 1-3 second timeout period. A middle-grey background was displayed during non-stimulus periods. In a given trial, every stimulus had an equal probability of being presented, but with the constraint that a maximum of three consecutive presentations of the same visual class was allowed to prevent the development of a bias towards one of the response ports. This trial structure was the same for all the training and test phases of the experiment, described in the next section. Stimulus presentation, response collection, and reward delivery were controlled by the MWORKS software (<https://mworks.github.io>).

4.4.6 Experimental design

Training

The animals were trained according to the following phases.

Habituation: the rats were first habituated to the operant chamber and were gradually trained to initiate trials autonomously and receive rewards from the lateral feeding needles. Initially, the animals were manually guided to the nose-poke ports, often reinforced with honey drops. A trial started once the animal touched the central port with its nose, signaled by a positive reinforcement tone, and followed by stevia-water rewards from both lateral feeding needles. This phase lasted about 7 days, progressing to the next phase once the animals initiated at least 100 trials autonomously in three consecutive sessions.

Phase 1: rats were conditioned to associate rewards with the presentation of paired auditory and visual stimuli. Initially, to facilitate the task, the visual stimuli were those with the lowest (0.25 Hz) and highest (4 Hz) temporal frequencies and were presented at full contrast alongside the fixed amplitude white noise sound for 2 seconds. Concomitantly, the reward was delivered from the correct response port, based on the class the visual stimulus belonged to. Advancement to the next phase was based on qualitative inspection of the behavior, ensuring that rats attended to the visual stimuli. Importantly, here, as in the rest of the training and test phases, only the visual stimuli were associated to a given reward port (i.e., to a given response class). The auditory stimuli were always task-irrelevant, i.e., were uninformative about the correct classification of the visual stimuli.

Phase 2: in this phase, the rats were actively reinforced based on their classification of the visual stimuli, presented again with a temporal frequency of either 0.25 or 4 Hz. That is, in this phase, the reward was delivered only after the rat had approached and licked the correct response port, while, in case of incorrect choice, the disturbing sound was delivered and the time-out period started. Again, both visual stimuli were paired with the same, uninformative fixed amplitude sound. Advancement to the next phase required that an animal correctly classified at least 70% of the visual stimuli in at least two consecutive sessions. Two of the 12 rats failed to meet this criterion and were excluded from further testing.

Phases 3, 4, and 5.: here, we gradually introduced the additional temporal frequencies of the visual stimuli. First, we added the gratings at 0.82 and 3.41 Hz (*Phase 3*), then those at 1.32 and 2.91 Hz (*Phase 4*), bringing to six the total of stimuli to classify. Finally, we presented all nine possible visual temporal frequencies (*Phase 5*). Also in this case, each advancement to the next phase depended on whether an animal was able to achieve an accuracy criterion of 70% correct responses in at least two consecutive training sessions. The 2.12 Hz stimulus, serving as an arbitrary boundary between the “low temporal frequency” and “high temporal frequency” classes, was randomly assigned to one of the classes in each trial and rewarded accordingly.

Upon achieving 70% accuracy in at least two consecutive sessions of Phase 5 (i.e., once all visual temporal frequencies had been introduced), indicating proficiency in classifying

all visual stimuli, we progressively reduced the contrast of the visual stimuli. The contrast reduction was performed in two main blocks based on accuracy criteria. In the initial exploratory block of contrast reduction, stimulus contrast was progressively lowered within 6-10 sessions, bringing it down to 15% for some animals (e.g., see example rat 1 in Fig 4.2A). During this phase, transient drops in accuracy occasionally occurred; in such cases, the contrast was temporarily increased to the preceding level or back to 100% to allow the animal to recover and consolidate its performance before resuming the reduction (e.g., see example rat 2 in Fig 4.2B). To further ensure the robustness of learning and to verify that animals generalized their classification strategy across contrast levels, all animals were subsequently re-exposed to full contrast before undergoing a second, final contrast reduction block. This final phase was implemented in three steps (50%, 35%, and 25% contrast), based on accuracy criteria: if a rat maintained $\geq 70\%$ correct choices for at least two consecutive sessions at a given contrast level, the contrast was further decreased. The learning trajectory during the training phase is illustrated in Fig 4.2 for two representative rats. All animals consistently reached stable performance at 25% contrast across temporal frequencies, which defined the entry point to the following test phase.

Each training session lasted 40-60 minutes (Caramellino et al., 2021; Matteucci et al., 2021), with the number of trials per session varying between ~ 100 and ~ 400 depending on the rat (with some animals being more prolific than others) and on the specific training day. As shown in Fig 4.2B, occasional drops in trial counts occurred, typically associated with a decrease in performance, which were likely caused by momentary reductions in the motivation of the animal or temporary failures of the behavioral rig (e.g., uneven delivery of liquid from the reward ports).

Test phase

After completing the training phase, the rats entered the test phase. Here, the visual stimuli presented in each trial could be paired not only with the fixed amplitude sound (as done during training) but also with the amplitude modulated auditory stimuli. Specifically, the trials in the test phase fell into four main categories, based on the combination of visual and auditory stimuli (see Fig 4.1C).

- In the “V + fixed amplitude sound” trials, which matched those in training Phase 5, the visual gratings of varying temporal frequencies were paired with the fixed amplitude white noise sound (green cells in Fig 4.1C).
- In the “V + AM sounds” trials, the visual gratings were paired with each of the amplitude modulated sounds at the various frequencies, resulting in all 81 possible

pairwise combinations of visual and auditory frequencies (cells in the squared matrix of Fig 4.1C).

- In the “V only” or unimodal visual trials, the drifting gratings were presented without any accompanying sound (purple cells in Fig 4.1C).
- In the “A only” or unimodal auditory trials, the amplitude modulated sounds were presented alone, without the visual gratings (cyan cells in Fig 4.1C). Since the rats had never been trained to classify the auditory stimuli, we simply collected their spontaneous responses to these sounds, without providing any feedback (i.e., neither reward nor time-out) about their choices.

All the stimulus conditions listed above were presented (pseudo-randomly interleaved) in every test session. As during the training phases, each test session lasted 40-60 minutes (Caramellino et al., 2021; Matteucci et al., 2021), with the number of trials per session varying between ~ 100 and ~ 400 depending on the rat and on the specific training day. The typical number of test sessions that were collected for each animal was ~ 90 .

Again, throughout both the training and test phases, the temporal frequency of the visual stimuli remained the only task-relevant feature for inferring the correct reward location. In contrast, the auditory stimuli were always task-irrelevant and did not indicate the reward side. We verified that the rats did not implicitly learn to classify the auditory stimuli according to their temporal frequency, by plotting their choices in the “A only” trials (see A.1 Fig). The resulting group average psychometric curve was flat, showing no modulation as a function of sound frequency.

4.4.7 Statistical analyses

The classification accuracies achieved by the rats in the test conditions described in the previous section were statistically compared using a repeated measures, two-way ANOVA (see Fig 4.3 and Tables 4.1 and 4.2). To perform the ANOVA, we first calculated the probabilities of each rat to classify a visual stimulus as belonging to the “High TF” category across all experimental conditions. This resulted in 10 data points for each of the stimulus conditions analyzed in Fig 4.3A and B, corresponding to the 10 rats in our study. The ANOVA analysis requires two key assumptions for the data distributions:

1. Normality, i.e., the data should follow a Gaussian distribution. This is particularly important for proportions of correct choices, as distributions near 0 and 1 can saturate and deviate from normality. To address this issue, we transformed the psychometric

data using the arcsine of the square root of each value. After the transformation, we tested the normality of the distributions with the Shapiro-Wilk test (Shapiro and Wilk, 1965) and we verified that it was satisfied by each data point distribution.

2. Homoscedasticity, i.e., the variance should be equal across experimental conditions. The arcsine transform also helps stabilize the variance for proportional data. This was assessed using Levene's test. Following the transform, the test could not reject the null hypothesis that all distributions had equal variance ($p = 0.3421$), indicating that the assumption of homoscedasticity was satisfied.

We then proceeded to calculate the two-way ANOVA with repeated measures. Our experimental design involved two independent variables: the TF of the visual stimuli, with 9 levels, and the experimental condition, with 3 levels. The latter were as follows. For the analysis carried out in Fig 4.3A: the “V + fixed amplitude sound”, the “V + AM sounds (congruent)” and the “V + AM sounds (anti-congruent)”; for the analysis carried out in Fig 4.3B: the “V + fixed amplitude sound”, the “V + AM sounds” and the “V only”. The dependent variable was rat response, measured as the probability of classifying a visual stimulus in the “High TF” category for each experimental condition, which was recorded for the same 10 animals across all the different levels of the independent variables. Tables 4.1A and 4.2A present the results of this analysis for the data shown, respectively, in Fig 4.3A and B, reporting, in both cases, significant effects for both the visual TF and the experimental condition variables, as well as a significant interaction between them. This indicates that the dependent variable changed across different visual TFs and experimental conditions, and that there was a combined influence of visual TF and experimental condition on the dependent variable.

Finally, to determine which specific levels of the experimental conditions contributed to the main effect reported in the ANOVA table, we conducted post-hoc tests. We performed pairwise comparisons of the different levels of the experimental conditions and evaluated their effects using the Tukey test to control for multiple comparisons, as summarized in Tables 4.1B and 4.2B.

4.4.8 Ideal observer model

In this section we derive the ideal observer model used to predict the responses of the rats in the visual rate classification task. The ideal observer follows a standard structure for a sensory classification task, based on the assumption that an internal, abstract representation of the visual stimulus, called the *measurement* and indicated by x , can be modeled as a Gaussian random variable (an approach that is firmly established in psychophysical modeling, going back to its roots in signal detection theory, see e.g. (Geisler, 1989, 2011; Green and Swets;

Kersten et al., 2004; Ma et al., 2023a)). The main nonstandard element of our analysis is the added hypothesis that x is scaled by a gain factor γ which can depend on the intensity of the auditory stimulus. This additional hypothesis stands on its own on an abstract level, but we show below that it can be derived as the consequence of a putative suppressive effect of sound on the activity of visual cortical neurons.

On each trial of the two-alternative forced choice (2AFC) task, the visual stimuli were assigned a temporal frequency which was randomly chosen from a pool of 9 possible values (0.25, 0.82, 1.32, 1.75, 2.12, 2.48, 2.91, 3.41 and 4 Hz). In the following, we will denote this visual temporal frequency by s , for stimulus. The goal of the animal was to accurately classify each visual stimulus as belonging to the low frequency class (L , for gratings with frequencies from 0.25 to 2.12 Hz) or to the high frequency class (H , from 2.12 to 4 Hz). Each of the 9 frequency values had an equal chance of being presented in a trial. The central value was randomly assigned to the H or L class, and therefore also H and L had an equal chance of being presented. The probabilistic structure of the task can also be restated as follows. On each trial, the stimulus class C is chosen with equal chance to be H or L , that is

$$p(C = H) = p(C = L) = \frac{1}{2} \quad (4.10)$$

Based on the class, the visual stimulus s presented to the animal is selected with probability

$$p(s|C) = \frac{\delta(s - s_0)}{2K + 1} + \frac{2}{2K + 1} \sum_{k=1}^K \delta(s - s_k^C) \quad (4.11)$$

where

$$s_0 = 2.12 \quad (4.12)$$

$$s_k^L \in \{0.25; 0.82; 1.32; 1.75\} \quad (4.13)$$

$$s_k^H \in \{2.48; 2.91; 3.41; 4\} \quad (4.14)$$

$$K = 4 \quad (4.15)$$

and for convenience with our derivations below we order the index k symmetrically around s_0 , such that $s_1^H = 2.48$, $s_2^H = 2.91$, etc and $s_1^L = 1.75$, $s_2^L = 1.32$, etc. Our task here is to compute the probability $P(\hat{C} = H | s, n)$ that the observer chooses the class “high visual frequency” ($\hat{C} = H$), when the frequency of the visual stimulus is s and the sound condition is n .

We start by modeling the process that on each trial gives rise to the measurement x , given the visual stimulus s and the sound condition n . We initially disregard the effect of the sound,

which we will add back later. We assume that a stimulus s gives rise to a certain pattern $\vec{r}$ of firing rates in V1, where r_i is the rate of the i -th neuron. The population response is normally distributed around a stimulus-dependent value $\vec{\mu}(s)$: $\vec{r} \sim \mathcal{N}(\vec{\mu}(s), \Sigma)$, where Σ is a generic (but stimulus-independent) covariance matrix.

The subject generates an internal representation (or *measurement*) x of the external stimulus s based on the activation pattern $\vec{r}$. We assume that this measurement is a projection of the vector of activities $\vec{r}$ on some measurement axis $y = \vec{u} + t\vec{v}$; in other words,

$$x = (\vec{r} - \vec{u}) \cdot \vec{v} \quad (4.16)$$

This is biologically plausible if, for example, the measurement x is the activity of a readout neuron with connectivity strength v_i with each neuron coding s . Ideally, the measurement axis y used by the subject is the one that allows to best discriminate the stimulus classes, therefore allowing for the best performance, but for the purpose of developing the model we do not need to make any particular assumptions on $\vec{u}$ and $\vec{v}$. Because x is an affine transformation of a normally-distributed random variable, it is itself a normally distributed random variable:

$$x \sim \mathcal{N}((\vec{\mu}(s) - \vec{u}) \cdot \vec{v}, \vec{v}^T \Sigma \vec{v}) \quad (4.17)$$

If the projection operation is itself noisy and introduces additional Gaussian noise of standard deviation τ , the distribution of the measurement becomes simply

$$x \sim \mathcal{N}((\vec{\mu}(s) - \vec{u}) \cdot \vec{v}, \vec{v}^T \Sigma \vec{v} + \tau^2) \quad (4.18)$$

In the following, we will lump both sources of variability into a single parameter ρ ,

$$\rho^2 = \vec{v}^T \Sigma \vec{v} + \tau^2 \quad (4.19)$$

We will conceptualize ρ as the general level of sensory noise present in the system for the purposes of this task. The distribution of the measurement will therefore be

$$x \sim \mathcal{N}((\vec{\mu}(s) - \vec{u}) \cdot \vec{v}, \rho^2) = \mathcal{N}(v(s), \rho^2) \quad (4.20)$$

where we have defined

$$v(s) = (\vec{\mu}(s) - \vec{u}) \cdot \vec{v} \quad (4.21)$$

for notational convenience.

To decide if a stimulus has high frequency (H) or low frequency (L), we assume that the subject performs Bayesian inference to compute a posterior probability over the classes H and L given the measurement and reports the class with highest posterior. In our case (see below for a detailed justification), this will be equivalent to finding a decision boundary x^* such that $P(H | x^*) = P(L | x^*)$, and reporting H if $x > x^*$ and L if $x < x^*$. On average over the distribution of the measurement given the stimulus, the probability of reporting a given class given the stimulus (which we may also call the psychometric function) will then be

$$p(\hat{C} = H | s) = \int_{x: p(H|x) > p(L|x)} p(x | s) dx \quad (4.22)$$

$$= p(x > x^* | s) \quad (4.23)$$

$$= \int_{x^*}^{\infty} p(x | s) dx \quad (4.24)$$

and $p(\hat{C} = L | s) = 1 - p(\hat{C} = H | s)$. We will now calculate x^* and consequently we will derive $p(\hat{C} = H | s)$.

Let's consider the simplest possible scenario for $\vec{\mu}$, namely the case where $\vec{\mu}$ is linear in s :

$$\vec{\mu}(s) = \vec{\alpha} + \vec{\beta}s \quad (4.25)$$

In this case, the mapping between s and x also becomes linear:

$$v(s) = (\vec{\mu}(s) - \vec{u}) \cdot \vec{v} = \vec{\alpha} \cdot \vec{v} + (\vec{\beta} \cdot \vec{v})s - \vec{u} \cdot \vec{v} = a + bs \quad (4.26)$$

with $a = (\vec{\alpha} - \vec{u}) \cdot \vec{v}$ and $b = \vec{\beta} \cdot \vec{v}$.

We will now show that when $\vec{\mu}$ is linear the symmetry of the problem and the shape of the response distributions imply that x^* is unique and $x^* = v(s_0)$.

If we compute the posteriors explicitly from the class-conditional stimulus distributions in Eq (4.11) we get

$$p(H|x) - p(L|x) = \frac{p(x|H)p(H)}{p(x)} - \frac{p(x|L)p(L)}{p(x)} \quad (4.27)$$

$$\propto p(x|H) - p(x|L) \quad (4.28)$$

$$= \frac{p(x|s_0)}{2K+1} + \frac{2}{2K+1} \sum_{k=1}^K p(x|s_k^H) - \frac{p(x|s_0)}{2K+1} - \frac{2}{2K+1} \sum_{k=1}^K p(x|s_k^L) \quad (4.29)$$

$$\propto \sum_{k=1}^K p(x|s_k^H) - p(x|s_k^L) \quad (4.30)$$

$$\propto \sum_{k=1}^K \exp \left[-\frac{(x - v(s_k^H))^2}{2\rho^2} \right] - \exp \left[-\frac{(x - v(s_k^L))^2}{2\rho^2} \right] \quad (4.31)$$

$$= \sum_{k=1}^K \exp \left[-\frac{(x - a - bs_k^H)^2}{2\rho^2} \right] - \exp \left[-\frac{(x - a - bs_k^L)^2}{2\rho^2} \right] \quad (4.32)$$

Note now that the symmetry of the class-conditional stimulus distributions in Eq (4.11) is such that

$$\begin{cases} (s_0 - s_k^H)^2 = (s_0 - s_k^L)^2 \forall k \\ (s - s_k^H)^2 < (s - s_k^L)^2 \forall k & \text{if } s > s_0 \\ (s - s_k^H)^2 > (s - s_k^L)^2 \forall k & \text{if } s < s_0 \end{cases} \quad (4.33)$$

Also, for any value of x and s ,

$$x - v(s) = a - bv^{-1}(x) - a - bs = b(v^{-1}(x) - s) \quad (4.34)$$

which means that, if $b > 0$ (and simply reversing the inequalities if $b < 0$),

$$\begin{cases} (x - v(s_k^H))^2 = (x - v(s_k^L))^2 \forall k \\ (x - v(s_k^H))^2 < (x - v(s_k^L))^2 \forall k & \text{if } x > v(s_0) \\ (x - v(s_k^H))^2 > (x - v(s_k^L))^2 \forall k & \text{if } x < v(s_0) \end{cases} \quad (4.35)$$

Therefore,

$$\begin{cases} p(H|x) = p(L|x) & \text{if } x = v(s_0) = a + bs_0 \\ p(H|x) > p(L|x) & \text{if } x > v(s_0) = a + bs_0 \\ p(H|x) < p(L|x) & \text{if } x < v(s_0) = a + bs_0 \end{cases} \quad (4.36)$$

Thus as a consequence of the symmetry of the problem the decision boundary x^* is simply

$$x^* = v(s_0) = a + bs_0 = (\vec{\alpha} + \vec{\beta}s_0 - \vec{u}) \cdot \vec{v} \quad (4.37)$$

Hence:

$$p(\hat{C} = H | s) = p(x > x^* | s) = \int_{x^*}^{\infty} p(x | s) dx \quad (4.38)$$

$$= \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi\rho^2}} \exp \left[-\frac{(x - (\vec{\mu}(s) - \vec{u}) \cdot \vec{v})^2}{2\rho^2} \right] dx \quad (4.39)$$

$$= \Phi \left[\frac{(\vec{\mu}(s) - \vec{u}) \cdot \vec{v} - x^*}{\rho} \right] \quad (4.40)$$

$$= \Phi \left[\frac{(\vec{\mu}(s) - \vec{u}) \cdot \vec{v} - (\vec{\mu}(s_0) - \vec{u}) \cdot \vec{v}}{\rho} \right] \quad (4.41)$$

$$= \Phi \left[\frac{(\vec{\mu}(s) - \vec{\mu}(s_0)) \cdot \vec{v}}{\rho} \right] \quad (4.42)$$

$$= \Phi \left[\frac{\vec{\beta}(s - s_0) \cdot \vec{v}}{\rho} \right] \quad (4.43)$$

$$= \Phi \left[\frac{s - s_0}{\sigma} \right] \quad (4.44)$$

with

$$\sigma = \frac{\rho}{\vec{\beta} \cdot \vec{v}} \quad (4.45)$$

As shown in Fig 4.4, we model the effect of an external sound stimulus n as a suppression of the activity of V1 neurons. This suppression may result from direct inhibition of V1 neurons by auditory neurons, but it could arise from indirect effects as well. We model it by multiplying the mean response $\vec{\mu}(s)$ of the neurons by a sound-dependent factor γ_n , while keeping fixed the decision boundary x^* learned during the training sessions. By substituting

$\vec{\mu}(s)$ with $\gamma_n \vec{\mu}(s)$ in the integral above, we get:

$$p(\hat{C} = H | s, n) = \Phi \left[\frac{(\gamma_n \vec{\alpha} + \gamma_n \vec{\beta} s - \vec{u}) \cdot \vec{v} - x^*}{\rho} \right] = \Phi \left[\frac{\gamma_n s - s_0 + (\gamma_n - 1)\lambda}{\sigma} \right] \quad (4.46)$$

with

$$\lambda = \frac{\vec{\alpha} \cdot \vec{v}}{\vec{\beta} \cdot \vec{v}} \quad (4.47)$$

Now, if the circuit computing x is adapted to extract the maximum amount of information from $\vec{r}$ in this task, at least in the case where the noise for $\vec{r}$ is isotropic, $\vec{v}$ will be parallel to $\vec{\beta}$ (this may not be true in presence of nontrivial noise correlation structure). On the other hand, if $\vec{\alpha}$ is a random high dimensional vector with no particular relation to $\vec{v}$, we can assume that with high probability $\vec{\alpha} \cdot \vec{v} = 0$. Therefore, under these simple assumptions, $\lambda = \vec{\alpha} \cdot \vec{v} / \vec{\beta} \cdot \vec{v} = 0$. In this case, the expression for the probability of reporting H in presence of auditory stimulus n reduces to

$$p(\hat{C} = H | s, n) = \Phi \left[\frac{\gamma_n s - s_0}{\sigma} \right] \quad (4.48)$$

where γ_n and σ are free parameters of the model. Finally, we note that while the linearity assumption for $\vec{\mu}(s)$ made in this section can seem restrictive, the derivations above remain valid as long as $\vec{\mu}(s) \cdot \vec{v}$ is (approximately) linear over the range of values of s that are relevant in the task.

Lapse rates. We extended the ideal observer to account for lapses by defining two lapse rates ε_H and ε_L , such that on each trial the observer has a probability $1 - \varepsilon_H - \varepsilon_L$ of actually engaging in the task, a probability ε_H of reporting H independently of the stimulus, and a probability ε_L of reporting L independently of the stimulus. Therefore,

$$p(\hat{C} = H | s, n) = \varepsilon_H + (1 - \varepsilon_H - \varepsilon_L) \Phi \left[\frac{\gamma_n s - s_0}{\sigma} \right] \quad (4.49)$$

and

$$p(\hat{C} = L | s, n) = 1 - p(\hat{C} = H | s, n) \quad (4.50)$$

4.4.9 Inferring the model parameters from experimental data

We used the ideal observer model derived above to describe the data from our experiment. According to the formula for $p(\hat{C} = H | s, n)$, the behavior of each rat r is characterized by some parameters $\theta^r = \{\sigma^r, \{\gamma_n^r\}, \epsilon_L^r, \epsilon_H^r\}$. We inferred the value of these parameters using a hierarchical Bayesian approach (Gelman et al., 2014a; Kruschke, 2015), for which we now give a brief conceptual overview (refer to A.3 Fig for a graphical depiction).

Denote by D the set of recorded behavioral data, with D_t^r being the choice (H or L) of rat number $r \in \{1, \dots, R\}$ on trial number $t \in \{1, \dots, T_r\}$. We start by defining, for any rat, a prior probability for the parameters:

$$p(\theta^r) = p(\sigma^r) p(\epsilon_L^r, \epsilon_H^r) \prod_n p(\gamma_n^r) \quad (4.51)$$

We also define a likelihood function for the parameters θ^r . The likelihood is simply the probability of observing the rat choices $D^r = \{D_t^r\}_{t=1}^{T_r}$, given a certain value of θ^r and of the visual and auditory stimuli provided to the rat on all trials (respectively, $s^r = \{s_t^r\}_{t=1}^{T_r}$ and $n^r = \{n_t^r\}_{t=1}^{T_r}$), seen as a function of θ^r :

$$\mathcal{L}^r(\theta^r) = p(D^r | \theta^r; s^r, n^r) \quad (4.52)$$

$$= \prod_{t=1}^{T_r} p(\hat{C} = D_t^r | \theta^r; s_t^r, n_t^r) \quad (4.53)$$

$$= \prod_{t: D_t^r = H} p(\hat{C} = H | \theta^r; s_t^r, n_t^r) \prod_{t: D_t^r = L} p(\hat{C} = L | \theta^r; s_t^r, n_t^r) \quad (4.54)$$

where $p(\hat{C} = H | s, n)$ and $p(\hat{C} = L | s, n)$ are given by Eq (4.49) and Eq (4.50). The likelihood function therefore encapsulates the ideal observer model, developed in the previous section. By applying Bayes' theorem, we compute a joint posterior probability distribution for the parameters:

$$p(\theta^r | D^r) \propto p(D^r | \theta^r) p(\theta^r) \quad (4.55)$$

(from now on, for convenience, we will omit the explicit dependency on s and n in the likelihood). However, since the rats belong to the same population and are subjected to the same experimental conditions, we assume that the parameters describing different animals will be related. In other words, gathering information about one rat tells us something about that specific rat, but should also inform us on what to expect about other rats, or rat behavior in general. We can build this assumption into the model by imposing that the parameters associated to different rats are all sampled from the same distribution, leaving

the parameters of this higher-level probability distribution to be also inferred from the data. Besides extracting “population-level” information from the data of each rat, this procedure will also help constraining the rat-level parameters to reasonable values given what we observed in the other rats, thus acting as a regularizer for the inference process. More formally, our posterior will be:

$$p(\theta, \eta \mid D) \propto p(D \mid \theta) p(\theta \mid \eta) p(\eta) \quad (4.56)$$

where $\theta = \{\theta^r\}_{r=1}^R$ are the rat specific parameters for all R rats, and η are the common (population) parameters.

More concretely, the priors we assign to the parameters ($p(\theta \mid \eta)$ in Eq (4.56)) are:

$$(\varepsilon_L^r, \varepsilon_H^r, 1 - \varepsilon_L^r - \varepsilon_H^r) \sim \text{Dirichlet}(1, 1, 1) \quad \forall r \quad (4.57)$$

$$\sigma^r \sim \text{Gamma}(k_\sigma, \theta_\sigma) \quad \forall r \quad (4.58)$$

$$\gamma_n^r \sim \mathcal{N}(\mu_{\gamma_n}, \sigma_{\gamma_n}^2) \quad \forall r, n \quad (4.59)$$

It follows that our population parameters η are the mean μ_{γ_n} and the standard deviation σ_{γ_n} of the γ_n parameters, and k_σ and θ_σ for the perceptual noise. Following standard practice (Gelman et al., 2014a; McElreath, 2020), we assign weakly informative priors $p(\eta)$ to the population-level parameters:

$$\mu_{\gamma_n} \sim \mathcal{N}(\mu = 0, \sigma^2 = 9) \quad \forall n \quad (4.60)$$

$$\sigma_{\gamma_n} \sim \text{Exponential}(\text{scale} = 3) \quad \forall n \quad (4.61)$$

$$k_\sigma \sim \text{Exponential}(3) \quad (4.62)$$

$$\theta_\sigma \sim \text{Exponential}(3) \quad (4.63)$$

Finally, in this hierarchical scheme we also need a global likelihood function ($p(D \mid \theta)$ in Eq (4.56)) which we obtain simply as the product of the rat-level likelihoods:

$$\mathcal{L}(\theta) = p\left(D \mid \{\vec{\gamma}_r, \sigma_r, \vec{\varepsilon}_r\}_{r=1}^R\right) \quad (4.64)$$

$$= \prod_{r=1}^R \prod_{t=1}^{T_r} p(\hat{C} = D_t^r \mid \theta^r; s_t^r, n_t^r) \quad (4.65)$$

Prior predictive check. We conducted a prior predictive check (Gelman et al., 2014a) to evaluate whether the model’s prior distributions were consistent with plausible observations, even before any actual data was considered. The check was performed by sampling 1000

sets of model parameters from their prior distributions and using them to generate synthetic data. The results showed that our assumed priors are flexible enough to allow a full range of behavior for the psychometric curves, and have no bias towards supporting the central claim of our paper, namely the dependence of the gain on the sound intensity (A.4 Fig).

Sampling the posterior distribution. We performed inference with a Markov Chain Monte Carlo (MCMC) method, using the NUTS algorithm (Hoffman and Gelman, 2014), as implemented in PYMC version 5.2.0 (Salvatier et al., 2016). This method samples parameters from the posterior distribution $p(\theta | D)$: with a high enough number of samples, the distribution of the samples matches closely the distribution of the posterior. We sampled 4 independent chains of 2000 draws each, 1000 of which we used for tuning the algorithm and subsequently discarded, and 1000 as actual samples of the target posterior distribution. The target acceptance probability parameter for NUTS was set at 0.99, and no divergences were detected during sampling.

Posterior predictive check. Following the model's inference, we performed a posterior predictive check (Gelman et al., 2014a) to evaluate the goodness-of-fit to the observed data. The check was performed by sampling a set of synthetic observations from each of the 1000 parameter posterior draws saved in the Markov chain. The results showed that the simulated data was consistent with the observed data, indicating that the model provides a robust description of the observed rats' behavior (A.5 Fig).

4.4.10 Model comparison

To verify that the derived model describes well the experimental data, we compared it with other models that could also reproduce the psychometric curves of the rats. To perform the comparison, we used a Leave-One-Out Cross-Validation (*LOO*) estimate of the Expected Log pointwise Predictive Density (*ELPD*) (Vehtari et al., 2017). *LOO* estimates how well, on average, the model describes data it hasn't been trained on. To do so, it computes how well a model trained on all the trials (across all animals) except one predicts the outcome of that trial, and averages this value over all the trials. The *ELPD* therefore automatically balances goodness of fit with model complexity, penalizing overly complex models that tend to overfit the noise in the data. More formally, the *LOO* estimate of the *ELPD* for a model $\mathcal{M}$ for which we can sample the posterior distribution with MCMC is defined as:

$$\text{ELPD}_{\text{LOO}}(\mathcal{M}) = \sum_{i=1}^N \log \left(\frac{1}{S} \sum_{j=1}^S p_{\mathcal{M}}(d_i | \theta_{(j; D_{-i})}) \right) \quad (4.66)$$

where $N = \sum_{r=1}^R T_r$ is the total number of trials in the data, S is the number of posterior samples in the Markov chain for model $\mathcal{M}$, $p_{\mathcal{M}}$ is the likelihood function of the model, d_i is the data from the i -th trial and $\theta_{(j;D_{-i})}$ is the j -th posterior sample of a Markov chain generated from all data except d_i (note that each posterior distribution sample, including $\theta_{(j;D_{-i})}$, is a high-dimensional vector containing one value for each parameter in the model). In Eq (4.66), the argument of the logarithm represents the likelihood of the model $\mathcal{M}$ for the datapoint d_i , averaged over the posterior distribution over θ (where this posterior is obtained without looking at datapoint d_i). Indeed, since the training of $\mathcal{M}$ results in a posterior distribution rather than a single parameter vector, all quantities related to $\mathcal{M}$ must be averaged over this distribution. For MCMC training, this average is computed as a sum over all sampled parameters, given that the samples follow the posterior distribution:

$$\frac{1}{S} \sum_{j=1}^S p_{\mathcal{M}}(d_i | \theta_{(j;D_{-i})}) \simeq \int p_{\mathcal{M}}(d_i | \theta) p(\theta | D_{-i}) d\theta = \langle p_{\mathcal{M}}(d_i | \theta) \rangle_{p(\theta|D_{-i})}$$

In practice, computing Eq (4.66) requires training the model N times, one per trial: given the large number of trials in our experiment, we approximated the leave-one-out procedure with pareto-smoothed importance sampling (PSIS-LOO) (Vehtari et al., 2017).

Computing the ELPD_{LOO} for different models gives us a relative measure of how well they describe the data. As shown in the inset of Fig 4.5A, the model with the highest ELPD is the one with 5 distinct γ parameters, one for each sound condition whose power is in a certain range (see main text for details).

We also compared this model with a model that incorporates the additional λ parameter discussed in the ideal observer derivation above (Eq (4.47)). This parameter is included with an analogous hierarchical structure to that of σ in the main model discussed above, such that each rat has its own λ^r , with

$$\lambda^r \sim \mathcal{N}(\mu_{\lambda}, \sigma_{\lambda}^2) \quad \forall r \quad (4.67)$$

$$\mu_{\lambda} \sim \mathcal{N}(\mu = 0, \sigma^2 = 9) \quad (4.68)$$

$$\sigma_{\lambda} \sim \text{Exponential}(3) \quad (4.69)$$

This comparison allowed us to test the hypothesis that, to the extent that the assumptions about stimulus encoding made in that section are valid, the vector of resting activities of neurons $\vec{\alpha}$ has no effect on the projection of measurement. The resulting $(5\gamma 1\lambda 1\sigma)$ model performance is indeed comparable to the $(5\gamma 1\sigma)$ model performance ($\Delta \text{ELPD} = 0.74 \pm 0.9$; posterior mean $\pm$ std. err.). Moreover, the inferred population value of λ in the $(5\gamma 1\lambda 1\sigma)$ model is consistent with 0: the posterior estimate is $\mu_{\lambda} = -0.17 \pm 0.19$ (mean $\pm$ std. err.),

and the probability of direction is $p(\mu_\lambda < 0|D) = 0.821$. The probability of direction is the proportion of the posterior distribution for a parameter that has the same sign as the median, and in simple regression scenarios it is related to a frequentist two-sided p-value ($p \simeq 2(1 - pd)$) (Makowski et al., 2019).

Data availability statement

The data and source code for model fitting and generating figures of this manuscript are available in a publicly accessible repository on Zenodo (doi:[10.5281/zenodo.17280352](https://doi.org/10.5281/zenodo.17280352)).

Chapter 5

A hierarchical Gaussian process model to study the activity of neuronal populations

5.1 Introduction

In the previous chapters, we examined how humans select between internal models to interpret noisy sensory data, and we described perceptual behavior in rats using a Bayesian model built around a potential neural implementation. However, while behavioral modeling strongly suggests the use of internal models, confirming their existence or adequately characterizing the kind of model implemented requires looking directly at the physiological substrate.

To investigate these representations, neuroscience has developed a diverse toolkit of methods offering each a distinct window into how values and quantities are coded in neural activity. Historically, electrophysiological recordings have been the gold standard, offering the temporal precision necessary to capture individual spiking events. It was through single-unit recordings that the field established the presence of hippocampal place cells, neurons that define a cognitive map of physical location ([O'Keefe and Dostrovsky, 1971](#)). Similarly, multi-unit recordings led to the finding that spontaneous population activity in the visual cortex statistically converges to match the structure of natural scenes, suggesting that the statistics of the visual environment are represented in an optimal internal model ([Berkes et al., 2011](#)). Similarly, calcium imaging techniques have allowed researchers to visualize the simultaneous activity of vast neuronal populations, shifting the focus from single or few neurons to ensemble statistics. For example, two-photon calcium imaging has been instrumental in realizing that some neurons in the visual cortex respond to the discrepancy between actual and predicted visual flow, suggesting that the visual cortex may use predictive coding strategies ([Keller et al., 2012](#)).

Recently, the field has seen a surge in the adoption of fiber photometry, a method that has rapidly become essential for dissecting the circuit dynamics of freely moving animals. Its rising popularity stems from its ability to record population-level activity from genetically defined cell types in deep brain structures, allowing molecular specificity in recording during complex behavior. The technique has been used to characterize dopamine reward prediction error signals in mice, revealing the precise temporal dynamics of dopamine release associated with cue, reward, and reward omission, providing deeper insights into reinforcement learning mechanisms ([Patriarchi et al., 2018](#)). Related work has dissected the dopamine signaling heterogeneity in subpopulations of midbrain dopamine neurons, showing distinct activity patterns linked to rewards, aversive stimuli, and accelerations and decelerations, challenging the notion of dopamine neurons as a homogeneous group ([Azcorra et al., 2023](#)). It has also been used to localize neuronal populations encoding winning and losing social behaviors ([Choi et al., 2024](#)), understand how associative learning uses a differential distribution of information across populations ([Avvisati et al., 2024](#)), and, in general, how population activity in particular cell groups relates to components of complex behavior ([Simpson et al., 2024](#)).

Fiber photometry's high success and applicability across experimental platforms have fueled a rapid expansion of its use in recent years. However, this quick adoption has outpaced standardization, leading to a lack of consensus on best practices for both experimental design and analytical methods ([Simpson et al., 2024](#)). In particular, many of the analysis procedures adopted derive from approaches designed for different methods, such as EEG and electrophysiology ([Loewinger et al., 2025](#)). Currently, the most common analytical strategy involves event-aligned analysis, where signals are time-locked to discrete behavioral events (e.g., the moment a reward is delivered) and averaged across trials. From the aligned traces, summary statistics (such as the Area Under the Curve) are used to perform standard hypothesis tests ([Kielbinski and Bernacka, 2024](#)). While parsimonious, this approach faces significant limitations: it relies on arbitrary time-window selection, which can obscure some temporal dynamics, and the process of trial-averaging can dilute important heterogeneity in learning rates or individual animal responses.

To address these complexities, the field has increasingly moved toward linear regression and Generalized Linear Models (GLMs). Regression is often implemented by running separate models at each time point (pointwise regression). However, standard pointwise approaches often fail to account for the hierarchical structure of the data (e.g., ignoring the shared structure and individual variability of the subjects) and do not exploit the correlation between signal values across time points. To capture the continuous evolution of the signal without discarding trial-level information, the use of Functional Linear Mixed Models (FLMM) has been proposed ([Loewinger et al., 2025](#)). Unlike standard summary analyses or

pointwise regressions, FLMM treats the photometry signal from every trial as a continuous functional data object. While it is based on a pointwise regression, it uses mixed models to account for hierarchical dependencies such as between-animal heterogeneity or within-session correlations. Moreover, it uses a smoothing techniques that exploit the inherent autocorrelation of the biological signal to provide continuous, non-noisy outputs.

Despite the significant conceptual advancements introduced by the FLMM framework, the method is not without its limitations. The approach relies on a complex, multi-stage estimation procedure that breaks the analysis into separate steps rather than treating the generative process organically. By treating each time point as independent during the initial regression, the model fails to take advantage of the information from neighboring time points, effectively ignoring the temporal continuity of the biological signal until the post-hoc smoothing phase. Another problem is that many behavioral tasks include multiple events of interest. If events happen at fixed times (e.g., if reward delivery happens after a fixed time window from the trial onset), time-locking to one event will inherently time-lock to the other events. However, this is not the case when trials are self-paced by the subject or variable intervals are imposed. In such cases, usual approaches to deal with the problem are to either separately align the signal to each event of interest or to align it to all the events in a trial by time-warping the signal between events. The first solution, however, requires running multiple regressions for the same data, while the second one can introduce artifacts. Also, neither solution gives conclusive information on the contribution of each event to the neural activation, as the activities related to multiple events remain entangled.

In this chapter, we propose a novel method based on Bayesian inference that addresses both issues. By using a hierarchical model of Gaussian processes (GP, [MacKay et al., 1998](#)), we explicitly describe the continuous nature of biological signals while simultaneously capturing both subject-specific variability and shared population structure. By assigning distinct GPs to different task events, we circumvent the artifacts introduced by traditional time-warping or rigid time-locking; this allows us to characterize not only if external variables affect neural activity, but also to which specific event that activity is tied. Finally, the nature of Bayesian inference allows us to naturally obtain measures of uncertainty for our results, enabling us to perform direct probabilistic comparisons between conditions without relying on asymptotic assumptions or ad hoc procedures.

We apply our methodology to photometric data from an experiment involving two groups of mice, modified to target two distinct neural populations of striatal interneurons. Our goal is to recognize which variables, external or internal, affect the activity of these populations during a value-based decision-making task. The experiment presents two main events, a *Go* cue and a reward delivery, allowing us to explore the capabilities of the proposed method.

5.2 Experiment

The value-based decision-making task was designed and performed by our collaborators Evan Iliakis and Marc Fuccillo at the Fuccillo Lab at the University of Pennsylvania. The experimental procedure is detailed in [Linares-García et al. \(2025\)](#). We give here a brief overview of the task, performed by mice with a custom platform, and a few details on the photometry recording procedure.

Each trial began with the extension of a retractable joystick that could either be pulled or pushed. The event was followed by a 100 ms delay and the onset of a "Go cue" light, which signaled the beginning of the trial. Head-fixed water-restricted mice were then required to displace the joystick with their right forepaw by at least 3 mm from baseline within a 10 s response window to register a choice. The rewarded direction (either "push" or "pull") was chosen at random, with one direction being four times more probable than the other. The session was divided into blocks where the same direction had the higher probability of reward. Each block lasted for a minimum of 17 rewarded trials, after which each trial had a 40% probability to end the block and start a new one with the opposite favored direction.

There were four possible trial outcomes: a *rewarded trial*, which delivered a reward (10 μ l water with 10% sucrose) followed by a variable inter-trial interval (ITI) of 2.5–8 seconds; an *unrewarded trial*, signaled by lights turning off followed by the same ITI; an *omission*, occurring if no choice was made within the window; and a *premature trial*, occurring if the choice was registered before the Go cue. Both omissions and premature trials resulted in the delivery of a white noise punishment signal, lights turning off, and a prolonged 15-second ITI.

The behavioral protocol included two phases. The first phase proceeded as described. The second phase included a varying amount of reward (2, 4, or 8 μ l) across blocks to assess motor vigor. For the current analysis, we only focused on the effect of probabilistic reversal, neglecting the values of varying reward.

Two groups of mice performed the task, each with a distinct genetic modification. These modifications were designed to render specific neuronal populations of the striatum accessible to fiber photometry. The first group was targeted to monitor low-threshold spiking interneurons (LTSIs), a class of somatostatin-expressing GABAergic interneurons that regulate corticostriatal integration and modulate striatal dopamine signaling implicated in goal-directed learning ([Holly et al., 2019, 2021](#)). The second group was targeted to monitor tyrosine hydroxylase-expressing interneurons (THINs), GABAergic interneurons that provide inhibitory inputs to LTSIs and affect the ability to update behavior based on new stimuli ([Assous et al., 2017; Kaminer et al., 2019](#)). Monitoring the activity of both interneu-

ron classes during behavior allows to relate their specific contributions and differences in decision-making and value assessment.

5.3 Bayesian Hierarchical Model

To analyze the data from the task, we first standardized the photometry signals individually for each subject to account for variability in fluorescence intensity and optic fiber placement. For each trial, we defined two alignment events: the Go cue (trial onset) and the Outcome (reward delivery or omission).

A set of predictor variables was recorded for each trial, characterizing task dynamics, the subject's internal state, and motor execution. The task dynamics variables included the outcome of both the current and the immediately preceding trial (rewarded vs. unrewarded), the duration of the inter-trial interval (ITI) of the previous trial, and the preferred direction (push vs. pull) of the current trial. To capture the animal's internal state, a Q-learning model was fitted to the behavioral data to compute the Reward Prediction Error (RPE), encoding the discrepancy between expected and realized rewards, (The computation of RPE has been done by our collaborators following the procedure defined in [Alabi et al., 2020](#)). This was complemented by the reward rate, representing the density of rewards per minute obtained preceding the trial. Finally, motor execution was recorded through continuous monitoring of the joystick position; from these traces, the direction of movement (push vs. pull), the peak joystick displacement (μm), and the response latency (i.e., the time in ms from the start of the trial to threshold crossing) were extracted. We did not directly include the joystick position in time in any analysis.

We categorized the behavioral predictors into binary variables (Previous Outcome, Current Outcome, Reward Direction, Push/Pull Choice) and continuous variables (Previous ITI, Reward Prediction Error, Reward Rate, Choice Latency, Peak Joystick Displacement). Binary variables were encoded as 0 (i.e., predictor absent) or 1 (i.e., predictor present). Continuous variables were standardized using the global mean and standard deviation across all subjects and trials.

Consistent with established approaches for analyzing photometry data ([Loewinger et al., 2025](#)), we adopted a Linear Regression framework to model the temporal dynamics of the signal. In a standard regression at a single time point t , for a subject s , the standardized signal $y_s(t)$ can be modeled as:

$$y_s(t) = \alpha^s(t) + \sum_p \beta_p^s(t) \cdot x_{s,p}^{(trial)} + \varepsilon \quad (5.1)$$

where $x_{s,p}^{(trial)}$ is the standardized value of predictor p for the current trial, $\alpha^s(t)$ represents the baseline response time at time t and the term ε represents residual noise.

To extend this into a Bayesian regression framework (Gelman et al., 2014a) and ensure that the resulting response functions are biologically plausible (i.e., to introduce smoothness), we placed a Gaussian Process prior on the coefficients $\alpha(t)$ and $\beta(t)$ (MacKay et al., 1998). Furthermore, since each trial involves multiple events (Go cue and Outcome) that have no fixed delay in time, we expanded the model to account for relative timing. We modeled the signal as the sum of responses aligned to each event:

$$y_s(t) = \sum_{E \in G, O} \left(\alpha_E^s(t - t_E) + \sum_p x_{s,p}^{(trial)} \cdot \beta_{p,E}^s(t - t_E) \right) + \varepsilon \quad (5.2)$$

where E represents the type of event (Go cue, G , or Outcome, O) and t_E is the time of that event within the trial. To simplify the notation, from here on we will refer to $\alpha_E^s(\tau)$ as $\beta_{0,E}^s(\tau)$, introducing a fictional predictor $p = 0$ that has value $x_{s,0}^{(trial)} = 1$ for all trials and subjects.

To capture both the overall group behavior of the mice and subject-specific variability, we used a hierarchical design. For each event type E , we defined a group-level intercept and a predictor responses $\beta_{E,p}(\tau)$, as zero-mean Gaussian Processes:

$$\beta_{E,p}(\tau) \sim GP(0, K(t, t'; \theta^g)) \quad \forall p \quad (5.3)$$

and, similarly, subject-specific responses as Gaussian Processes centered on the group-level dynamics:

$$\beta_{E,p}^s(\tau) \sim GP(\beta_{E,p}(\tau), K(t, t'; \theta^s)) \quad \forall p \quad (5.4)$$

Where $K(t, t'; \theta)$ represents the specific kernel function of the GPs, encoding our assumptions about the smoothness and variability of the function, θ^g are its group-level parameters and θ^s its subject-level parameters.

Compared to other methods used to study photometric data, this regression architecture lets us account at once for the presence of multiple events, subject variability, and the continuous nature of the data, while also allowing us to evaluate the uncertainty of the inferred values through the posterior distribution. However, it presents a significant computational challenge: Gaussian Processes scale poorly with sample size. The time complexity of $O(N^3)$ makes exact inference practically impossible, given the high number of trials and dense sampling rate of this dataset. We adopted a series of measures to mitigate this problem.

First, we restricted the temporal support of the response functions to a shorter window. This is common practice in regression models for time-resolved regression models, including non-GP-based ones (see for instance Balzani et al., 2020; Choi et al., 2024; Park et al., 2014; Parker et al., 2022; Pillow et al., 2008; Runyan et al., 2017), which can be easily motivated by considering what would happen if we allowed the functions $\beta(\tau)$ to span the maximum possible length over time. This would mean that the functions would be at least as long as the longest trial¹, which would be problematic not only for the long time required by the inference, but also for its quality: the tails of the inferred functions would effectively only describe the longest trials, overfitting their shape or regressing to the prior. We therefore fixed a three-second window around each event: $[0, 3]$ s relative to the Go cue and $[-1, 2]$ s relative to the Outcome.

Second, we reduced the dimensionality of the input data. The photometry signal was sampled at 1 kHz: we binned it to reduce the 3000 photometry samples of the three-second window of each function to 100. In practice, we averaged every 30 samples of the signal to keep the downsampled data smooth and prevent potential aliasing effects.

However, even with these reductions, the computational cost of standard GP inference remained prohibitive. To overcome this, we implemented a Hilbert Space Gaussian Process (HSGP) approximation (Riutort-Mayol et al., 2023; Solin and Särkkä, 2020; see also Section A.1.1). This approximation allows us to substitute each Gaussian Process with a linear combination of M sinusoidal basis functions, reducing the time complexity to $O(MN + M)$. The resulting function is given by:

$$f(\tau) = \sum_{k=1}^M \tilde{S}_k(\theta) \phi_k(\tau) z_k \quad (5.5)$$

where $\tilde{S}_k(\theta)$ is a function that depends on the hyperparameters θ of the GP kernel function $K(t, t'; \theta)$ (for details, see Section A.1.1), $\phi_k(\tau)$ is a basis function whose shape only depends on the index k and the time window over which the inference is performed, and the spectral coefficients $z_k \sim \mathcal{N}(0, 1)$ are regression parameters. When inferring multiple functions, as in our case, the added advantage of HSGP is that they all share the same basis functions: only the hyperparameters and the values of z_k have to be inferred. This greatly reduces the parameter space and the model complexity.

As in the case of GPs, the HSGP prior is determined by the choice of the covariance kernel $K(t, t'; \theta)$. Rather than the commonly used Squared Exponential kernel (that we used

¹Since the Outcome-aligned functions are centered on the outcome time, they develop both in negative time, to describe the activity preceding the event, and positive time. To have the function span all the possible times, it has to be very long in negative times, for trials that have an extended delay in the choice, and in positive times, for trials that have a long post-trial ITI, making the overall function possibly longer than the longest trial.

in Chapter 3), we selected the Matern 5/2 kernel (Williams and Rasmussen, 2006):

$$k(t, t'; \eta, l) = \eta^2 \left(1 + \frac{\sqrt{5}|t - t'|}{l} + \frac{5(t - t')^2}{3l^2} \right) \exp \left(-\frac{\sqrt{5}|t - t'|}{l} \right) \quad (5.6)$$

While the Squared Exponential generates infinitely differentiable functions, resulting in a degree of smoothness that is often unrealistic for biological data, the Matern 5/2 kernel allows for rougher dynamics. Given the nature of neural signals, characterized by rapid changes rather than perfectly smooth oscillations, we considered this kernel more appropriate.

A known limitation of the HSGP approximation is that the sinusoidal basis functions require Dirichlet boundary conditions (i.e., the function has to have a value of zero at the limits of the domain). To prevent this from distorting the inferred neural responses at the beginning and end of our chosen time windows, we extended the GP window by a factor of $c = 1.5$ beyond the three-second window of our functions, following standard practice (Riutort-Mayol et al., 2023). We also fixed the number of basis functions to $M = 50$.

We assigned the following priors to the hyperparameters of the population-level functions $\beta_{p,E}(\tau)$, i.e., to the signal amplitude η_p^g and the length scale l_p^g :

$$\eta_p^g \sim \text{Exponential}(\lambda = 0.5) \quad \forall p \quad (5.7)$$

$$l_p^g \sim \text{Gamma}(\alpha = 4, \beta = 3) \quad \forall p \quad (5.8)$$

Notice that the same hyperparameter is shared across functions for each predictor. Similarly, subject-level functions have priors:

$$\eta_p^s \sim \text{Exponential}(\lambda = 0.5) \quad \forall p \quad (5.9)$$

$$l_p^s \sim \text{Gamma}(\alpha = 4, \beta = 3) \quad \forall p \quad (5.10)$$

Notice that these hyperparameters are the same for all subjects, as was the case with the GPs in Chapter 2. As showed in Eq. 5.5, the group-level intercept and parameter functions are given by:

$$\beta_{p,E}(\tau) = \sum_{k=1}^M \phi_k(\tau) \gamma_{p,E,k} \quad (5.11)$$

Where the prior of the parameter $\gamma_{p,E,k}$, incorporating both z_k and $\tilde{S}_k$, is:

$$\gamma_{p,E,k} \sim \mathcal{N}(0, \tilde{S}_k(\eta_p^g, l_p^g)) \quad (5.12)$$

To take advantage of the structure of the HSGP functions, and in particular of the common $\phi_k(\tau)$ functions, we parameterized the subject-level responses as a sum of their mean and a zero-centered deviation $\Delta\beta_{p,E}^s(\tau)$:

$$\beta_{p,E}^s(\tau) = \beta_{p,E}(\tau) + \Delta\beta_{p,E}^s(\tau) \quad (5.13)$$

$$= \sum_{k=1}^M \phi_k(\tau) \gamma_{p,E,k} + \sum_{k=1}^M \phi_k(\tau) \gamma_{p,E,k}^s \quad (5.14)$$

$$= \sum_{k=1}^M \phi_k(\tau) (\gamma_{p,E,k} + \gamma_{p,E,k}^s) \quad (5.15)$$

Where $\gamma_{p,E,k}^s$ are the subject deviation parameters. To avoid identifiability issues with the model, we used a Zero-Sum Normal distribution as prior for these values:

$$\vec{\gamma}_{p,E,k} \sim \mathcal{N} \left(0, \tilde{S}_k(1, l_p) \left(\mathbf{I} - \frac{1}{N_{subj}} \mathbf{J} \right) \right) \forall p, E, k \quad (5.16)$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{J}$ is a matrix of ones, and N_{subj} is the number of subjects: the samples of all subjects from this singular multivariate distribution always sum to 0. Without this trick, the population average reverts to its prior, giving a consistently null result after inference.

Finally, we modeled the residual noise ε using a Student-T likelihood:

$$y_{obs}(t) \sim \mathcal{T}(\mathbf{v}, \mu(t), \sigma) \quad (5.17)$$

where $\mu(t)$ is the predicted signal from the regression and $\mathbf{v}$ represents the degrees of freedom. The Student-T distribution has heavier tails than the Normal distribution, making the inference robust to outliers. The priors assigned to the group-level and subject-level standard deviation were:

$$\sigma_\sigma \sim \text{HalfNormal}(\sigma = 0.2) \quad (5.18)$$

$$\sigma^s \sim \text{HalfNormal}(\sigma = \sigma_\sigma) \forall s \quad (5.19)$$

We didn't give $\mathbf{v}$ a hierarchical structure. Its prior was set to:

$$\mathbf{v} \sim \text{Gamma}(\alpha = 10, \beta = 0.5). \quad (5.20)$$

We performed inference with a Markov Chain Monte Carlo (MCMC) method, using the NUTS algorithm ([Hoffman and Gelman, 2014](#)) in PYMC version 5.25.1 ([Salvatier et al.,](#)

2016). To accelerate inference, we employed the JAX backend (Bradbury et al., 2018), specifically using NUTS as implemented by the BlackJAX library (Cabezas et al., 2024), on two Nvidia L4 GPUs with 24Gb of VRAM each.

For each inference run, we sampled 4 independent chains of either 2000 (for the inferences with only binary predictors) or 3000 (for the inferences with both binary and continuous predictors) draws each, half of which we used for tuning the algorithm and subsequently discarded, and half as actual samples of the target posterior distribution. The target acceptance probability parameter for NUTS was set at 0.95, and no divergences were detected during sampling.

5.4 Results

As of the time of writing, we can only present some initial results, as a complete analysis has not been concluded yet. Nevertheless, these preliminary findings proved promising.

We first performed the inference using only the four binary predictors: Previous Outcome, Current Outcome, Reward Direction, and Push/Pull Choice. We applied the regression to the LTSI group data (13 mice, 5786 trials) and the THIN group data (13 mice, 5909 trials), independently using the hierarchical architecture defined in the previous section (Fig. 5.1).

For both groups, the dominant predictor resulting from the regression is the Current Outcome (Fig. 5.1, third row). Both populations show a rapid change in activity following reward delivery, yet with opposite trends. LTSI interneurons show a sharp increase in activity after a rewarded outcome, significantly exceeding the moderate activity generated by an unrewarded one. Conversely, THIN interneurons display a marked suppression of activity following a reward, compared to a moderate increase following unrewarded trials. The strong response to the delivery of reward observed for LTSI is consistent with result in literature (Holly et al., 2019, 2021). Similarly, a reduced activity of THIN interneurons following a reward has been observed for single cells (Stanfield, 2023). Importantly, the Current Outcome coefficient for both populations of interneurons remained at baseline (zero) throughout the Go cue time window, as well as during the pre-reward phase of the Outcome alignment. This confirms that our model correctly respects the temporal structure of the task, showing no modulation by the outcome before it is revealed, and that it correctly disentangles the signals generated by the Go cue and the reward delivery.

The Push/Pull choice elicited a significant, albeit smaller, modulation in both populations (Fig. 5.1, last row). As a result of our encoding of the predictor, the trends show the response to a pulling motion. In LTSI neurons, we observed no significant modulation after the Go cue, while an observable effect is present in the Outcome time window: a first burst of

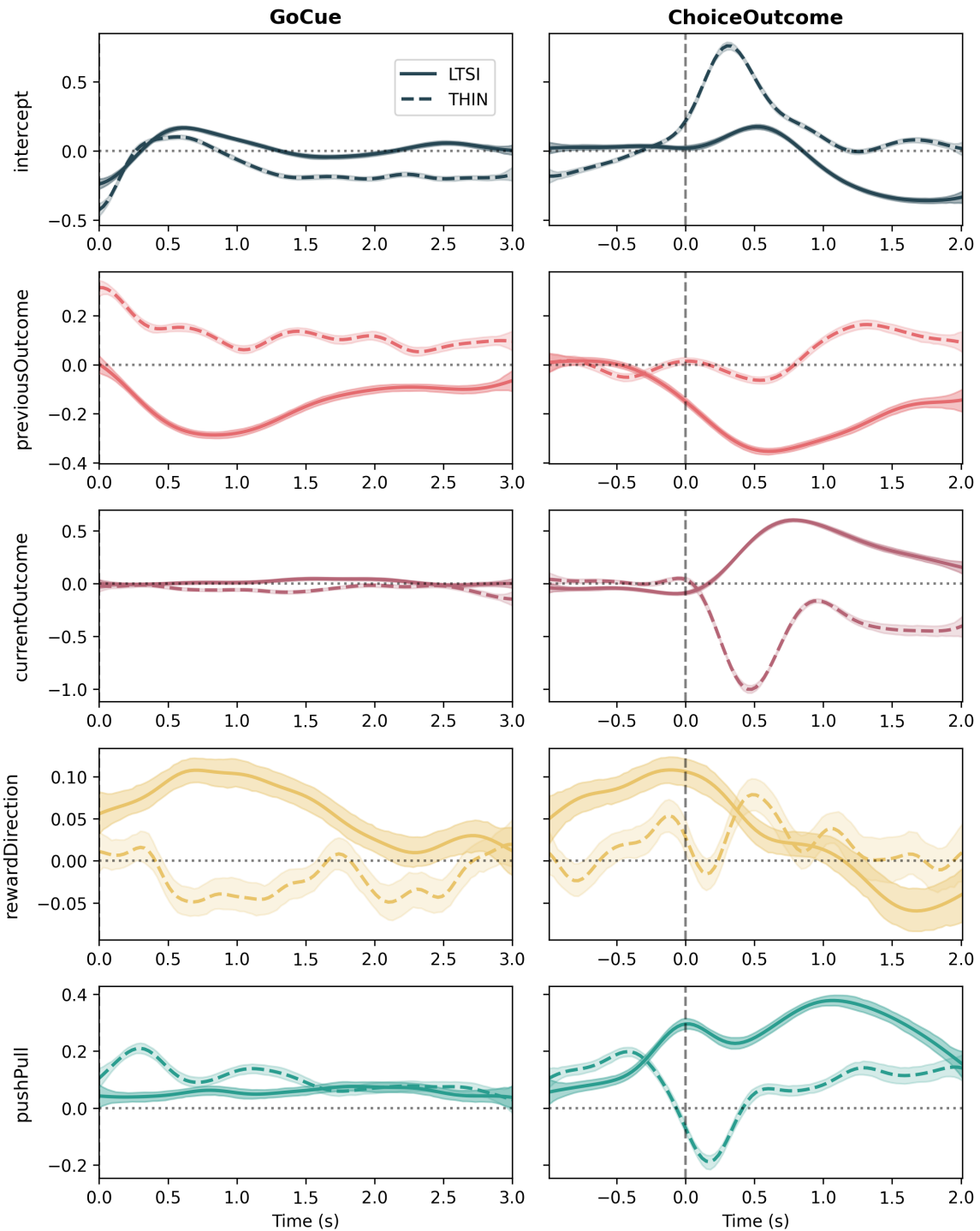


Fig. 5.1 Inferred responses in time of the LTSI (continuous lines) and THIN (dashed lines) neuronal population to binary-only predictors. Each row represents one of the variables included in the model. The left column shows the response after the start of the trial. The right column shows the response around the reward delivery. The time of reward delivery is marked by a vertical dashed line. On each box of the plot, the colored line represents the posterior mean of the inferred function, and the shaded area represents a 90% CI.

activity, peaked exactly at the moment of reward delivery, is followed by a second, slower activation. THIN neurons again displayed the inverse pattern: a sudden drop in activity at reward delivery, followed by a gradual recovery. Additionally, THIN neurons showed early-trial activity immediately following the Go cue, possibly reflecting movement planning or initiation.

Neither population showed a strong or consistent activation in response to Reward Direction (Fig. 5.1, fourth row). Although subjects presumably maintain an internal representation of the rewarding action, these striatal interneurons appear dedicated to encoding the specific action and outcome more than the abstract direction.

Finally, the Previous Outcome generated qualitatively different responses in the two groups (Fig. 5.1, second row). LTSI neurons displayed similar temporal dynamics after both the Go cue and reward delivery (a slow suppression followed by a rebound). In contrast, THIN neurons responded strongly only at trial onset, with activity gradually decaying; the previous outcome does not elicit a strong reaction during the reward phase, save for a mild increase after 1 s.

We later extended the model to include continuous behavioral predictors: Reward Prediction Error (RPE), Reward Rate, and Choice Latency. Due to the increased computational load, we excluded the less informative Reward Direction predictor and reduced the number of basis functions to 40 (Fig. 5.2).

The binary predictors yielded results largely consistent with the previous iteration. However, we observed minor fluctuations in the baseline (pre-outcome) periods for coefficients that were previously negligible. This likely reflects mild collinearity between some of the binary and continuous predictors. We can see the effect of these collinearity from the correlation between the posterior samples of the spectral coefficients z_k (see Eq. 5.5) for each pair of predictors (Figs. 5.3 and 5.4): other than the expected correlation between the intercept and the binary variables parameters (a consequence of the asymmetric $[0, 1]$ coding we used), there is some correlation between predictors (e.g. between Current Outcome and RPE), explaining the small effect on the binary predictors coefficients.

Overall, continuous predictors displayed a weaker effect on neural activation compared to the binary ones. We consider two potential explanations for this disparity. First, this could be a methodological effect. For example, it could stem from the aforementioned collinearity between predictors, as the model may attribute the bulk of the shared variance to the binary predictor; however, this would only be valid for a few of them, and the effect would be small. Alternatively, measurement noise in behavioral tracking can lead to attenuation bias, naturally suppressing the coefficients of continuous predictors relative to the binary ones.

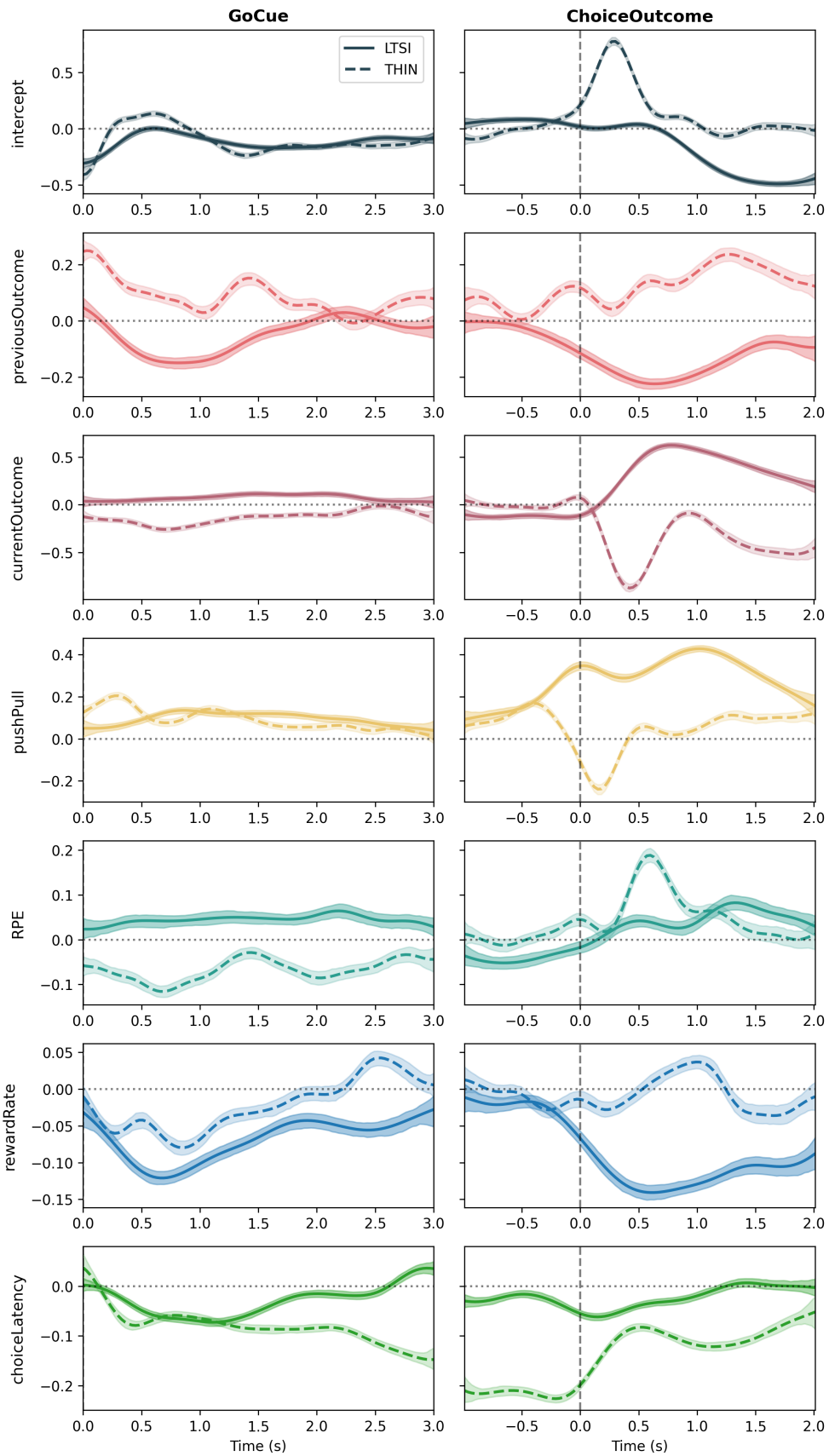


Fig. 5.2 Inferred responses in time of the LTSI (continuous lines) and THIN (dashed lines) neuronal population to binary and continuous predictors. Details as in Fig. 5.1.

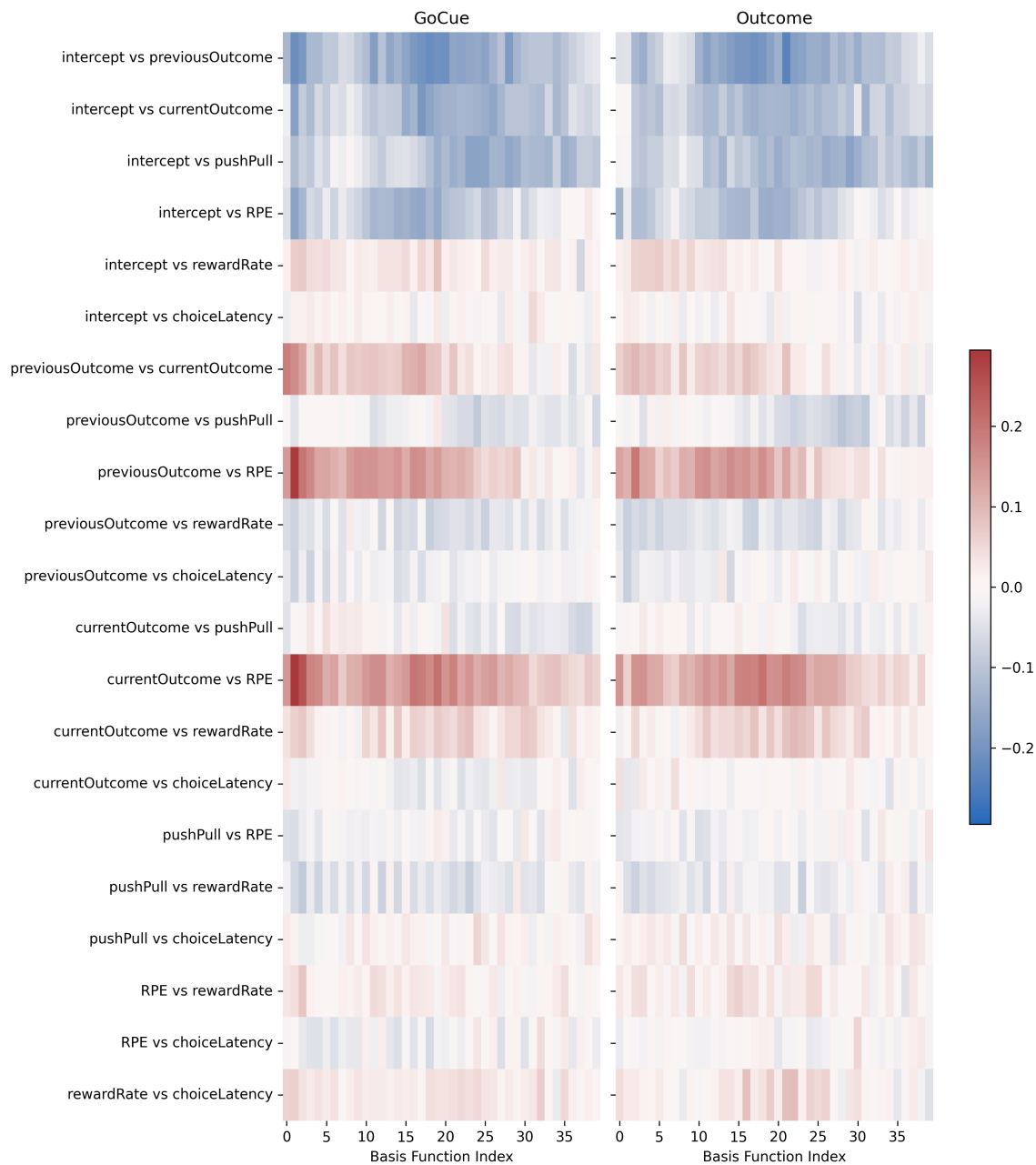


Fig. 5.3 Posterior correlations of spectral coefficients across behavioral predictors in LTSI. The heatmaps display the pairwise Pearson correlation coefficients between the population-level basis function weights (z_k) for the Go cue (left) and Outcome (right) kernels. Rows represent pairs of predictors, while columns correspond to the indices of the Hilbert space basis functions, ordered from low to high frequency. The color scale is common to both panels, centered at zero (white), with red indicating positive correlation and blue indicating negative correlation.

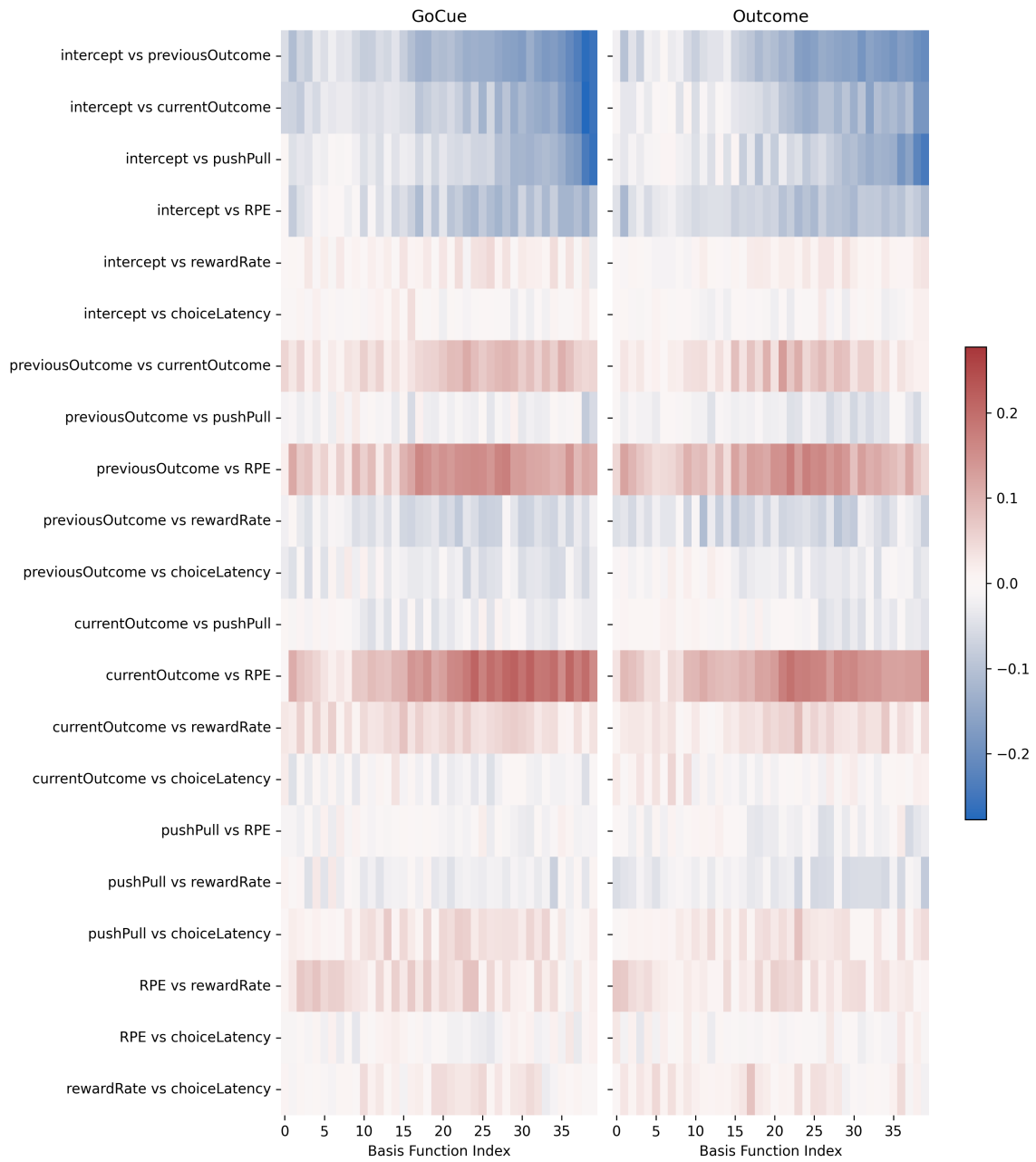


Fig. 5.4 Posterior correlations of spectral coefficients across behavioral predictors in THIN. The heatmaps display the pairwise Pearson correlation coefficients between the population-level basis function weights (z_k) for the Go cue (left) and Outcome (right) kernels. Rows represent pairs of predictors, while columns correspond to the indices of the Hilbert space basis functions, ordered from low to high frequency. The color scale is common to both panels, centered at zero (white), with red indicating positive correlation and blue indicating negative correlation.

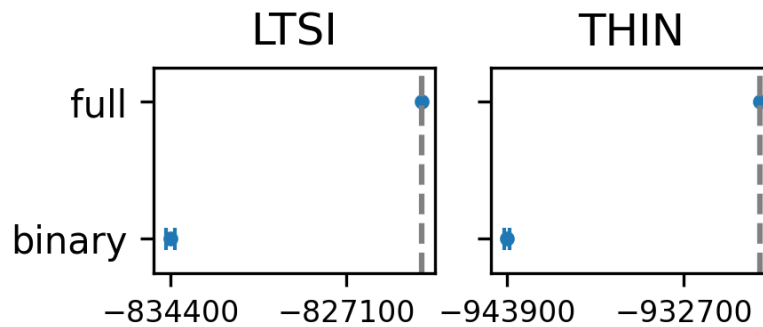


Fig. 5.5 **Model comparison between binary-only and complete models.** Pareto-Smoothed-Important-Sampling Leave-One-Out Cross-Validation (PSIS LOO CV) comparison between the binary-only model ("binary") and the model with both binary and discrete predictors ("full") for LTSI (left) and THIN (right). The x-axis shows the Expected Log Predictive Density (ELPD) of each description. The errorbars represent the standard error of the ELPD differences with the best model, $ELPD_{\text{best}} - ELPD_M$. The vertical dashed line marks the ELPD value of the best model.

Furthermore, the different coding applied to binary and continuous predictors, with the latter allowing for higher values than the former, could explain this difference.

Second, physiologically, this could suggest that continuous predictors give little contribution to the predictive power of the model, and could be disregarded without incurring a significant loss. We tested this hypothesis using Leave One Out Cross Validation (LOO-CV, see Section 2.4.4 and 4.4.10) between the regressions with only binary predictors and the one with binary and continuous (Fig. 5.5): for both LTSI and THIN data, the model including continuous variables showed a significantly higher performance in predicting new data (measured as Expected Log Predictive Density, ELPD). This, while not excluding that neurons' activity favours binary predictors, indicates that including the continuous ones leads to a significant improvement in the model's descriptive power.

Furthermore, an examination of the temporal structure of their coefficients reveals subtle but distinct modulations. RPE displays a mostly negligible effect on LTSI neurons, with some possible, minor modulation after reward delivery (Fig. 5.2, fifth row, continuous line). In THIN neurons, this modulation is more marked, with a clear peak shortly after the outcome of the trial is revealed (Fig. 5.2, fifth row, dashed line). Both populations have a low response to the Reward Rate, but LTSI displays a consistent pattern of gradual decrease in activity after both the Go cue and the outcome (Fig. 5.2, fifth row).

Finally, Choice Latency elicited the strongest reaction among continuous predictors for THIN interneurons (Fig. 5.2, last row, dashed line). Specifically, we observed a distinct negative deflection prior to the choice outcome. Given that latency is z-scored (where faster

trials correspond to negative values), this negative coefficient indicates that quicker responses elicit overall higher neural activity, while slower trials are associated with suppression. A similar, but weaker, trend is present in the LTISI group (Fig. 5.2, last row, continuous line).

Beyond interpreting the kernels individually, the GP regression allows us to perform a Posterior Predictive Check, reconstructing the shape of full trials based on their specific event timings and predictor values. This generative step allows us to verify whether the sum of our inferred components actually reproduces the macroscopic dynamics of the raw signal. By sampling parameters from the posterior distribution, we generated a range of simulated trials, creating a distribution of predicted traces that can be directly overlaid on the ground truth (Fig. 5.6). Visually, the model successfully captures the global structure and the relevant trends in the majority of the trials. The high noise fluctuations of the real data are well captured by the inferred variance.

This predictive accuracy is even more evident when aggregating data across all mice to examine specific behavioral conditions. When stratifying trials based on predictor states, distinguishing, for example, between trials where the Current Outcome is rewarded or not, the model faithfully reproduces the average trajectories observed in each condition, both for the THIN and the LTISI group (Figs. 5.7 and 5.8). The close alignment between the empirical and predicted means demonstrates that the inferred functions successfully account for the contribution of individual variables, accurately capturing how each predictor modulates the neural responses of the mice population.

5.5 Discussion

In this chapter, we presented a novel method to analyze photometry data within a Bayesian framework, employing a Hierarchical Hilbert Space Gaussian Process (HSGP) regression. This approach overcomes the limitations of traditional trial-averaging techniques by explicitly modeling the additive contributions of overlapping task events and the continuous nature of the neural response. By projecting the neural time-series onto a computationally efficient orthogonal basis, we were able to perform full probabilistic inference across large datasets, capturing both the shared population-level dynamics and subject-specific variability. We successfully applied this framework to characterize the functional properties of LTISI and THIN striatal interneurons in mice, demonstrating the model's capacity to disentangle the effect of different predictors and the effect of the same predictor at different times.

By decoupling neural responses from fixed trial structures, this method is particularly well-suited for analyzing naturalistic or continuous behaviors, where events do not occur in fixed or non-overlapping windows. The computational efficiency of the Hilbert Space ap-

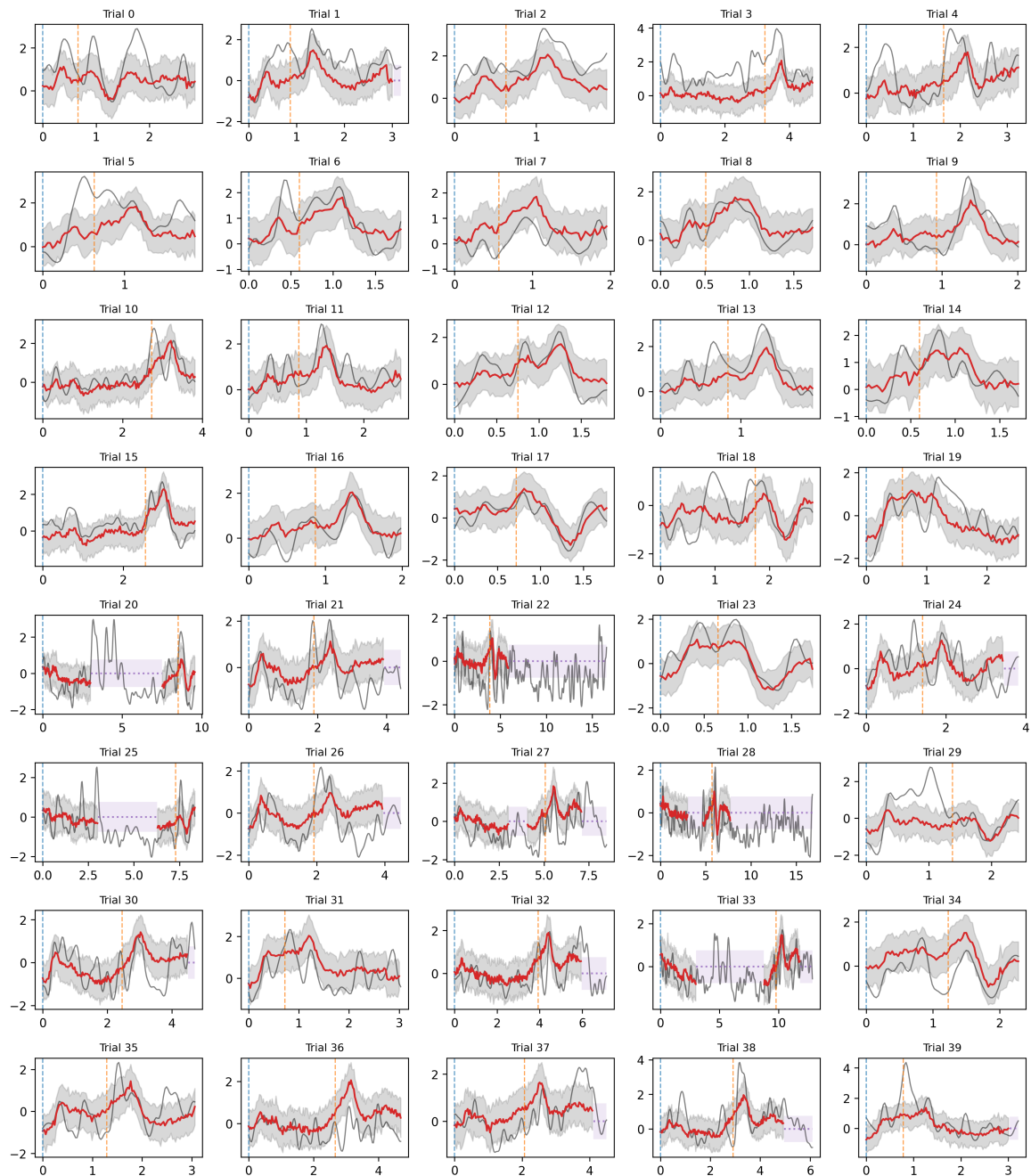


Fig. 5.6 Posterior predictive simulation of the model with continuous predictors on THIN data Each box represents one trial of subject 1 of the THIN population. Only the first 60 trials of the dataset are shown. On each box, the gray line represents the true neuronal activity, the red line represents the average of the posterior draws, and the gray shaded area represents one standard deviation from the average. For each trial, 1500×4 simulations were drawn, one for each sample of the four MCMC chains. A yellow line on each box shows the reward delivery time. In areas outside of both GP windows, where no simulation was possible, a dashed purple line represents the best guess for the data (i.e., a flat line at 0, since the data was standardized), and the light purple shaded area shows the inferred noise of the neural activity.

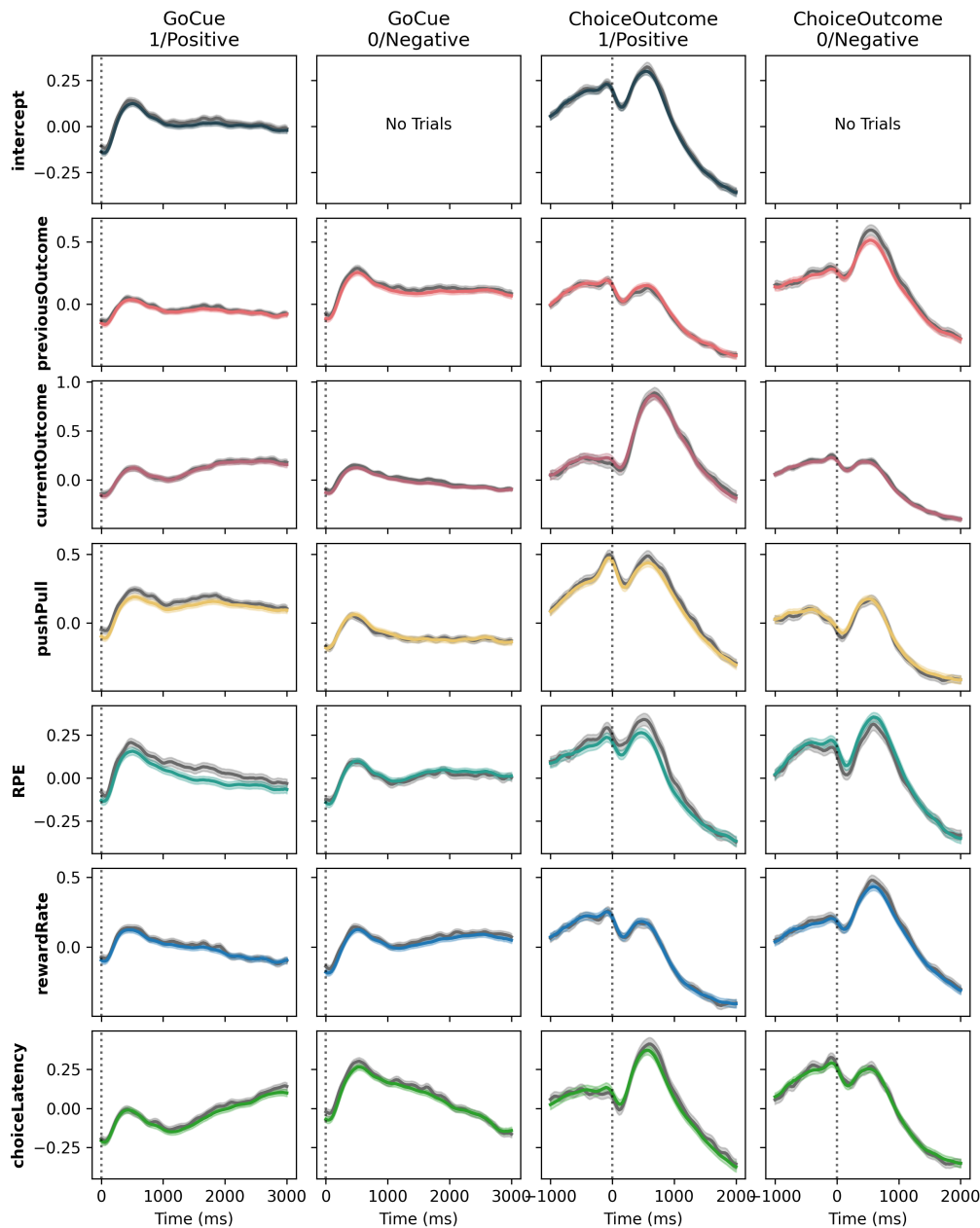


Fig. 5.7 Posterior Predictive Check (PPC) of the average LTISI population responses by predictor and condition. Each row represents one of the variables included in the model. The two leftmost columns show the trials aligned with the Go cue. The two rightmost columns show the trials aligned to the choice outcome. The times of alignment are marked by a vertical dashed line. The first of each pair of columns represents the trials where the variable was 1 (for binary predictors) or greater than 0 (for continuous predictors). The second represents the trials where the variable was 0 (for binary predictors) or less than 0 (for continuous predictors). On each box of the plot, the gray line represents the trials where the corresponding variable followed the corresponding condition, averaged. The gray shaded area represents the standard error of the mean (SEM). The colored lines represent the posterior expectation of the mean response. Shaded colored regions indicate the Posterior Standard Deviation of the Mean.

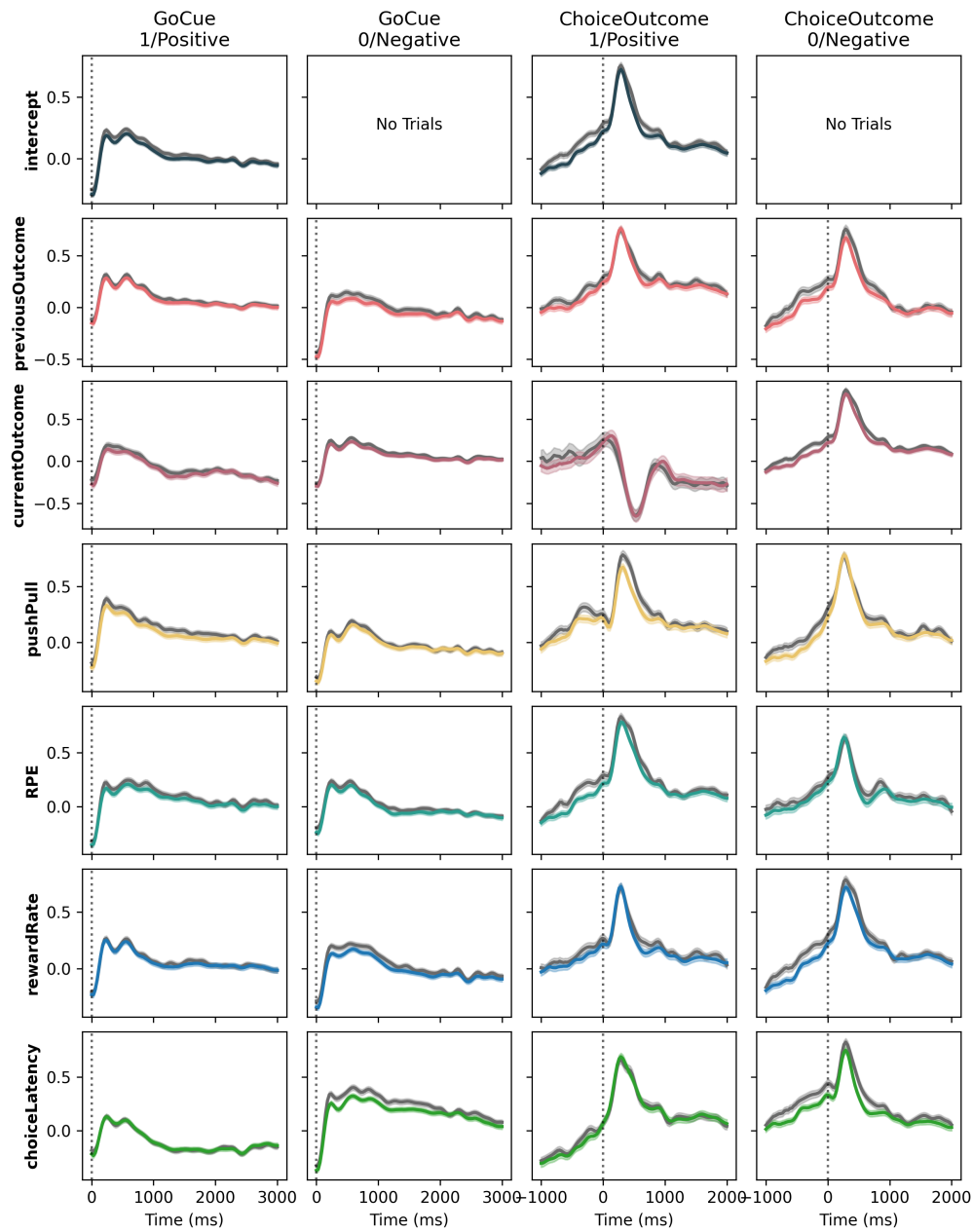


Fig. 5.8 Posterior Predictive Check (PPC) of the average THIN population responses by predictor and condition. Details as in Fig. 5.7.

proximation allows this regression to scale to reasonably big datasets without the prohibitive computational cost usually associated with Gaussian Processes. Therefore, this approach offers a generalizable tool for photometry data from experimental designs where traditional trial-averaging techniques would fail to resolve overlapping dynamics or to properly account for continuous signals.

Furthermore, the adoption of a fully Bayesian formalism offers distinct advantages over typical approaches based on General Linear Models (GLMs). Unlike frequentist approaches, our framework generates full posterior distributions for every model parameter, offering a principled and intuitive quantification of uncertainty at every time point. Moreover, posterior predictive checks let us evaluate the goodness of the fit, both on the level of the single trial and on a global level. Additionally, as we have seen, this formalism allows us to study possible collinearities between parameters and to compare different models, taking advantage of the posterior chains inferred. In particular, going forward with the study, LOO CV will allow us to recognize which of the predictors we are using, as well as which of those we haven't tested yet, are necessary to properly account for neural activation. Finally, the hierarchical structure of the model imposes a soft regularization: by sharing information across subjects, the model produces robust population-level inferences that are less susceptible to outliers, while simultaneously respecting and preserving the biological variability unique to each animal.

However, the architecture still presents a few areas of improvement, most notably, it still requires downsampling the data. The linear dependence of HSGP on sample size makes it so that the final running time would still be manageable, in the order of a few days. However, each predictor added increases the execution time, so the complete model would require a considerable duration. Nevertheless, there might be further optimizations of the code that can reduce the running time, so it would be worth it to look for more optimized implementations of the method. If indeed a further speed-up is possible, this could allow us to include more predictors, for instance, deriving directly from finer measurements of the animal's movement and behavior, such as the position of the joystick in time. Another possible improvement in the model architecture is the addition of one more hierarchical layer to account for the session-by-session differences within the same mouse. This could increase the precision of the model and give us a measure of how consistent each animal is. Finally, we plan on including a model-based "tuning index" in order to summarize the degree to which a population represents a certain variable (external or internal), and that can be compared across interneural populations, mice, or sessions.

In conclusion, this study demonstrates that fiber photometry data can be effectively analyzed within a Bayesian hierarchical framework. We showed that this approach offers

concrete advantages over traditional methods, particularly the ability to disentangle overlapping events and provide a clear measure of statistical uncertainty. Given the increasing prevalence of photometry in neuroscience, the availability of robust analysis pipelines is becoming more important. The method presented here offers a reliable, scalable alternative for extracting consistent physiological trends from complex experimental designs, providing a solid foundation for future investigations into neural dynamics.

Chapter 6

Conclusion

In this thesis, we explored the idea that the brain learns, selects, and applies probabilistic models to understand reality. To do so, we used a normative approach, using the tools and methods of Bayesian statistics.

A florid branch of cognitive neuroscience literature developed around the concept of internal models. Ever since the idea was popularized in the field ([Craig, 1943](#)), a vast amount of evidence has supported the hypothesis, from the discovery of cognitive maps ([Hafting et al., 2005](#); [O’Keefe and Dostrovsky, 1971](#); [Tolman, 1948](#)), to the successes of model-based reinforcement learning ([Dayan and Daw, 2008](#); [Glimcher, 2011](#)). In particular, probabilistic models, when cast in a Bayesian framework, can explain how the brain perceives stimuli ([Fischer and Peña, 2011](#); [Ma et al., 2023b](#); [Weiss et al., 2002](#)) and integrates them ([Alais and Burr, 2004](#); [Ernst and Banks, 2002](#); [Helbig and Ernst, 2008](#); [Kording et al., 2004](#)), how humans can understand each other’s actions and goals ([Alanqary et al., 2021](#); [Asakura and Inui, 2016](#); [Bhui and Gershman, 2020](#); [Diaconescu et al., 2014](#); [Ying et al., 2023](#)), how the mind can structure the understanding of complex concepts ([Bill et al., 2022](#); [Griffiths et al., 2024](#); [Kemp et al., 2007](#); [Yang et al., 2021](#)) or gain an intuition of physical causes and effects ([Sanborn et al., 2013](#)), and the way all this can be implemented in the neural substrate ([Abbott and Griffiths, 2011](#); [Bourgin et al., 2014](#); [Chater et al., 2020](#); [Clark, 2013](#); [Daw et al., 2005](#); [Findling et al., 2025](#); [Gershman et al., 2012](#); [Ullman et al., 2012](#)). The normative framework used in most of these studies, characterizing the optimal behavior that humans and animals seem to implement, is Bayesian Inference: a brain interpreting sensory data is described as if it inverting a generative internal model of the environment, combining prior expectations with current sensory input to estimate the state of the world. In Chapter 4, we used the same framework to explain the behavior of rats in a visual frequency discrimination task. In particular, the goal of our study was to understand if and how task-irrelevant sounds impacted stimulus perception: despite numerous, contrasting reports that they affect (enhancing or

suppressing) visually evoked neural responses (Deneux et al., 2019; Ibrahim et al., 2016; Iurilli et al., 2012; Oude Lohuis et al., 2024), there is a lack of evidence that any of these effects surfaces to the perceptual level.

In our visual-frequency-discrimination task, a sound-induced effect was already appreciable from the psychometric curves of the rats' choices (Fig. 4.3B); however, the origin and nature of this bias were unclear. Our hypothesis, consistent with that of Iurilli et al. (2012), was that the presence of sounds inhibits the activation of frequency-encoding neurons, changing the perception of the stimuli. We implemented this hypothesis in a Bayesian ideal observer model of the rat's choices. We assumed that the rats infer the class of a stimulus (either High or Low frequency relative to a reference value) by modeling their frequency-evoked neural activity as a Gaussian-distributed variable. Integrating the likelihood of this activity over all the possible frequencies within each class, the rat can find the class with the highest posterior probability. In our specific setup, this integration simplifies to a comparison against a single learned decision boundary: the rat reports 'High' only if its internal measurement (i.e., the evoked neural activity) exceeds a specific threshold. Given the noisy nature of the measurement, this mechanism naturally produces the sigmoidal shape of the psychometric function. If a sound inhibits neural activation in the visual cortex, it shifts the measurement relative to the learned boundary, distorting the sigmoid.

We used this ideal observer model to describe the rats' choices, and we inferred the model's parameters using a hierarchical Bayesian regression (Gelman et al., 2014b; Kruschke, 2015). This methodological approach, the same one used throughout all chapters of the thesis, offered several key advantages to the analysis. First of all, it allowed us to simultaneously characterize single-subject and population-level behavior, while at the same time regularizing the inference process. Moreover, by estimating the full posterior distribution, we obtained a principled and intuitive measure of parameter uncertainty and the necessary diagnostic tools to ensure reliability (Vehtari et al., 2021). Finally, it allowed us to evaluate and compare different models using Leave One Out Cross Validation (LOO CV), a robust selection framework that relies on minimal assumptions (Vehtari et al., 2017, 2024). The inferred posterior distribution precisely captured the behavioral data across the full spectrum of visual frequencies and noise stimuli (Fig. A.5). To further validate our hypothesis, we compared the ideal observer against alternative explanations, such as models where auditory stimuli affected the perceptual noise rather than compressing the perceived frequency space. The model selection process decisively favored the frequency compression mechanism. Furthermore, the analysis revealed that sound intensity drove this effect: models that constrained the compression parameter to be identical for stimuli of similar intensity outperformed those where no such constraint was imposed. Finally, the relation between sound intensity and

perceptual compression appeared to be linear: indeed, a model implementing such a linear relationship performed comparably to a fully unconstrained model, suggesting that the auditory cortex exerts a monotonic, inhibitory influence on visual perception.

To our knowledge, this result provides the first demonstration of a significant impact of task-irrelevant sounds on visual perception in rats. Using a classical ideal observer approach (Ma et al., 2023b) paired with a Bayesian data analysis approach, we were able to bridge this behavioral effect to existing reports of physiological effects (Jurilli et al., 2012). Assuming a neural implementation of the internal measurement and modeling of the visual stimuli, we formulated a prediction on the specific neuronal mechanisms underlying these sound-evoked behavioral phenomena.

However, there are limits to what can be inferred about neural mechanisms solely from behavioral data. While our model selection favored the direct auditory-to-visual inhibition hypothesis, we cannot exclude the possibility that distinct cognitive strategies or physiological processes, such as attentional shifts or indirect sound-evoked responses, might generate psychometric functions mathematically similar to our ideal observer. We did not undertake specific analytical or experimental tests to definitively rule out such mechanisms, and it remains possible that an alternative framework could offer an even more parsimonious explanation, for instance, by explicitly incorporating the relationship between sound intensity and frequency perception rather than inferring it from the data. Consequently, while the inhibition model provides a compelling functional explanation consistent with physiological literature, it remains a sufficient description of the behavior rather than a uniquely proven mechanism.

Resolving ambiguities such as these requires going beyond the analysis of behavioral output; to study how the brain implements internal models and what effect the specific implementation has on perception and behavior, it is necessary to demonstrate if and how the variables constituting these models are explicitly encoded in the activity of neural populations. This requirement motivates the work presented in Chapter 5. Here, we leverage the same rigorous statistical machinery established in the previous chapters, namely hierarchical Bayesian modeling and cross-validation, and apply it directly to neural data. Shifting our focus from the output of the internal model (behavior) to its substrate (neural activity), we developed a method to decode how and which variables are represented in the fluorescence traces of interneuron populations during a value-based decision-making task.

The data we studied were recorded using fiber photometry, a calcium imaging technique rapidly gaining prominence due to its neural specificity and versatility (Kielbinski and Bernacka, 2024; Simpson et al., 2024). We developed a novel method to analyze this kind of traces, employing a Hierarchical Hilbert Space Gaussian Process (HSGP, Solin and Särkkä,

2020) regression. Our approach addresses several key limitations of current methods for photometry data. Unlike standard methods that estimate single scalar parameters for every trace timepoint, Gaussian Processes (GP) allow us to specify priors over entire functions. This enables the inference of continuous response kernels for specific contextual or behavioral variables, capturing the temporal structure and dynamics of neural population activity.

A significant advantage of this functional regression, compared to more common point-by-point estimations (Markowitz et al., 2023; Willmore et al., 2022), is the ability to determine not only if, but also when a variable modulates the neural signal. While advanced frequentist smoothing methods can achieve similar results, they often require complex, unintuitive computations to establish confidence intervals (Loewinger et al., 2025); in contrast, our Bayesian framework provides a direct, principled measure of uncertainty via the posterior distribution. Moreover, our method creates a precise estimate of neural responses to multiple asynchronous events (e.g., Go cue and Reward delivery during a trial), without requiring to warp the data or to perform multiple inferences with different alignments. This results in a more precise estimation of the neural response, as we can decouple the effect elicited by different events.

The major drawback of using Gaussian Processes is their computational cost, as inference time grows cubically with the number of data points. To solve this bottleneck, we used the aforementioned Hilbert Space approximation. This Fourier-like decomposition of the inferred functions, combined with optimized parameterization, reduces the computational complexity from cubic to linear, making the method efficient and practically viable.

We demonstrated the practical utility of this framework by applying it to data from mice performing a value-based decision-making task (two-armed bandit). This paradigm is particularly relevant to the central theme of this thesis, as it requires the animal to construct and update an internal model of reward contingencies to maximize rewards (Huh et al., 2009; Qü et al., 2025). We analyzed activity from two distinct subpopulations of striatal interneurons, Tyrosine Hydroxylase-expressing Interneurons (THIN) or Low-Threshold Spiking Interneurons (LTSI), both involved in goal-directed behavior (Assous et al., 2017; Holly et al., 2019, 2021; Kaminer et al., 2019).

Our analysis, albeit preliminary, shows promising results. The kernels inferred had reasonable dynamics and displayed expected features. For example, the response to the trial outcome only resulted non-null after the reward delivery, suggesting that the method properly disentangles the activation generated by different variables and linked to different events. Moreover, our method successfully reproduced known physiological dynamics. Consistent with prior studies, the model captured a reward-induced excitation in LTSI interneurons (Holly et al., 2019, 2021) and a reduction in THIN activity linked to reward contexts (Stanfield,

2023). Additionally, the Bayesian framework enabled us to test the necessity of specific variables by comparing full and reduced models via LOO-CV. The approach also provided diagnostic tools to assess potential collinearity: by computing correlations between posterior samples of different parameters, we could directly quantify the degree of confounding.

While this methodology is currently in an active stage of development and refinement, our preliminary results demonstrate the theoretical soundness of the approach. We have identified specific limitations regarding the inference time of the method: these issues are the subject of ongoing work. Nevertheless, the current iteration successfully establishes the method's validity and its potential to extract information from population-level data.

Ideally, the maturation of this framework will allow for a shift from methodological validation to a more detailed physiological investigation. A compelling direction for future research would be to further develop the analysis of the data at hand to rigorously map the relationship between striatal activity and behavioral strategies. Such an analysis could unveil precisely how the variables of an internal model, such as reward prediction error or value, are dynamically represented in neural circuits. Furthermore, it could point us toward understanding which of the many possible strategies the animals are enacting, and even if more than one strategy is employed (Le et al., 2022). This last option is of particular interest in the context of this thesis, as a difference of strategies could relate to a multitude of interpretations of the task dynamics the animal is generating and comparing, and therefore to a process of model selection.

Indeed, while there is extensive interest in the study of internal models in the brain, less attention has been given to the way animals and humans intuitively decide which model to use. While some related psychological literature investigated the qualitative heuristics that guide human explanations (Einhorn and Hogarth, 1986; Read and Marcus-Newhall, 1993; Thagard, 1989), few studies tried to establish normative frameworks and strategies to quantify the selection process. Often, the choice of one interpretation over another is framed in Bayesian terms as a causal inference guided by priors (Körding et al., 2007; Peters et al., 2016; Rohe and Noppeney, 2015). However, the origin and structure of these priors are frequently left unexplored.

The robust evidence for a "simplicity bias" in perception and cognition (Feldman, 2016; Gershman and Niv, 2013; Lombrozo, 2007) hints, however, at specific optimal strategies: strategies that mirror the principles of statistical Model Selection (MS). Just as Bayesian inference proved successful in explaining perception, MS provides a parallel normative framework for understanding intuitive interpretation. Indeed, previous work has shown that human choices between different models align with statistical criteria (Blanchard et al., 2018; Genewein and Braun, 2014; Griffiths and Tenenbaum, 2005), providing a theoretical

justification for the cognitive penalization of complexity. Following the quantitative approach developed in [Piasini et al. \(2025\)](#) to measure this simplicity bias, Chapters 2 and 3 of this thesis investigate to what extent human model selection aligns with these theoretical principles.

The first study focused on the effect of sample size. A common characteristic of most MS criteria is that a larger amount of data weakens the relative contribution of complexity penalties compared to the goodness-of-fit of the model ([Akaike, 1998](#); [Balasubramanian, 1997](#); [Grünwald, 2007](#); [MacKay, 1992](#); [Schwarz, 1978](#)). By comparing how this balance shifts in humans, we were able to challenge the hypothesis that the brain implements a theoretical computation based on the guiding principles of MS frameworks.

We designed a decision-making task where participants had to choose one among five possible models to explain noisy data. To interpret participants' answers we used the Akaike Information Criterion (AIC, [Akaike, 1998](#)) and the Bayesian Information Criterion (BIC, [Schwarz, 1978](#)) as normative frameworks: both criteria, despite deriving from different optimality principles, characterize the best choice as the one with the highest log-likelihood, penalized by a quantity dependent on its number of parameters. We used the same measures of goodness-of-fit and complexity penalty to model human choices. This allowed us to quantitatively show that our participants display a simplicity bias, and that log-likelihood is a sensible measure for their goodness-of-fit, compared to easier-to-compute heuristics. Most importantly, we could prove that human choices depend on the sample size, but that this dependence is qualitatively different from the one induced by theoretical computations. Indeed, contrary to the linear scaling of the log-likelihood we observe in AIC and BIC, our participants' behavior demonstrates a sublinear dependence on the quantity of data presented. We further showed that the shape of this trend, inferred using Gaussian Processes, is not arbitrary, but closely follows the biases reported in the literature on numerosity perception: not only it grows logarithmically, as prescribed by Weber's law ([Fechner, 1860](#); [Merten and Nieder, 2009](#)), but after a certain threshold (i.e., for high enough quantity of data presented) it plateaus, mirroring the switch from numerosity-based to density-based estimation documented in [Anobile et al. \(2015\)](#) and [Burr et al. \(2018\)](#).

Overall, the picture presented by our results is that of a suboptimal selection process. The brain *does* exhibit the hallmarks of Model Selection: a preference for simplicity that relaxes as data accumulates. However, this behavior doesn't appear to organically manifest, but rather to be explicitly implemented in approximate heuristics. Indeed, it is vulnerable to perceptual biases that wouldn't directly affect a principled theoretical computation. However, as we showed in the chapter 3, this heuristic is based on a nuanced evaluation of model complexity.

In this second study on intuitive model selection, based on a fixed-sample-size variant of the task from Chapter 2, we investigated which model features shape the simplicity bias. Our analysis used again BIC as a starting framework, but modified it to infer the "internal penalty" used by the participants. By replacing the theoretical complexity term with a fitted penalty, we revealed another apparent divergence from optimality: unlike standard metrics, where the penalty grows linearly with the number of parameters, the fitted one only grew sublinearly. However, the high between-participant variability of this trend, together with the impossibility of relating this presumed bias to any perceptual effect, strongly suggested an alternative explanation. This intuition led us to find a new theoretically justified term that explained the sublinear trend with reduced variability. This term depended explicitly on the specific symmetries of the models used in our experiment, strongly suggesting that human model selection is sensitive to distinctive features of the problem at hand. In other words, it doesn't rely on a general heuristic that, for example, only depends on the number of model parameters, but it adapts to the class of models considered. A similar result was already hinted at in [Piasini et al. \(2025\)](#), where participants demonstrated a non-null sensitivity to terms that are only applicable to some models. Using a similar approach as the one in their study, based on the Laplace approximation of Bayesian Model Selection (BMS, [Balasubramanian, 1997](#); [MacKay, 1992](#)), we obtained yet another complexity term, this time penalizing a model the more its parameters are fine-tuned to the data. A very similar term induced a generally consistent sensitivity in the participants of [Piasini et al. \(2025\)](#). However, in our study, it once again resulted in a high variability between inferred sensitivities and showed, on average, a null effect. Somewhat surprisingly, the reason we found behind this incongruence is that in [Piasini et al. \(2025\)](#) the term properly reflected BMS, while in our case it produced an unreliable approximation. This would indicate that, once again, the reliance on this model feature changes based on the context, and more importantly, that it does so in a way that matches the optimal Bayesian integration.

Taken together, the results from Chapter 2 and 3 reveal a complex but fascinating picture of the mechanism used by the brain to choose between competing models. On one hand, the strategies employed by humans appear to be based on a heuristic computation, which possibly makes the task more resource-rational compared to a direct theoretically-principled one. This is corroborated by the fact that in both of these studies, as well as in [Piasini et al. \(2025\)](#), human sensitivity to complexity terms deviates quantitatively from the theoretical optimum. On the other hand, we observed that model-specific features, which are naturally accounted for in optimal statistical computations, have a measurable influence on behavior.

Similarly to what we proposed in Chapter 2, a possible explanation is that the brain uses a heuristic constructed from learned MS characteristics and model-specific features. For

instance, the visual system might estimate goodness-of-fit via the match between data density and model shape, weight it by the perceived sample size, and apply a series of penalties derived from the specific model family. Alternatively, the brain may indeed be implementing a formal approach, such as the integration of BMS, but through a radically approximated algorithm to minimize cognitive load. The literature on probabilistic cognition supports this view: many departures from rational norms in decision-making are well explained by the use of random sampling and stochastic search algorithms that rely on a very limited number of samples to represent probability distributions (Chater et al., 2020; Daw and Courville, 2008; Sanborn and Chater, 2016; Zhu et al., 2020).

While validating this stochastic hypothesis without introducing ad-hoc assumptions remains challenging, it offers a promising avenue for future research. Specifically, it could be insightful to investigate how sampling-based statistical metrics, such as LOO-CV or WAIC (Watanabe, 2013), behave when constrained to a minimal number of posterior draws. Such an analysis could reveal whether computational resource limits reproduce any of the deviations from optimality we observed in human subjects.

Beyond the possible link between the conflicting results of the two chapters, the robustness of our conclusions must be weighed against specific experimental and analytical limitations present in our work. Indeed, both studies present gaps in the validation of the proposed mechanisms. In Chapter 2, the link between the sublinear sample-size scaling and numerosity perception remains based on a qualitative correlation with the logarithmic trend of Fechner's law. A more direct evaluation of the participants' numerosity perception would have provided a stronger link between the observed trend and perceptual biases. While aggregating both the current trials and a sample-size-estimation experiment would have been challenging, given the extended length of the current task, a separate experiment on the same kind of visual data would provide more direct evidence or disprove our hypothesis. Moreover, our analyses only considered the effect of the total sample size; we did not model how the perceived numerosity of different clusters in the same trial might distort the decision process. Similarly, in Chapter 3, while the symmetry-based complexity term offers a sufficient explanation for the data, we have not proven it is a necessary one. We lacked a counterfactual experiment involving models devoid of such symmetries to definitively prove that this feature is the causal driver of the bias. Additionally, regarding the "Remainder" term, we attributed its high between-subject variability to the limitations of the Laplace approximation. However, we did not validate this; we could have, for instance, computed the exact Bayesian integral via numerical methods to isolate the true residual term, defined as the difference between the exact evidence and the remaining Laplace terms. Consequently, we cannot definitively

determine whether the discrepancy arises from the faulty mathematical approximation or a fundamental difference in human processing.

Our modeling approach also shows some limitations. We based our conclusions on an evaluation of the predictive power of multiple models of behavior. However, we lack an absolute estimate of how well the models perform. To evaluate how close they get to the maximum explainable variance in the behavioral data, for example, we could have employed a flexible, generic model, such as a Deep Neural Network. This would have allowed us to determine if our models capture the true structure of human strategy or simply outperform rigid alternatives. Furthermore, our analysis assumes that participants employ a single strategy to solve the task. While we controlled that their average performance didn't decline during the entire length of the task, this doesn't exclude the possibility that the same participant switched between different strategies during the experiment. For example, participants might employ different policies when facing visually different trials. The observed behaviors may therefore be statistical artifacts of averaging these distinct strategies.

To resolve these uncertainties, future research must prioritize validating the specific mechanisms we proposed while expanding the scope of the investigation. Concurrently, it would be equally insightful to test whether the features and biases of human model selection we studied in this thesis are consistent across different tasks. Our results confirm the presence of a simplicity bias, a now well-established guiding principle, in situations that require abstract reasoning: differently from previous research, in fact, simple considerations about the geometry of the data space are not sufficient to solve the tasks we designed (in other words, there are no simple "decision boundaries" in data space delimiting data that is likely to come from one model from data that is likely to come from another model). Moreover, we confirmed quantitatively that human behavior over-relies on the number of parameters. However, we still don't know if the results we drew from our experiments are specific to our setup or if they generalize to multiple scenarios. This is the case, for example, of the sample size scaling: we know from [Griffiths and Tenenbaum \(2005\)](#) that there are other contexts where sample size has an effect on human evaluations, but we have no quantitative measures of this effect. After confirming its link to numerosity, future research should test if the sublinear trend we observed in this work is a general one. Alternatively, a version of the experiment in Chapter 2 that only reports the quantity of data verbally (similarly to [Griffiths and Tenenbaum, 2005](#)) could tell us if indeed the trend originates in numerosity perception: in fact, the symbolic number mapping in adults with formal schooling is linear ([Dehaene et al., 2008](#)) and should therefore generate a different scaling. Similarly, in future work, we should check whether the sublinearity of the penalty term we found in Chapter 3 is indeed a consequence of the symmetries of the models, therefore supporting the idea

that specific theoretical features affect the selection process, or if the same trend is found in models without said symmetries.

Overall, these experiments would build toward answering the broader questions opened by this relatively new field. We could start to understand whether the brain uses a common strategy for diverse tasks, making model selection a dedicated cognitive function, or if every task results in distinct behaviors, suggesting that selection is merely the outcome of task-specific computations. Even if the latter proves to be the case, we could still seek to identify properties, conserved quantitatively or qualitatively, which would point toward overarching computational principles. Moreover, we could probe the persistence of these strategies to determine exactly how far specific biases generalize. For instance, we could test whether the brain maintains a consistent selection strategy across tasks with visually similar data but involving different classes of models, or conversely, with the same underlying models but visually distinct inputs. We could further expand this exploration by observing if these behaviors transfer to different sensory modalities, or how the selection process shifts in tasks where available time or attentional resources are limited.

Furthermore, we could consider tasks that don't present a fixed set of models. While the comparison with theoretical model selection requires a finite set of candidate models, in naturalistic scenarios the brain usually is not given such a set. It either generates multiple interpretations or draws them from memory. Future experiments should test how strongly this impacts the model selection process, and if the generation and selection process are intertwined (e.g. the brain might start by generating simple hypotheses, and if they don't reach a certain threshold of goodness-of-fit try again with more complex ones) or separate processes.

All these points raise physiological counterparts: how and where, in the brain, is model selection implemented, and at what level of description does it operate? It remains open whether a single, general-purpose mechanism supports model selection across domains, or whether different tasks recruit distinct brain networks. The behavioral patterns outlined above can help distinguish these possibilities, but stronger evidence will come from pairing these tasks with noninvasive neuroimaging (EEG, fMRI) during testing. In parallel, rodent-friendly versions of model-selection tasks, such as two-armed bandits, can support targeted tests of specific circuits and mechanisms. In experiments of this kind, analytical tools like the decoding framework developed in Chapter 5 could prove instrumental in linking behavioral strategies to their neural population-level substrates.

In conclusion, the diverse lines of inquiry presented in this thesis converge on a single fundamental concept: the brain's reliance on internal models to navigate an uncertain world. We explored this concept from different points of view, spanning all three of Marr's levels

(Marr, 1982). We discussed how it is often necessary to compare and choose between multiple such probabilistic models, and explored how humans achieve this objective. To do so, we considered the computational and theoretical features of model selection, such as its scaling with sample size, and linked them to human behavior, gaining insight into the possible algorithms that the brain employs. Furthermore, we explored how internal models can help make sense of perceptual biases, linking the inferential process of rats to possible neural implementations. Finally, we developed a tool to study which variables and components of these models are represented in different neuronal populations.

The multiple domains we touched on were studied through a common methodological lens. Not only did we employ a consistent analytical framework across all chapters, applying Bayesian regression even to experimental domains where it remains historically underutilized, and illustrating the advantages it offers over conventional frequentist approaches. We also extensively employed normative frameworks to characterize the optimal computational strategies underlying these behaviors. These reference models define optimal behavior under a set of assumptions: in our sensory perception study, the ideal observer model offered a rigorous baseline of optimal integration, enabling us to isolate the specific inhibitory mechanism by which auditory cues alter visual processing. Likewise, in our exploration of cognitive model selection, we employed statistical information criteria as the standard for rational inference. As opposed to the former case, however, this highlighted a reliance on heuristics and perceptual biases that guide human intuition away from theoretical optimality. These deviations show us that there are necessary biological constraints missing from the theoretical assumptions we used. To fully understand the approximations relied on by the brain, however, one must eventually use direct neural measurements. This necessity highlights the role of the methodological advances proposed in Chapter 5. While psychophysics can illuminate the computational goals and the algorithmic strategies of the brain, rigorously probing their physical implementation requires the ability to track internal variables within neural dynamics. The Gaussian Process framework developed contributes to building this bridge.

Altogether, this thesis argues that internal models are not only learned and applied, but also selected; it showcases how normative modeling and Bayesian analysis can expose the essential strategies used by humans and animals to resolve the inherent ambiguity of sensory input into a structured and meaningful interpretation of reality.

References

- Joshua T Abbott and Thomas L Griffiths. Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.
- Edward H Adelson, Alex P Pentland, et al. The perception of shading and reflectance. *Perception as Bayesian inference*, 1:409–424, 1996.
- Hirotougu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.
- Opeyemi O Alabi, M Felicia Davatolhagh, Mara Robinson, Michael P Fortunato, Luigim Vargas Cifuentes, Joseph W Kable, and Marc Vincent Fuccillo. Disruption of $nrxn1\alpha$ within excitatory forebrain circuits drives value-based dysfunction. *elife*, 9:e54838, 2020.
- David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.
- Arwa Alanqary, Gloria Z Lin, Joie Le, Tan Zhi-Xuan, Vikash K Mansinghka, and Joshua B Tenenbaum. Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*, 2021.
- Nitay Alon, Lion Schulz, Vaughan Bell, Michael Moutoussis, Peter Dayan, and Joseph M Barnby. (mal) adaptive mentalizing in the cognitive hierarchy, and its link to paranoia. *Computational Psychiatry*, 8(1):159, 2024.
- Giovanni Anobile, Guido Marco Cicchini, and David C Burr. Linear mapping of numbers onto space requires attention. *Cognition*, 122(3):454–459, 2012.
- Giovanni Anobile, Marco Turi, Guido Marco Cicchini, and David C Burr. Mechanisms for perception of numerosity or texture-density are governed by crowding-like effects. *Journal of Vision*, 15(5):4–4, 2015.
- Sylvain Arlot and Alain Celisse. A survey of cross validation procedures for model selection. *Statistics Surveys*, 4, 07 2009. doi:[10.1214/09-SS054](https://doi.org/10.1214/09-SS054).
- Nobuhiko Asakura and Toshio Inui. A bayesian framework for false belief reasoning in children: a rational integration of theory-theory and simulation theory. *Frontiers in psychology*, 7:2019, 2016.
- Maxime Assous, Jaime Kaminer, Fulva Shah, Arpan Garg, Tibor Koós, and James M Tepper. Differential processing of thalamic information via distinct striatal interneuron circuits. *Nature communications*, 8(1):15860, 2017.

- Riccardo Avvisati, Anna-Kristin Kaufmann, Callum J Young, Gabriella E Portlock, Sophie Cancemi, Rui Ponte Costa, Peter J Magill, and Paul D Dodson. Distributional coding of associative learning in discrete populations of midbrain dopamine neurons. *Cell reports*, 43(4), 2024.
- Maite Azcorra, Zachary Gaertner, Connor Davidson, Qianzi He, Hailey Kim, Shivathmihai Nagappan, Cooper K Hayes, Charu Ramakrishnan, Lief Fenno, Yoon Seok Kim, et al. Unique functional responses differentially map onto genetic subtypes of dopamine neurons. *Nature neuroscience*, 26(10):1762–1774, 2023.
- Vijay Balasubramanian. Statistical inference, occam’s razor, and statistical mechanics on the space of probability distributions. *Neural computation*, 9(2):349–368, 1997.
- Edoardo Balzani, Kaushik Lakshminarasimhan, Dora Angelaki, and Cristina Savin. Efficient estimation of neural tuning during naturalistic behavior. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12604–12614. Curran Associates, Inc., 2020.
- Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.
- Rahul Bhui and Samuel J Gershman. Paradoxical effects of persuasive messages. *Decision*, 7(4):239, 2020.
- Johannes Bill, Samuel J Gershman, and Jan Drugowitsch. Visual motion perception as online hierarchical inference. *Nature communications*, 13(1):7403, 2022.
- Célian Bimbard, Timothy PH Sit, Anna Lebedeva, Charu B Reddy, Kenneth D Harris, and Matteo Carandini. Behavioral origin of sound-evoked activity in mouse visual cortex. *Nature neuroscience*, 26(2):251–258, 2023.
- Michelangelo Bin, Jie Huang, Alberto Isidori, Lorenzo Marconi, Matteo Mischiati, and Eduardo Sontag. Internal models in control, bioengineering, and neuroscience. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):55–79, 2022.
- Jennifer K Bizley, Fernando R Nodal, Victoria M Bajo, Israel Nelken, and Andrew J King. Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral cortex*, 17(9):2172–2189, 2007.
- Jennifer K Bizley, Ross K Maddox, and Adrian KC Lee. Defining auditory-visual objects: behavioral tests and physiological mechanisms. *Trends in neurosciences*, 39(2):74–85, 2016.
- Sarah-Jayne Blakemore, Daniel Wolpert, and Chris Frith. Why can’t you tickle yourself? *Neuroreport*, 11(11):R11–R16, 2000.
- Thomas Blanchard, Tania Lombrozo, and Shaun Nichols. Bayesian occam’s razor is a razor of the people. *Cognitive science*, 42(4):1345–1359, 2018.
- Elizabeth Baraff Bonawitz and Tania Lombrozo. Occam’s rattle: children’s use of simplicity and probability to constrain inference. *Developmental psychology*, 48(4):1156, 2012.

- David Bourgin, Joshua Abbott, Tom Griffiths, Kevin Smith, and Ed Vul. Empirical evidence for markov chain monte carlo in memory search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.
- James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: composable transformations of python+ numpy programs. 2018.
- David C Burr, Giovanni Anobile, and Roberto Arrighi. Psychophysical evidence for the number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1740):20170045, 2018.
- Alberto Cabezas, Adrien Corenflos, Junpeng Lao, Rémi Louf, Antoine Carnec, Kaustubh Chaudhari, Reuben Cohn-Gordon, Jeremie Coullon, Wei Deng, Sam Duffield, et al. Blackjax: composable bayesian inference in jax. *arXiv preprint arXiv:2402.10797*, 2024.
- Riccardo Caramellino, Eugenio Piasini, Andrea Buccellato, Anna Carboncino, Vijay Balasubramanian, and Davide Zoccolan. Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes. *Elife*, 10:e72081, 2021.
- Vicente Vivanco Cepeda, Joshua B Tenenbaum, Kevin A Smith, and Vivian C Paulun. Ensemble physics: perceiving the mass of groups of objects is more than the sum of its parts. *Journal of Vision*, 25(9):2520–2520, 2025.
- Nick Chater, Jian-Qiao Zhu, Jake Spicer, Joakim Sundh, Pablo León-Villagr a, and Adam Sanborn. Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, 29(5):506–512, 2020.
- Daiwei Chen, Weikai Chang, and Pratik Chaudhari. Learning Capacity: A Measure of the Effective Dimensionality of a Model. URL <http://arxiv.org/abs/2305.17332>.
- Samuel J Cheyette and Steven T Piantadosi. A unified account of numerosity perception. *Nature human behaviour*, 4(12):1265–1272, 2020.
- Tae-Yong Choi, Hyungseok Jeon, Sejin Jeong, Eum Ji Kim, Jeongseop Kim, Yun Ha Jeong, Byungsoo Kang, Murim Choi, and Ja Wook Koo. Distinct prefrontal projection activity and transcriptional state conversely orchestrate social competition and hierarchy. *Neuron*, 112(4):611–627, 2024.
- Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.
- Simon Clavagnier, Arnaud Falchier, and Henry Kennedy. Long-distance feedback projections to area v1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive, Affective, & Behavioral Neuroscience*, 4(2):117–126, 2004.
- Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.
- Kenneth James Williams Craik. *The nature of explanation*. Cambridge University Press, 1943.

- Nathaniel Daw and Aaron Courville. The pigeon as particle filter. *Advances in neural information processing systems*, 20:369–376, 2008.
- Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.
- Peter Dayan and Nathaniel D Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, 2008.
- Clélia de Mulatier and Matteo Marsili. Bayesian inference of minimally complex models with interactions of arbitrary order. *Physical Review E*, 111(5):054307, 2025.
- Russell L. De Valois and Karen K. De Valois. *Spatial Vision*. Number 14 in Oxford Psychology Series. Oxford University Press. ISBN 978-0-19-505019-6.
- Stanislas Dehaene, Véronique Izard, Elizabeth Spelke, and Pierre Pica. Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures. *science*, 320(5880):1217–1220, 2008.
- Lisa R Dehner, Leslie P Keniston, H Ruth Clemo, and M Alex Meredith. Cross-modal circuitry between auditory and somatosensory areas of the cat anterior ectosylvian sulcal cortex: a ‘new’ inhibitory form of multisensory convergence. *Cerebral Cortex*, 14(4):387–403, 2004.
- Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.
- Thomas Deneux, Evan R Harrell, Alexandre Kempf, Sebastian Ceballo, Anton Filipchuk, and Brice Bathellier. Context-dependent signaling of coincident auditory and visual events in primary visual cortex. *Elife*, 8:e44006, 2019.
- Andreea O Diaconescu, Christoph Mathys, Lilian AE Weber, Jean Daunizeau, Lars Kasper, Ekaterina I Lomakina, Ernst Fehr, and Klaas E Stephan. Inferring on the intentions of others by hierarchical bayesian learning. *PLoS computational biology*, 10(9):e1003810, 2014.
- Anthony Dickinson and Bernard Balleine. The role of learning in the operation of motivational systems. *Stevens’ handbook of experimental psychology*, 3:497–533, 2002.
- Mark A Eckert, Nirav V Kamdar, Catherine E Chang, Christian F Beckmann, Michael D Greicius, and Vinod Menon. A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fmri connectivity analysis. *Human brain mapping*, 29(7):848–857, 2008.
- Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483, 1978.
- Hillel J Einhorn and Robin M Hogarth. Judging probable cause. *Psychological bulletin*, 99(1):3, 1986.

- Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.
- Marc O Ernst and Heinrich H Bülthoff. Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169, 2004.
- Arnaud Falchier, Simon Clavagnier, Pascal Barone, and Henry Kennedy. Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, 22(13):5749–5759, 2002.
- Gustav Theodor Fechner. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel, 1860.
- Jacob Feldman. The simplicity principle in human concept learning. *Current directions in psychological science*, 12(6):227–232, 2003.
- Jacob Feldman. The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(5):330–340, 2016.
- Christopher R Fetsch, Amanda H Turner, Gregory C DeAngelis, and Dora E Angelaki. Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience*, 29(49):15601–15612, 2009.
- Christopher R Fetsch, Gregory C DeAngelis, and Dora E Angelaki. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nature Reviews Neuroscience*, 14(6):429–442, 2013.
- Charles Findling, Felix Hubert, International Brain Laboratory, Luigi Acerbi, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, E Kelly Buchanan, Sebastian Bruijns, et al. Brain-wide representations of prior information in mouse decision-making. *Nature*, 645(8079):192–200, 2025.
- Brian J Fischer and José Luis Peña. Owl’s behavior and neural representation predicted by bayesian inference. *Nature neuroscience*, 14(8):1061–1066, 2011.
- Bruce A Francis and Walter Murray Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976.
- Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.
- Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3):70–87, 2006.
- Aleena R Garner and Georg B Keller. A cortical circuit for audio-visual predictions. *Nature neuroscience*, 25(1):98–105, 2022.
- Wilson S Geisler. Ideal observer theory in psychophysics and physiology. *Physica Scripta*, 39(1):153–160, January 1989. ISSN 0031-8949, 1402-4896. doi:[10.1088/0031-8949/39/1/025](https://doi.org/10.1088/0031-8949/39/1/025).
- Wilson S. Geisler. Contributions of ideal observer theory to vision research. *Vision Research*, 51(7):771–781, April 2011. ISSN 0042-6989. doi:[10.1016/j.visres.2010.09.027](https://doi.org/10.1016/j.visres.2010.09.027).

- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2014a. ISBN 978-1-4398-4095-5.
- Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014b.
- Tim Genewein and Daniel A Braun. Occam’s razor in sensorimotor learning. *Proceedings of the Royal Society B: Biological Sciences*, 281(1783):20132952, 2014.
- Samuel J Gershman and Yael Niv. Perceptual estimation obeys occam’s razor. *Frontiers in psychology*, 4:623, 2013.
- Samuel J Gershman, Edward Vul, and Joshua B Tenenbaum. Multistability and perceptual inference. *Neural computation*, 24(1):1–24, 2012.
- Stephanie Gleiss and Christoph Kayser. Audio-visual detection benefits in the rat. *PLOS ONE*, 7(9):1–8, 09 2012. doi:[10.1371/journal.pone.0045677](https://doi.org/10.1371/journal.pone.0045677). URL <https://doi.org/10.1371/journal.pone.0045677>.
- Paul W Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108 (supplement_3):15647–15654, 2011.
- Joshua I. Gold and Michael N. Shadlen. The Neural Basis of Decision Making. 30(1). doi:[10.1146/annurev.neuro.29.051605.113038](https://doi.org/10.1146/annurev.neuro.29.051605.113038). URL <http://dx.doi.org/10.1146/annurev.neuro.29.051605.113038>.
- David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley. ISBN 978-0-932146-23-6.
- Thomas L Griffiths and Joshua B Tenenbaum. Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384, 2005.
- Thomas L Griffiths, Nick Chater, and Joshua B Tenenbaum. *Bayesian models of cognition: Reverse engineering the mind*. MIT Press, 2024.
- Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.
- Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.
- Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*, 2009.
- Gary Hatfield and William Epstein. The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin*, 97(2):155, 1985.
- Hannah B Helbig and Marc O Ernst. Visual-haptic cue weighting is independent of modality-specific attention. *Journal of vision*, 8(1):21–21, 2008.
- Nicholas J Higham. Cholesky factorization. *Wiley interdisciplinary reviews: computational statistics*, 1(2):251–254, 2009.

- Rebecca J Hirst, David P McGovern, Annalisa Setti, Ladan Shams, and Fiona N Newell. What you see is what you hear: Twenty years of research using the sound-induced flash illusion. *Neuroscience & Biobehavioral Reviews*, 118:759–774, 2020.
- Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15 (47):1593–1623, 2014. URL <http://jmlr.org/papers/v15/hoffman14a.html>.
- Elizabeth N Holly, M Felicia Davatolhagh, Kyuhyun Choi, Opeyemi O Alabi, Luigim Vargas Cifuentes, and Marc V Fuccillo. Striatal low-threshold spiking interneurons regulate goal-directed learning. *Neuron*, 103(1):92–101, 2019.
- Elizabeth N Holly, M Felicia Davatolhagh, Rodrigo A España, and Marc V Fuccillo. Striatal low-threshold spiking interneurons locally gate dopamine. *Current Biology*, 31(18): 4139–4147, 2021.
- Namjung Huh, Suhyun Jo, Hoseok Kim, Jung Hoon Sul, and Min Whan Jung. Model-based reinforcement learning under concurrent schedules of reinforcement in rodents. *Learning & Memory*, 16(5):315–323, 2009.
- Leena A Ibrahim, Lukas Mesik, Xu-ying Ji, Qi Fang, Hai-fu Li, Ya-tang Li, Brian Zingg, Li I Zhang, and Huizhong Whit Tao. Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition. *Neuron*, 89(5):1031–1045, 2016.
- Giuliano Iurilli, Diego Ghezzi, Umberto Olcese, Glenda Lassi, Cristiano Nazzaro, Raffaella Tonini, Valter Tucci, Fabio Benfenati, and Paolo Medini. Sound-driven synaptic inhibition in primary visual cortex. *Neuron*, 73(4):814–828, 2012.
- Harold Jeffreys. Some tests of significance, treated by the theory of probability. In *Mathematical proceedings of the Cambridge philosophical society*, volume 31, pages 203–222. Cambridge University Press, 1935.
- Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186 (1007):453–461, 1946.
- Samuel Johnson, Andy Jin, and Frank Keil. Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.
- Jaime Kaminer, Diego Espinoza, Shaznaan Bhimani, James M Tepper, Tibor Koos, and Michael W Shiflett. Loss of striatal tyrosine-hydroxylase interneurons impairs instrumental goal-directed behavior. *European Journal of Neuroscience*, 50(4):2653–2662, 2019.
- Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks. 33(8):2274–2307. ISSN 0899-7667. doi:[10.1162/neco_a_01411](https://doi.org/10.1162/neco_a_01411). URL https://doi.org/10.1162/neco_a_01411.
- Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. The discrimination of visual number. *The American journal of psychology*, 62(4):498–525, 1949.

- Christoph Kayser and Nikos K Logothetis. Do early sensory cortices integrate cross-modal information? *Brain structure and function*, 212:121–132, 2007.
- Christoph Kayser, Christopher I Petkov, Mark Augath, and Nikos K Logothetis. Integration of touch and sound in auditory cortex. *Neuron*, 48(2):373–384, 2005.
- Christoph Kayser, Christopher I Petkov, and Nikos K Logothetis. Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, 18(7):1560–1574, 2008.
- Christoph Kayser, Nikos K Logothetis, and Stefano Panzeri. Visual enhancement of the information representation in auditory cortex. *Current Biology*, 20(1):19–24, 2010.
- Julian Keil. Double flash illusions: current findings and future directions. *Frontiers in Neuroscience*, 14:298, 2020.
- Georg B Keller, Tobias Bonhoeffer, and Mark Hübener. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74(5):809–815, 2012.
- Charles Kemp, Andrew Perfors, and Joshua B Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental science*, 10(3):307–321, 2007.
- Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object Perception as Bayesian Inference. *Annual Review of Psychology*, 55(Volume 55, 2004):271–304, January 2004. ISSN 0066-4308, 1545-2085. doi:[10.1146/annurev.psych.55.090902.142005](https://doi.org/10.1146/annurev.psych.55.090902.142005).
- Michał Kielbinski and Joanna Bernacka. Fiber photometry in neuroscience research: principles, applications, and future directions. *Pharmacological Reports*, 76(6):1242–1255, 2024.
- David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.
- David C. Knill and Whitman Richards, editors. *Perception as Bayesian Inference*. Cambridge University Press, 1 edition. ISBN 978-0-521-46109-2 978-0-521-06499-6 978-0-511-98403-7. doi:[10.1017/CBO9780511984037](https://doi.org/10.1017/CBO9780511984037). URL <https://www.cambridge.org/core/product/identifier/9780511984037/type/book>.
- Eric B Knudsen and Joni D Wallis. Taking stock of value in the orbitofrontal cortex. *Nature Reviews Neuroscience*, 23(7):428–438, 2022.
- Adám Koblínger, József Fiser, and Máté Lengyel. Representations of uncertainty: Where art thou? 38:150–162. ISSN 2352-1546. doi:[10.1016/j.cobeha.2021.03.009](https://doi.org/10.1016/j.cobeha.2021.03.009). URL <https://www.sciencedirect.com/science/article/pii/S2352154621000577>.
- Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013.
- Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.
- Konrad P Kording, Shih-pi Ku, and Daniel M Wolpert. Bayesian integration in force estimation. *Journal of neurophysiology*, 92(5):3161–3165, 2004.

- Konrad P Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS one*, 2(9):e943, 2007.
- Lester E Krueger. Perceived numerosity. *Perception & Psychophysics*, 11:5–9, 1972.
- John K. Kruschke. *Doing Bayesian Data Analysis*. Academic Press, 2 edition, 2015. ISBN 978-0-12-405888-0.
- Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33):1143, 2019. doi:[10.21105/joss.01143](https://doi.org/10.21105/joss.01143).
- Nhat Minh Le, Murat Yildirim, Yizhi Wang, Hiroki Sugihara, Mehrdad Jazayeri, and Mriganka Sur. Mixture of learning strategies underlies rodent behavior in dynamic foraging. *bioRxiv*, pages 2022–03, 2022.
- Michael D Lee. A hierarchical bayesian model of human decision-making on an optimal stopping problem. *Cognitive science*, 30(3):1–26, 2006.
- Ivan Linares-García, Evan A Iliakis, Sofia E Juliani, Alexandra N Ramirez, Joel Woolley, Edgar Díaz-Hernández, Marc V Fuccillo, and David J Margolis. An open-source joystick platform for investigating forelimb motor control, auditory-motor integration, and value-based decision-making in head-fixed mice. *eneuro*, 12(4), 2025.
- Alma Lindborg and Tobias S Andersen. Bayesian binding and fusion models explain illusion and enhancement effects in audiovisual speech perception. *PLoS One*, 16(2):e0246986, 2021.
- Kevin Lloyd, Adam Sanborn, David Leslie, and Stephan Lewandowsky. Why higher working memory capacity may help you learn: Sampling, search, and degrees of approximation. *Cognitive Science*, 43(12):e12805, 2019.
- Gabriel Loewinger, Erjia Cui, David Lovinger, and Francisco Pereira. A statistical framework for analysis of trial-level temporal dynamics in fiber photometry experiments. *Elife*, 13:RP95802, 2025.
- Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257, 2007.
- Liqun Luo, Edward M Callaway, and Karel Svoboda. Genetic dissection of neural circuits. *Neuron*, 57(5):634–660, 2008.
- Wei Ji Ma and Mehrdad Jazayeri. Neural Coding of Uncertainty and Probability. 37(1):205–220. ISSN 0147-006X, 1545-4126. doi:[10.1146/annurev-neuro-071013-014017](https://doi.org/10.1146/annurev-neuro-071013-014017). URL <https://www.annualreviews.org/content/journals/10.1146/annurev-neuro-071013-014017>.
- Wei Ji Ma, Masud Husain, and Paul M Bays. Changing concepts of working memory. *Nature neuroscience*, 17(3):347–356, 2014.
- Wei Ji Ma, Konrad Kording, and Daniel Goldreich. *Bayesian Models of Perception and Action: An Introduction*. MIT Press, 2023a. ISBN 978-0-262-04759-3.

- Wei Ji Ma, Konrad Paul Kording, and Daniel Goldreich. *Bayesian models of perception and action: An introduction*. MIT press, 2023b.
- David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.
- David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.
- J MacQueen. Multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.
- John F Magnotti, Wei Ji Ma, and Michael S Beauchamp. Causal inference of asynchronous audiovisual speech. *Frontiers in psychology*, 4:798, 2013.
- John F Magnotti, Kristen B Dzeda, Kira Wegner-Clemens, Johannes Rennig, and Michael S Beauchamp. Weak observer-level correlation and strong stimulus-level correlation between the mcgurk effect and audiovisual speech-in-noise: A causal inference explanation. *Cortex*, 133:371–383, 2020.
- Tiago V Maia and Michael J Frank. From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, 14(2):154–162, 2011.
- Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02767>.
- Jeffrey E Markowitz, Winthrop F Gillis, Maya Jay, Jeffrey Wood, Ryley W Harris, Robert Cieszkowski, Rebecca Scott, David Brann, Dorothy Koveal, Tomasz Kula, et al. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, 614(7946):108–117, 2023.
- David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 1982.
- Roberto Martuzzi, Micah M Murray, Christoph M Michel, Jean-Philippe Thiran, Philippe P Maeder, Stephanie Clarke, and Reto A Meuli. Multisensory interactions within human primary cortices revealed by bold dynamics. *Cerebral Cortex*, 17(7):1672–1679, 2007.
- Giulio Matteucci and Davide Zoccolan. Unsupervised experience with temporal continuity of the visual environment is causally involved in the development of v1 complex cells. *Science Advances*, 6(22):eaba3742, 2020.
- Giulio Matteucci, Rosilari Bellacosa Marotti, Margherita Riggi, Federica B Rosselli, and Davide Zoccolan. Nonlinear processing of shape information in rat lateral extrastriate cortex. *Journal of Neuroscience*, 39(9):1649–1670, 2019.
- Giulio Matteucci, Margherita Riggi, Davide Zoccolan, et al. A template-matching algorithm for laminar identification of cortical recording sites from evoked response potentials. *J Neurophysiol*, 124:102–114, 2020.

- Giulio Matteucci, Benedetta Zattera, Rosilari Bellacosa Marotti, and Davide Zoccolan. Rats spontaneously perceive global motion direction of drifting plaids. *PLoS Computational Biology*, 17(9):e1009415, 2021.
- Giulio Matteucci, Rosilari Bellacosa Marotti, Benedetta Zattera, and Davide Zoccolan. Truly pattern: Nonlinear integration of motion signals is required to account for the responses of pattern cells in rat visual cortex. *Science Advances*, 9(45):eadh4690, 2023.
- John P McClure Jr and Pierre-Olivier Polack. Pure tones modulate the representation of orientation and direction in the primary visual cortex. *Journal of neurophysiology*, 121(6):2202–2214, 2019.
- Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, second edition edition, 2020. ISBN 978-0-367-13991-9.
- Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.
- Martha T Mednick and Sharon Halpern. Remote associates test. *Psychological Review*, 1968.
- Guido T Meijer, Jorrit S Montijn, Cyriel MA Pennartz, and Carien S Lansink. Audiovisual modulation in mouse primary visual cortex depends on cross-modal stimulus configuration and congruency. *Journal of Neuroscience*, 37(36):8783–8796, 2017.
- M Alex Meredith, Leslie R Keniston, Lisa R Dehner, and H Ruth Clemo. Crossmodal projections from somatosensory area siv to the auditory field of the anterior ectosylvian sulcus (faes) in cat: further evidence for subthreshold forms of multisensory processing. *Experimental Brain Research*, 172:472–484, 2006.
- Katharina Merten and Andreas Nieder. Compressed scaling of abstract numerosity representations in adult humans and monkeys. *Journal of Cognitive Neuroscience*, 21(2):333–346, 2009.
- Michael L Morgan, Gregory C DeAngelis, and Dora E Angelaki. Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron*, 59(4):662–673, 2008.
- In Jae Myung. The importance of complexity in model selection. *Journal of mathematical psychology*, 44(1):190–204, 2000.
- Bence Nanay. The importance of amodal completion in everyday perception. *i-Perception*, 9(4):2041669518788887, 2018.
- Nader Nikbakht, Azadeh Tafreshiha, Davide Zoccolan, and Mathew E Diamond. Supralinear and supramodal integration of visual and tactile signals in rats: psychophysics and neuronal mechanisms. *Neuron*, 97(3):626–639, 2018.
- Uta Noppeney. Perceptual inference, learning, and attention in a multisensory world. *Annual review of neuroscience*, 44(1):449–473, 2021.

- John O’Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.
- Matthijs N Oude Lohuis, Pietro Marchesi, Umberto Olcese, and Cyriel MA Pennartz. Triple dissociation of visual, auditory and motor processing in mouse primary visual cortex. *Nature Neuroscience*, 27(4):758–771, 2024.
- Michael Pacer and Tania Lombrozo. Ockham’s razor cuts to the root: simplicity in causal explanation. *Journal of Experimental Psychology: General*, 146(12):1761, 2017.
- Stefano Panzeri, Christopher D Harvey, Eugenio Piasini, Peter E Latham, and Tommaso Fellin. Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507, 2017.
- Il Memming Park, Miriam L R Meister, Alexander C Huk, and Jonathan W Pillow. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience*, 17(10):1395–1403, October 2014. ISSN 1097-6256, 1546-1726. doi:[10.1038/nn.3800](https://doi.org/10.1038/nn.3800).
- Nathan F. Parker, Avinash Baidya, Julia Cox, Laura M. Haetzel, Anna Zhukovskaya, Malavika Murugan, Ben Engelhard, Mark S. Goldman, and Ilana B. Witten. Choice-selective sequences dominate in cortical relative to thalamic inputs to NAc to support reinforcement learning. *Cell Reports*, 39(7), May 2022. ISSN 2211-1247. doi:[10.1016/j.celrep.2022.110756](https://doi.org/10.1016/j.celrep.2022.110756).
- Tommaso Patriarchi, Jounhong Ryan Cho, Katharina Merten, Mark W Howe, Aaron Marley, Wei-Hong Xiong, Robert W Folk, Gerard Joey Broussard, Ruqiang Liang, Min Jee Jang, et al. Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science*, 360(6396):eaat4422, 2018.
- Megan AK Peters, Wei Ji Ma, and Ladan Shams. The size-weight illusion is not anti-bayesian after all: a unifying bayesian account. *PeerJ*, 4:e2124, 2016.
- Eugenio Piasini, Shuze Liu, Pratik Chaudhari, Vijay Balasubramanian, and Joshua I Gold. How occam’s razor guides human decision-making. 2025. doi:[10.7554/elife.109113.1](https://doi.org/10.7554/elife.109113.1). URL <http://dx.doi.org/10.7554/eLife.109113.1>.
- Jonathan W. Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M. Litke, E. J. Chichilnisky, and Eero P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, August 2008. doi:[10.1038/nature07140](https://doi.org/10.1038/nature07140).
- Sashank Pisupati, Lital Chartarifsky-Lynn, Anup Khanal, and Anne K Churchland. Lapses in perceptual decisions reflect exploration. *Elife*, 10:e55490, 2021.
- Albert J Qü, Lung-Hao Tai, Christopher D Hall, Emilie M Tu, Maria K Eckstein, Karyna Mishchanchuk, Wan Chen Lin, Juliana B Chase, Andrew F MacAskill, Anne GE Collins, et al. Nucleus accumbens dopamine release reflects bayesian inference during instrumental learning. *PLOS Computational Biology*, 21(7):e1013226, 2025.
- David Raposo, John P Sheppard, Paul R Schrater, and Anne K Churchland. Multisensory decision-making in rats and humans. *Journal of neuroscience*, 32(11):3726–3735, 2012.

- David Raposo, Matthew T Kaufman, and Anne K Churchland. A category-free neural population supports evolving demands during decision-making. *Nature neuroscience*, 17(12):1784–1792, 2014.
- Stephen J Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3):429, 1993.
- Francesco G. Rinaldi. OCCABS. OSF Preregistration, May 2025. URL osf.io/cjt92. Retrieved from osf.io/cjt92.
- Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
- Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R Andersen, Arno Solin, and Aki Vehtari. Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17, 2023.
- Kathleen S Rockland and Hisayuki Ojima. Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, 50(1-2):19–26, 2003.
- Tim Rohe and Uta Noppeney. Cortical hierarchies perform bayesian causal inference in multisensory perception. *PLoS biology*, 13(2):e1002073, 2015.
- Caroline A. Runyan, Eugenio Piasini, Stefano Panzeri, and Christopher D. Harvey. Distinct timescales of population coding across cortex. *Nature*, 548(7665):92–96, August 2017. ISSN 0028-0836, 1476-4687. doi:[10.1038/nature23020](https://doi.org/10.1038/nature23020).
- John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016. doi:[10.7717/peerj-cs.55](https://doi.org/10.7717/peerj-cs.55).
- Majed Samad, Albert Jin Chung, and Ladan Shams. Perception of body ownership is driven by bayesian sensory inference. *PloS one*, 10(2):e0117178, 2015.
- Adam N Sanborn and Nick Chater. Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893, 2016.
- Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, 120(2):411, 2013.
- Charles E Schroeder and John Foxe. Multisensory contributions to low-level, ‘unisensory’ processing. *Current opinion in neurobiology*, 15(4):454–458, 2005.
- Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.
- Robert L Seilheimer, Ari Rosenberg, and Dora E Angelaki. Models and processes of multisensory cue combination. *Current opinion in neurobiology*, 25:38–46, 2014.
- Ladan Shams and Ulrik R Beierholm. Causal inference in perception. *Trends in cognitive sciences*, 14(9):425–432, 2010.

- Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. What you see is what you hear. *Nature*, 408(6814):788–788, 2000.
- Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.
- Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):749–760, 1995.
- Eleanor H Simpson, Thomas Akam, Tommaso Patriarchi, Marta Blanco-Pozo, Lauren M Burgeno, Ali Mohebi, Stephanie J Cragg, and Mark E Walton. Lights, fiber, action! a primer on in vivo fiber photometry. *Neuron*, 112(5):718–739, 2024.
- Kevin A Smith, David E Huber, and Edward Vul. Multiply-constrained semantic search in the remote associates test. *Cognition*, 128(1):64–75, 2013.
- Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 30(2):419–446, 2020.
- Roger Wolcott Sperry. Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of comparative and physiological psychology*, 43(6):482, 1950.
- David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.
- Briana R Stanfield. *The Role of Tyrosine Hydroxylase-Expressing Interneurons in Striatal Activity and Operant Behavior*. PhD thesis, Rutgers The State University of New Jersey, Graduate School-Newark, 2023.
- Jimmy Stehberg, Phat T Dang, and Ron D Frostig. Unimodal primary sensory cortices are directly connected by long-range horizontal projections in the rat sensory cortex. *Frontiers in neuroanatomy*, 8:93, 2014.
- Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.
- Ke Sun and Frank Nielsen. A geometric modeling of Occam’s razor in deep learning. 8(1): 233–273. ISSN 2511-249X. doi:[10.1007/s41884-025-00167-2](https://doi.org/10.1007/s41884-025-00167-2). URL <https://doi.org/10.1007/s41884-025-00167-2>.
- Paul Thagard. Explanatory coherence. *Behavioral and brain sciences*, 12(3):435–467, 1989.
- Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.
- Tomer D Ullman, Noah D Goodman, and Joshua B Tenenbaum. Theory learning as stochastic search in the language of thought. *Cognitive Development*, 27(4):455–480, 2012.
- RJ Van Lier, PA Van der Helm, and ELJ Leeuwenberg. Competing global and local completions in visual occlusion. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):571, 1995.

- Walter Vanzella, Natalia Grion, Daniele Bertolini, Andrea Perissinotto, Marco Gigante, and Davide Zoccolan. A passive, camera-based head-tracking system for real-time, three-dimensional estimation of head position and orientation in rodents. *Journal of neurophysiology*, 122(6):2220–2242, 2019.
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. ISSN 1573-1375. doi:[10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4). URL <https://doi.org/10.1007/s11222-016-9696-4>.
- Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-Normalization, Folding, and Localization: An Improved $R^{\hat{}}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718, June 2021. ISSN 1936-0975, 1931-6690. doi:[10.1214/20-BA1221](https://doi.org/10.1214/20-BA1221).
- Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.
- Simone Viganò, Valerio Rubino, Marco Buiatti, and Manuela Piazza. The neural representation of absolute direction during mental navigation in conceptual spaces. *Communications biology*, 4(1):1294, 2021.
- John Von Neumann. Method in the physical sciences. *Collected Works*, 6:491–498, 1955.
- Sumio Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013.
- Shams Watkins, Ladan Shams, Sachiyo Tanaka, J-D Haynes, and Geraint Rees. Sound alters activity in human v1 in association with illusory visual perception. *Neuroimage*, 31(3):1247–1256, 2006.
- Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604, 2002.
- Quentin Welniarz, Yulia Worbe, and Cecile Gallea. The forward model: A unifying theory for the role of the cerebellum in motor control and sense of agency. *Frontiers in Systems Neuroscience*, 15:644059, 2021.
- Peter A White. Making causal judgments from the proportion of confirming instances: the pci rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4):710, 2003.
- David Whitney and Allison Yamanashi Leib. Ensemble perception. *Annual review of psychology*, 69(1):105–129, 2018.
- Aaron M Williams, Christopher F Angeloni, and Maria N Geffen. Sound improves neuronal encoding of visual stimuli in mouse primary visual cortex. *Journal of Neuroscience*, 43(16):2885–2906, 2023.
- Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

- Lindsay Willmore, Courtney Cameron, John Yang, Ilana B Witten, and Annegret L Falkner. Behavioural and dopaminergic signatures of resilience. *Nature*, 611(7934):124–132, 2022.
- Zachary Wojtowicz and Simon DeDeo. From probability to consilience: How explanatory values implement bayesian reasoning. *Trends in Cognitive Sciences*, 24(12):981–993, 2020.
- Sichao Yang, Johannes Bill, Jan Drugowitsch, and Samuel J Gershman. Human visual motion perception shows hallmarks of bayesian structural inference. *Scientific reports*, 11(1):3714, 2021.
- Lance Ying, Tan Zhi-Xuan, Vikash Mansinghka, and Joshua B Tenenbaum. Inferring the goals of communicating agents from actions and instructions. In *Proceedings of the AAAI symposium series*, volume 2, pages 26–33, 2023.
- Jian-Qiao Zhu, Adam N Sanborn, and Nick Chater. The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological review*, 127(5):719, 2020.
- Davide Zoccolan and Alessandro Di Filippo. Chapter 5 - Methodological Approaches to the Behavioural Investigation of Visual Perception in Rodents. In Abdel Ennaceur and Maria Angelica de Souza Silva, editors, *Handbook of Behavioral Neuroscience*, volume 27 of *Handbook of Object Novelty Recognition*, pages 69–101. Elsevier, January 2018. doi:[10.1016/B978-0-12-812012-5.00005-7](https://doi.org/10.1016/B978-0-12-812012-5.00005-7).

Appendix A

Supplementary material

Chapter 2

A.0.1 Supplementary Tables

Table A.1 Posterior summary statistics for the key population-level parameters of the AIC “computation” description. The columns represent the posterior mean (**mean**), posterior standard deviation (**sd**), the 3% and 97% boundaries of the Highest Density Interval (**hdi_3%**, **hdi_97%**), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (**mcse_mean**, **mcse_sd**) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (**ess_bulk**, **ess_tail**) (Vehtari et al., 2021), and the R-hat convergence diagnostic (**r_hat**) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019).

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.230	0.015	0.201	0.257	0.000	0.000	4006.6	3870.9	1.001
μ_{β_D}	0.711	0.031	0.654	0.768	0.000	0.000	4125.5	3965.1	1.000
σ_{β_L}	0.149	0.012	0.126	0.171	0.000	0.000	3994.9	3962.9	1.000
σ_{β_D}	0.299	0.022	0.257	0.339	0.000	0.000	3738.5	3798.4	1.000
μ_{ϵ}	0.023	0.003	0.017	0.028	0.000	0.000	3325.6	3650.5	1.000

Table A.2 Posterior summary statistics for the key population-level parameters of the BIC “computation” description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.236	0.016	0.207	0.267	0.000	0.000	4252.5	3998.3	1.0
μ_{β_D}	0.798	0.034	0.736	0.863	0.001	0.000	3888.8	3738.9	1.0
σ_{β_L}	0.156	0.012	0.134	0.180	0.000	0.000	4076.6	3945.5	1.0
σ_{β_D}	0.343	0.026	0.295	0.390	0.000	0.000	3770.2	3643.3	1.0
μ_{ε}	0.023	0.003	0.018	0.029	0.000	0.000	3529.0	3156.3	1.0

Table A.3 Posterior summary statistics for the key population-level parameters of the AIC “independent” description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	9.032	0.523	8.007	9.944	0.008	0.006	4001.5	3812.8	1.000
μ_{β_D}	0.738	0.032	0.678	0.799	0.001	0.000	3734.2	3841.7	1.000
σ_{β_L}	5.297	0.361	4.659	5.981	0.006	0.004	3970.2	4078.2	1.001
σ_{β_D}	0.315	0.024	0.273	0.364	0.000	0.000	4018.8	3608.4	1.000
μ_{ε}	0.023	0.003	0.017	0.029	0.000	0.000	2660.8	3630.7	1.000

Table A.4 Posterior summary statistics for the key population-level parameters of the AIC “heuristic scaling” description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	2.672	0.171	2.350	2.983	0.003	0.002	4066.9	3760.1	1.000
μ_{β_D}	1.469	0.064	1.353	1.592	0.001	0.001	3756.5	3645.7	1.001
σ_{β_L}	1.719	0.135	1.459	1.963	0.002	0.001	4237.0	3889.2	1.001
σ_{β_D}	0.625	0.047	0.535	0.711	0.001	0.001	4028.6	3889.8	1.000
μ_{ε}	0.022	0.003	0.017	0.027	0.000	0.000	2798.2	3595.8	1.001

Table A.5 Posterior summary statistics for the key population-level parameters of the BIC “heuristic scaling” description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	2.791	0.534	1.773	3.770	0.009	0.006	3823.6	3958.7	1.000
μ_{β_D}	1.712	0.075	1.571	1.847	0.001	0.001	3999.5	3834.3	1.001
σ_{β_L}	17.240	0.551	16.183	18.245	0.009	0.006	4017.7	4011.8	1.001
σ_{β_D}	0.742	0.057	0.638	0.847	0.001	0.001	4015.7	3995.4	1.001
μ_{ε}	0.022	0.003	0.017	0.028	0.000	0.000	2827.3	3430.4	1.000

Table A.6 Posterior summary statistics for the key population-level parameters of the AIC “free scaling” description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.351	0.025	0.304	0.397	0.000	0.000	3855.0	3688.1	1.001
μ_{β_D}	1.468	0.063	1.357	1.594	0.001	0.001	3874.7	3756.8	1.002
σ_{β_L}	0.254	0.017	0.222	0.288	0.000	0.000	3963.2	3635.8	1.001
σ_{β_D}	0.626	0.047	0.540	0.713	0.001	0.001	4207.0	3987.7	1.000
μ_{ϵ}	0.022	0.003	0.017	0.028	0.000	0.000	2568.8	3446.5	1.000
η_{pop}	2.652	0.532	1.662	3.633	0.008	0.006	4017.9	3828.6	1.001
l_{pop}	1.826	0.440	1.009	2.643	0.007	0.005	3664.0	3645.4	1.000

Table A.7 Posterior summary statistics for the key population-level parameters of the BIC “free scaling” description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.472	0.033	0.408	0.532	0.001	0.000	3627.5	3807.3	1.000
μ_{β_D}	0.974	0.043	0.895	1.058	0.001	0.000	3735.4	3826.8	1.001
σ_{β_L}	0.302	0.025	0.255	0.348	0.000	0.000	3890.5	3598.1	1.000
σ_{β_D}	0.417	0.032	0.358	0.480	0.000	0.000	4149.5	3812.9	1.000
μ_{ϵ}	0.022	0.003	0.017	0.028	0.000	0.000	3018.6	3320.4	1.001
η_{pop}	2.854	0.529	1.927	3.875	0.009	0.006	3747.5	3701.8	1.0
l_{pop}	1.442	0.378	0.770	2.112	0.006	0.005	3498.2	3691.1	1.0

Table A.8 Posterior summary statistics for the key population-level parameters of the L-only description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.057	0.006	0.047	0.069	0.000	0.000	3699.5	3485.5	1.000
σ_{β_L}	0.058	0.005	0.049	0.067	0.000	0.000	4046.5	3802.3	1.001
μ_{ϵ}	0.014	0.002	0.010	0.019	0.000	0.000	3531.9	3812.3	1.000

Table A.9 Posterior summary statistics for the key population-level parameters of the W-I description. Details are as in Table A.1.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_W}	-3.714	0.235	-4.152	-3.275	0.004	0.003	4182.6	4031.0	1.000
μ_{β_I}	0.109	0.054	0.013	0.215	0.001	0.001	4331.0	3826.4	1.000
σ_{β_W}	2.292	0.186	1.943	2.637	0.003	0.002	4208.5	4102.4	1.001
σ_{β_I}	0.536	0.041	0.463	0.618	0.001	0.000	4066.3	3930.0	1.001
μ_{ϵ}	0.004	0.001	0.002	0.007	0.000	0.000	3642.1	3930.8	1.000

Chapter 3

A.0.2 Marginal Likelihood of Gaussian Mixture Models

The marginal likelihood for the Gaussian Mixture Models used in our experiment is given by:

$$p(\mathbf{X} | K_M) = \int_{-1}^1 d\mu_1 \dots \int_{-1}^1 d\mu_K \prod_{x_i \in \mathbf{X}} \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{j=1}^{K_M} \frac{1}{K_M} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma^2}\right) \frac{1}{T^{K_M}} \quad (\text{A.1})$$

with $\sigma = 0.1$ fixed standard deviation of the components, and $T = 2$ length of the space $[-1, 1]$ where the means can be located.

To turn the integration into Gaussian integrals, whose solution is known, one should switch the sum over the K_M means with the product over the N datapoints x_i . This allows to later switch the integration and the resulting sum, thus obtaining integrals of products of Gaussians, which are themselves Gaussian integrals. The switch between the product and the sum can be obtained by summing over the *assignment* $z_i \in \{1, \dots, K_M\}$ of each point x_i to one of the K_M means, meaning that we sum over all the possible ways each point could be assigned to (i.e., could have been generated by) each of the means:

$$P(\mathbf{X} | \vec{\mu}; \sigma, K_M) = C \prod_{x_i \in \mathbf{X}} \sum_{j=1}^{K_M} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma^2}\right) = \quad (\text{A.2})$$

$$= C \sum_{z_1=1}^{K_M} \dots \sum_{z_N=1}^{K_M} \prod_{x_i \in \mathbf{X}} \exp\left(-\frac{(x_i - \mu_{z_i})^2}{2\sigma^2}\right) = \quad (\text{A.3})$$

$$= C \sum_{\vec{z} \in \{1, \dots, K_M\}^N} \prod_{j=1}^K \prod_{x_i: z_i=j} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma^2}\right) \quad (\text{A.4})$$

Where $C = \left(\frac{1}{K_M \sqrt{2\pi\sigma^2}}\right)^N$. If we integrate this term, each product on x_i is easily solvable as a Gaussian integral:

$$I_G(S_j) = \int_{-1}^1 \prod_{x_i \in S_j} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma^2}\right) d\mu_j = \int_{-1}^1 \exp\left(-\frac{\sum_{x_i \in S_j} (x_i - \mu_j)^2}{2\sigma^2}\right) d\mu_j \quad (\text{A.5})$$

Where $S_j = \{x_i : z_i = j\}$

If we focus on the argument of the exponential, we get:

$$-\frac{\sum_{x_i \in S_j} (x_i - \mu_j)^2}{2\sigma^2} = -\frac{1}{2\sigma^2} (N_j \mu_j^2 - 2(\sum x_i) \mu_j + \sum x_i^2) \quad (\text{A.6})$$

$$= -\frac{1}{2\sigma^2} \left(N_j \left(\mu - \frac{\sum x_i}{N_j} \right)^2 + \sum x_i^2 - \frac{(\sum x_i)^2}{N_j} \right) \quad (\text{A.7})$$

Where N_j is the number of data points in the partition S_j . If we substitute this back in the integral, we get:

$$I_G(S_j) = \exp \left(-\frac{1}{2\sigma^2} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{N_j} \right) \right) \int_{-T}^T \exp \left(-\frac{N_j \left(\mu - \frac{\sum x_i}{N_j} \right)^2}{2\sigma^2} \right) d\mu_j. \quad (\text{A.8})$$

The Gaussian integral has as solution:

$$I_G(S_j) = \exp \left(-\frac{1}{2\sigma^2} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{N_j} \right) \right) \sqrt{\frac{\pi\sigma^2}{2N_j}} \times \left[\text{erf} \left(\sqrt{\frac{N_j}{2\sigma^2}} \left(1 - \frac{\sum x_i}{N_j} \right) \right) + \text{erf} \left(\sqrt{\frac{N_j}{2\sigma^2}} \left(1 + \frac{\sum x_i}{N_j} \right) \right) \right] \quad (\text{A.9})$$

The marginal likelihood is therefore:

$$P(\mathbf{X} | M) = \left(\frac{1}{K_M \sqrt{2\pi\sigma^2}} \right)^N \sum_{\vec{z} \in \{1, \dots, K_M\}^N} \prod_{j=1}^K I_G(S_j) \quad (\text{A.10})$$

The derivation provides an exact analytical expression for the marginal likelihood of the Gaussian Mixture Model. However, its practical application is severely limited by its computational complexity. The formula requires summing over all possible assignments of N datapoints to K_M mixture components, resulting in $(K_M)^N$ terms. This number grows exponentially with the size of the dataset: for the model with $K_M = 5$, having $N = 50$ data-points, we would have to sum $\sim 10^{35}$ terms. This makes the direct computation of the marginal likelihood intractable for all but the most trivial of problems.

A.0.3 Higher-Order Correction to the Laplace Approximation

In the main text, we derived the Fisher Information Approximation (FIA) by taking the leading terms of the Laplace approximation of the marginal likelihood. This approximation

can be refined by including higher-order terms from the Taylor series expansion. Here we compute the next significant term in this expansion, the $O(N^{-1})$ correction, and discuss its applicability to Gaussian Mixture Models (GMMs).

The Laplace approximation is derived by approximating the function in the exponent, $N\psi(\vec{\theta})$, with a second-order Taylor expansion around its maximum $\vec{\theta}^*$. This turns the integrand into a Gaussian function. To improve this approximation, we can extend the Taylor series for $\psi(\vec{\theta})$ to include third and fourth-order terms.

For simplicity, let's first consider the one-dimensional case. The integral we wish to solve is:

$$I = \int \exp(N\psi(\theta))f(\theta)d\theta \quad (\text{A.11})$$

where, as previously discussed, $f(\theta)$ is the prior $P(\theta)$, which in our case is constant (and in general can be approximated as constant over the region of interest around the maximum likelihood estimate θ^*).

The Taylor expansion of $\psi(\theta)$ is:

$$\psi(\theta) \simeq \psi(\theta^*) + \frac{1}{2}(\theta - \theta^*)^2\psi''(\theta^*) + \frac{1}{3!}(\theta - \theta^*)^3\psi'''(\theta^*) + \frac{1}{4!}(\theta - \theta^*)^4\psi^{(4)}(\theta^*) + \dots \quad (\text{A.12})$$

By exponentiating this series, we can approximate the exponential of the higher-order terms with their own series expansion, as for large N their contribution is small:

$$\exp\left(N\left[\frac{(\theta - \theta^*)^3}{6}\psi'''(\theta^*) + \frac{(\theta - \theta^*)^4}{24}\psi^{(4)}(\theta^*) + \dots\right]\right) \simeq \quad (\text{A.13})$$

$$1 + N\left[\frac{(\theta - \theta^*)^3}{6}\psi'''(\theta^*) + \dots\right] + \frac{N^2}{2}\left[\frac{(\theta - \theta^*)^3}{6}\psi'''(\theta^*) + \dots\right]^2 + \dots \quad (\text{A.14})$$

The integral I thus becomes a series of Gaussian-like integrals involving powers of $(\theta - \theta^*)$. The odd-powered terms integrate to zero, and by collecting the terms of the same order in N , we arrive at a corrected approximation:

$$I \simeq \sqrt{\frac{2\pi}{N|\psi''(\theta^*)|}} \exp(N\psi(\theta^*))f(\theta^*) \left[1 + \frac{C_1}{N}\right] \quad (\text{A.15})$$

where the first-order correction term C_1 is given by:

$$C_1 = \frac{\psi^{(4)}}{8(\psi'')^2} - \frac{5(\psi^{(3)})^2}{24(\psi'')^3} \quad (\text{A.16})$$

with all derivatives evaluated at the maximum likelihood estimate θ^* .

This result can be generalized to the K_M -dimensional integral for our GMMs. The marginal likelihood is then approximated as:

$$P(\mathbf{X} | M) \simeq \left(\frac{2\pi}{N} \right)^{\frac{K_M}{2}} \det(\hat{F})^{-\frac{1}{2}} P(\vec{\mu}^*) \exp(\log P(\mathbf{X} | \vec{\mu}^*)) \left(1 + \frac{C_1}{N} \right) \quad (\text{A.17})$$

In the multivariate case, the correction term C_1 is significantly more complex, involving tensor contractions of the higher-order derivatives of $\psi(\vec{\mu})$. Following the notation of (Shun and McCullagh, 1995), it can be expressed as:

$$C_1 = \frac{1}{24}(\psi^{(ijkl)} H_{ij}^{-1} H_{kl}^{-1})[3] - \frac{1}{72}(\psi^{(ijk)} \psi^{(lpq)} H_{ij}^{-1} H_{kl}^{-1} H_{pq}^{-1})[15] \quad (\text{A.18})$$

where $\psi^{(ijkl)}$ represents the fourth partial derivative $\frac{\partial^4 \psi}{\partial \mu_i \partial \mu_j \partial \mu_k \partial \mu_l}$, H_{ij}^{-1} is the (i, j) -th element of the inverse of the Hessian matrix $\hat{H} = -\nabla^2 \psi(\vec{\mu}^*)$, and the bracketed numbers indicate the number of terms obtained by permutation of the indices.

To compute C_1 for our GMM, we need the partial derivatives of the single-point log-likelihood, which we define as $r(x_i, \vec{\mu}) = \log P(x_i | \vec{\mu})$, up to the fourth order. The total log-likelihood is the sum over all points, $\sum_i r(x_i, \vec{\mu})$, and its derivatives are the sum of the individual derivatives.

The derivatives with respect to the means μ_j are as follows. We use the shorthand $r^{(i)} = \frac{\partial r}{\partial \mu_i}$, $r^{(ij)} = \frac{\partial^2 r}{\partial \mu_i \partial \mu_j}$, and so on.

First derivative:

$$r^{(i)} = \frac{\exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right)}{\sum_{k=1}^{K_M} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)} \frac{x - \mu_i}{\sigma^2} \quad (\text{A.19})$$

Second derivatives:

$$r^{(ij)} = -r^{(i)} r^{(j)} \quad (\text{for } i \neq j) \quad (\text{A.20})$$

$$r^{(ii)} = -r^{(i)} \left(r^{(i)} - \frac{x - \mu_i}{\sigma^2} \right) - \frac{r^{(i)}}{x - \mu_i} \quad (\text{A.21})$$

Third derivatives:

$$r^{(ijk)} = 2r^{(i)} r^{(j)} r^{(k)} \quad (\text{for } i, j, k \text{ distinct}) \quad (\text{A.22})$$

$$r^{(iij)} = r^{(j)} \left((r^{(i)})^2 - r^{(ii)} \right) \quad (\text{for } i \neq j) \quad (\text{A.23})$$

$$r^{(iii)} = -r^{(ii)} \left(r^{(i)} - \frac{x - \mu_i}{\sigma^2} \right) - r^{(i)} r^{(ii)} - \frac{\partial}{\partial \mu_i} \left(\frac{r^{(i)}}{x - \mu_i} \right) \quad (\text{A.24})$$

Fourth derivatives:

$$r^{(ijkl)} = -6r^{(i)}r^{(j)}r^{(k)}r^{(l)} \quad (\text{for } i, j, k, l \text{ distinct}) \quad (\text{A.25})$$

$$r^{(iijk)} = 2r^{(ii)}r^{(j)}r^{(k)} - 4(r^{(i)})^2r^{(j)}r^{(k)} \quad (\text{for } i, j, k \text{ distinct}) \quad (\text{A.26})$$

$$r^{(iijj)} = -r^{(ii)}r^{(jj)} + (r^{(i)})^2r^{(jj)} + (r^{(j)})^2r^{(ii)} - 3(r^{(i)})^2(r^{(j)})^2 \quad (\text{for } i \neq j) \quad (\text{A.27})$$

$$r^{(iiij)} = -r^{(j)}r^{(i)} \left((r^{(i)})^2 - r^{(ii)} \right) + r^{(j)} \left(2r^{(i)}r^{(ii)} - r^{(iii)} \right) \quad (\text{for } i \neq j) \quad (\text{A.28})$$

$$= r^{(j)} \left[-(r^{(i)})^3 + 3r^{(i)}r^{(ii)} - r^{(iii)} \right] \quad (\text{A.29})$$

$$r^{(iiii)} = 6(r^{(i)})^2r^{(ii)} + r^{(ii)} \frac{(x - \mu_i)^2}{\sigma^4} - 6r^{(i)}r^{(ii)} \frac{x - \mu_i}{\sigma^2} + 3f_i r^{(ii)} \quad (\text{A.30})$$

$$- 3 \frac{r^{(ii)}}{\sigma^2} - 2r^{(i)} \frac{x - \mu_i}{\sigma^4} - 3f_i (r^{(i)})^2 + 6 \frac{(r^{(i)})^2}{\sigma^2} \quad (\text{A.31})$$

While the term C_1 provides a more accurate approximation, it does not resolve the fundamental issue that plagues the Laplace approximation for GMMs. The standard FIA term, which depends on $\det(\hat{F})^{-\frac{1}{2}}$, fails when the model becomes non-identifiable. This occurs, for instance, when two component means are very close (but not quite overlapping perfectly), i.e., $\mu_i \simeq \mu_j$ for $i \neq j$.

In this scenario, the likelihood function develops a ridge, and the parameters are no longer locally distinct. As a consequence, the Hessian matrix $\hat{H}$ (and thus the Fisher Information Matrix $\hat{F}$) becomes singular at the maximum likelihood estimate $\vec{\mu}^*$. A singular Hessian means its determinant is zero, causing the $\log(\det \hat{F})$ term in the FIA formula to diverge to $-\infty$, or more practically to result in a very negative value, if the means are close.

The higher-order correction term C_1 is critically dependent on H_{ij}^{-1} , the inverse of the Hessian matrix. If the Hessian matrix is singular, its inverse does not exist. Therefore, the very same condition of parameter non-identifiability that causes the failure of the standard FIA also makes the $O(N^{-1})$ term unreliable. Adding more terms to the approximation does not cure the underlying problem stemming from the geometric properties of the GMM likelihood surface.

A.0.4 Supplementary Tables

Table A.10 Posterior summary statistics for the key population-level parameters of the BIC “computation” description. The columns represent the posterior mean (**mean**), posterior standard deviation (**sd**), the 3% and 97% boundaries of the Highest Density Interval (**hdi_3%**, **hdi_97%**), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (**mcse_mean**, **mcse_sd**) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (**ess_bulk**, **ess_tail**) (Vehtari et al., 2021), and the R-hat convergence diagnostic (**r_hat**) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019).

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.161	0.012	0.138	0.182	0.000	0.000	4497.3	2705.7	1.001
μ_{β_D}	0.698	0.052	0.599	0.795	0.001	0.001	4833.0	2511.6	1.002
σ_{β_L}	0.079	0.009	0.062	0.095	0.000	0.000	4199.3	3047.5	1.001
σ_{β_D}	0.349	0.041	0.278	0.426	0.001	0.000	4386.0	2989.0	1.001
μ_{ϵ}	0.028	0.005	0.019	0.037	0.000	0.000	764.1	1553.1	1.002

Table A.11 Posterior summary statistics for the key population-level parameters of the “free penalty” description. Details are as in Table A.10.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.164	0.012	0.143	0.188	0.000	0.000	7174.9	2732.5	1.001
μ_{β_D}	0.553	0.035	0.489	0.621	0.000	0.000	8244.1	2778.7	1.001
$\mu_{\hat{\kappa}_1}$	0.000	0.000	0.000	0.000	0.000	0.000	4000.0	4000.0	-
$\mu_{\hat{\kappa}_2}$	1.506	0.189	1.145	1.859	0.002	0.001	8241.6	2853.9	1.001
$\mu_{\hat{\kappa}_3}$	2.917	0.254	2.424	3.365	0.003	0.002	9640.1	2742.5	1.000
$\mu_{\hat{\kappa}_4}$	4.023	0.280	3.483	4.539	0.003	0.002	10308.2	2690.5	1.001
$\mu_{\hat{\kappa}_5}$	4.783	0.303	4.233	5.369	0.003	0.002	8244.1	2778.7	1.001
σ_{β_L}	0.077	0.009	0.061	0.094	0.000	0.000	8766.2	3326.5	1.000
σ_{β_D}	0.241	0.027	0.191	0.292	0.000	0.000	8076.7	2701.7	1.003
$\sigma_{\hat{\kappa}_1}$	0.000	0.000	0.000	0.000	0.000	0.000	4000.0	4000.0	-
$\sigma_{\hat{\kappa}_2}$	1.276	0.142	1.034	1.560	0.002	0.001	6767.4	2762.4	1.001
$\sigma_{\hat{\kappa}_3}$	1.720	0.190	1.390	2.087	0.002	0.002	6959.6	2194.3	1.001
$\sigma_{\hat{\kappa}_4}$	1.928	0.210	1.535	2.313	0.003	0.002	8810.7	2578.5	1.003
$\sigma_{\hat{\kappa}_5}$	2.084	0.233	1.655	2.528	0.003	0.002	8076.7	2701.7	1.003
μ_{ϵ}	0.025	0.004	0.018	0.034	0.000	0.000	1430.9	2519.1	1.000

Table A.12 Posterior summary statistics for the key population-level parameters of the “naive correction” description. Details are as in Table A.10.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.159	0.012	0.137	0.181	0.000	0.000	6708.3	2877.8	1.000
μ_{β_D}	1.238	0.182	0.904	1.574	0.002	0.001	8257.9	3129.0	1.001
$\mu_{\beta_{S_{naive}}}$	0.736	0.258	0.253	1.214	0.003	0.002	6832.4	3080.3	1.000
σ_{β_L}	0.077	0.009	0.062	0.094	0.000	0.000	6158.9	3277.4	1.001
σ_{β_D}	1.241	0.098	1.053	1.418	0.001	0.001	5585.3	3217.1	1.000
$\sigma_{\beta_{S_{naive}}}$	3.007	2.954	0.000	8.361	0.035	0.034	5603.1	2019.5	1.002
μ_{ϵ}	0.025	0.004	0.016	0.033	0.000	0.000	1018.8	1482.7	1.004

Table A.13 Posterior summary statistics for the key population-level parameters of the “correction” description. Details are as in Table A.10.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.117	0.009	0.101	0.134	0.000	0.000	7866.7	2877.4	1.001
μ_{β_D}	1.203	0.080	1.052	1.354	0.001	0.001	6712.2	3212.9	1.001
μ_{β_S}	1.035	0.105	0.841	1.227	0.001	0.001	6511.8	2865.4	1.002
σ_{β_L}	0.058	0.007	0.046	0.071	0.000	0.000	6060.2	3270.9	0.999
σ_{β_D}	0.553	0.064	0.438	0.672	0.001	0.001	5448.7	3132.7	1.000
σ_{β_S}	0.680	0.079	0.533	0.823	0.001	0.001	6107.4	3173.8	0.999
μ_{ϵ}	0.022	0.004	0.014	0.029	0.000	0.000	967.1	1397.2	1.006

Table A.14 Posterior summary statistics for the key population-level parameters of the “FIA” description. Details are as in Table A.10.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_{β_L}	0.103	0.007	0.090	0.118	0.000	0.000	7725.8	2932.1	1.000
μ_{β_D}	1.127	0.076	0.977	1.261	0.001	0.001	7971.5	2914.2	1.001
μ_{β_S}	0.984	0.099	0.799	1.171	0.001	0.001	7565.9	2416.7	1.001
μ_{β_R}	-0.030	0.073	-0.169	0.110	0.001	0.001	7588.1	2780.3	1.002
σ_{β_L}	0.047	0.006	0.035	0.058	0.000	0.000	5577.1	3613.5	1.001
σ_{β_D}	0.492	0.058	0.393	0.607	0.001	0.001	6044.3	3513.7	1.000
σ_{β_S}	0.647	0.074	0.509	0.786	0.001	0.001	6780.5	3155.7	1.000
σ_{β_R}	0.466	0.056	0.371	0.577	0.001	0.001	5592.6	3072.2	1.001
μ_{ϵ}	0.021	0.004	0.015	0.028	0.000	0.000	1282.3	2238.2	1.001

Chapter 4

A.0.5 Supplementary Figures

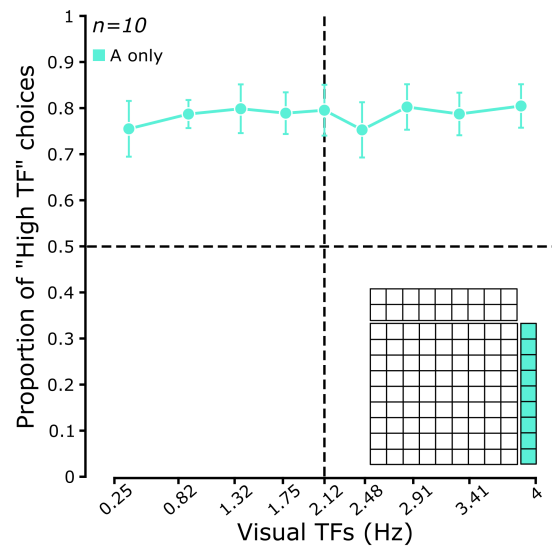
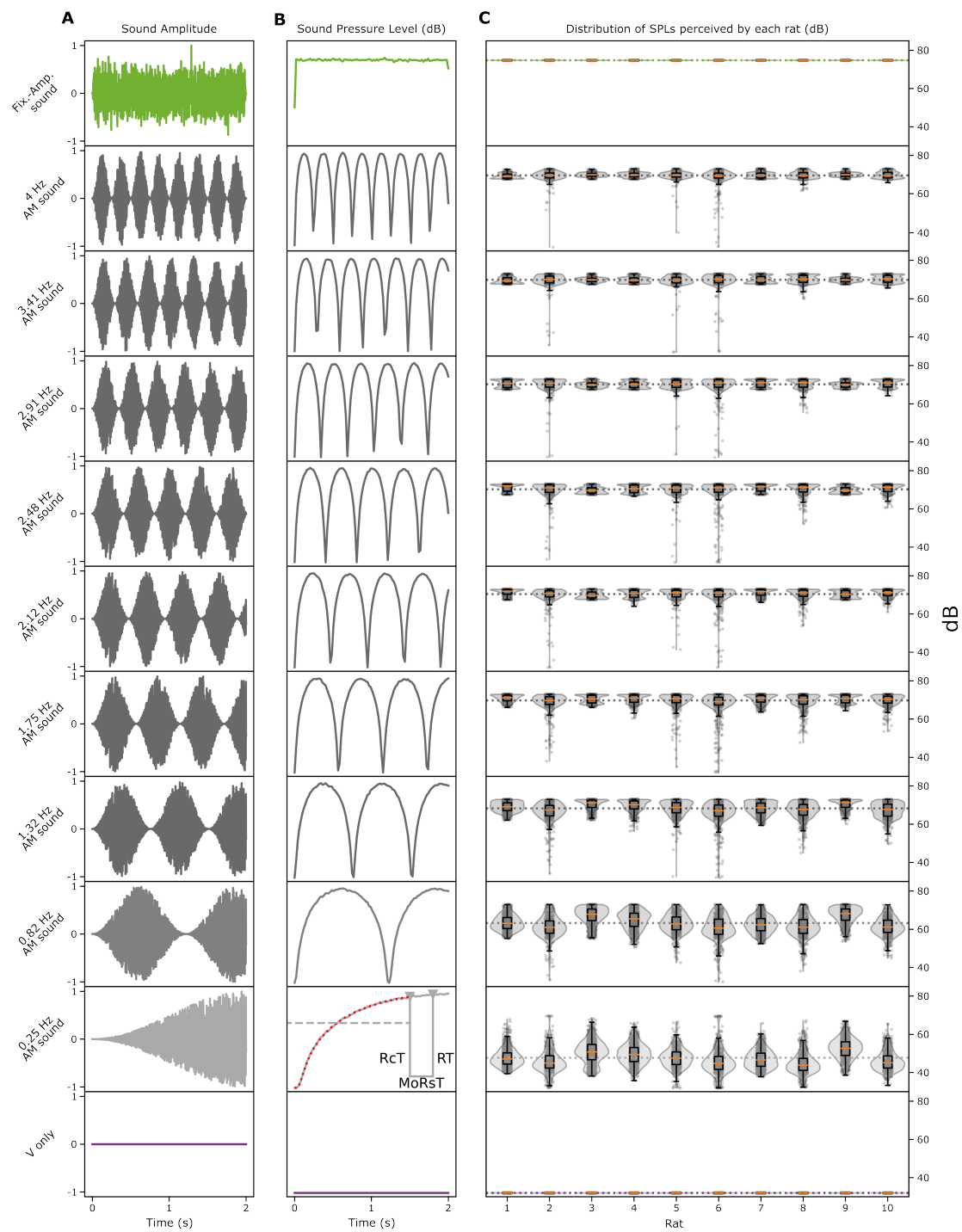


Fig. A.1 **Rats are not sensitive to the temporal frequency of the unimodal, purely auditory stimuli.** Group average proportion of "high TF" choices ($n=10$ rats) as a function of the TF of the unimodal auditory stimuli. Error bars are SEM.



(Caption on next page.)

Fig. A.2 The auditory stimuli and their intensity. Each row refers to a distinct auditory stimulus used in the experiment: 1) the white noise burst with fixed maximal amplitude (top row; green); 2) the amplitude modulated white noise bursts, with the 9 different temporal frequencies of the sinusoidal envelopes (middle rows; gray); and 3) the absence of the auditory stimulus (bottom row; purple). **A.** Waveforms of the auditory stimuli used to drive the speakers that delivered the sounds to the rats (note that, for the sake of visualization, the sounds are not shown at their actual sampling frequency of 44.1 KHz, but they have been downsampled to 500 Hz). **B.** Intensities of the auditory stimuli in dB as a function of time, as computed based on the waveforms shown in **A** and the measured intensity of 10 s long white noise bursts with different amplitudes (see the Materials and Methods for details). The panel referring to the TF of 0.25 Hz illustrates how the average sound intensity experienced by a rat in a given trial was computed, based on the reaction time RcT of the animal. The latter was obtained by subtracting the estimated motor response time $MoReT$ to the measured response time RT . **C.** Distributions and estimates of the average sound intensities experienced by each rat across all trials recorded for any given stimulus condition. Scatter plots show the estimated sound intensity experienced by the rats in each individual trial. Violin plots represent the full probability density distributions of these estimates while box plots show their median and interquartile range. The dashed line indicates the across-rat average intensity per stimulus condition.

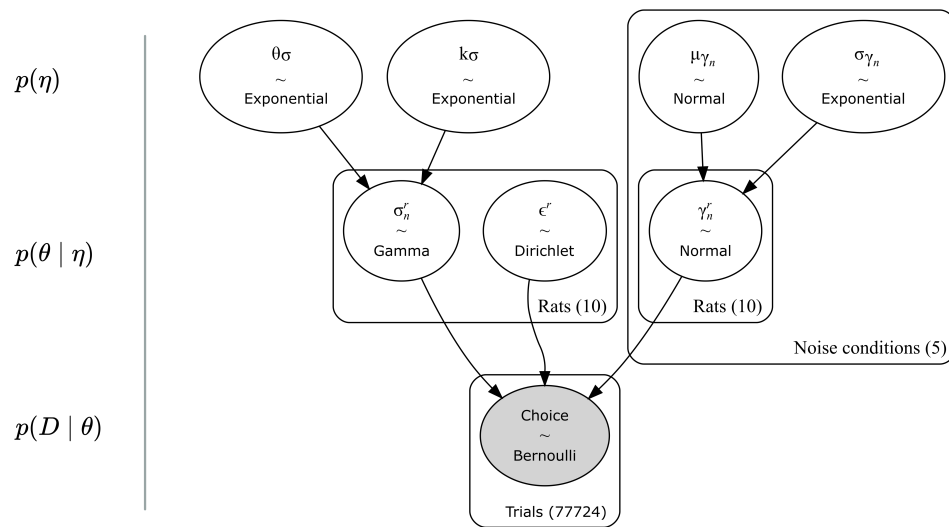


Fig. A.3 **Schematic of Bayesian hierarchical inference for our ideal observer model.** On the left, each row is labeled as representing the population parameters priors ($p(\eta)$), the subject-level parameters priors ($p(\theta | \eta)$) or the likelihood ($p(D | \theta)$). On the right, each bubble represents a parameter of the model. The name of the parameter and its distribution are reported. The rectangular plates indicate that multiple iid variables have been grouped. The group name and the corresponding number of variables are indicated on the bottom left of each plate. Each arrow represents the dependencies of the model. Gray coloring indicates that the variable is conditioned on observations.

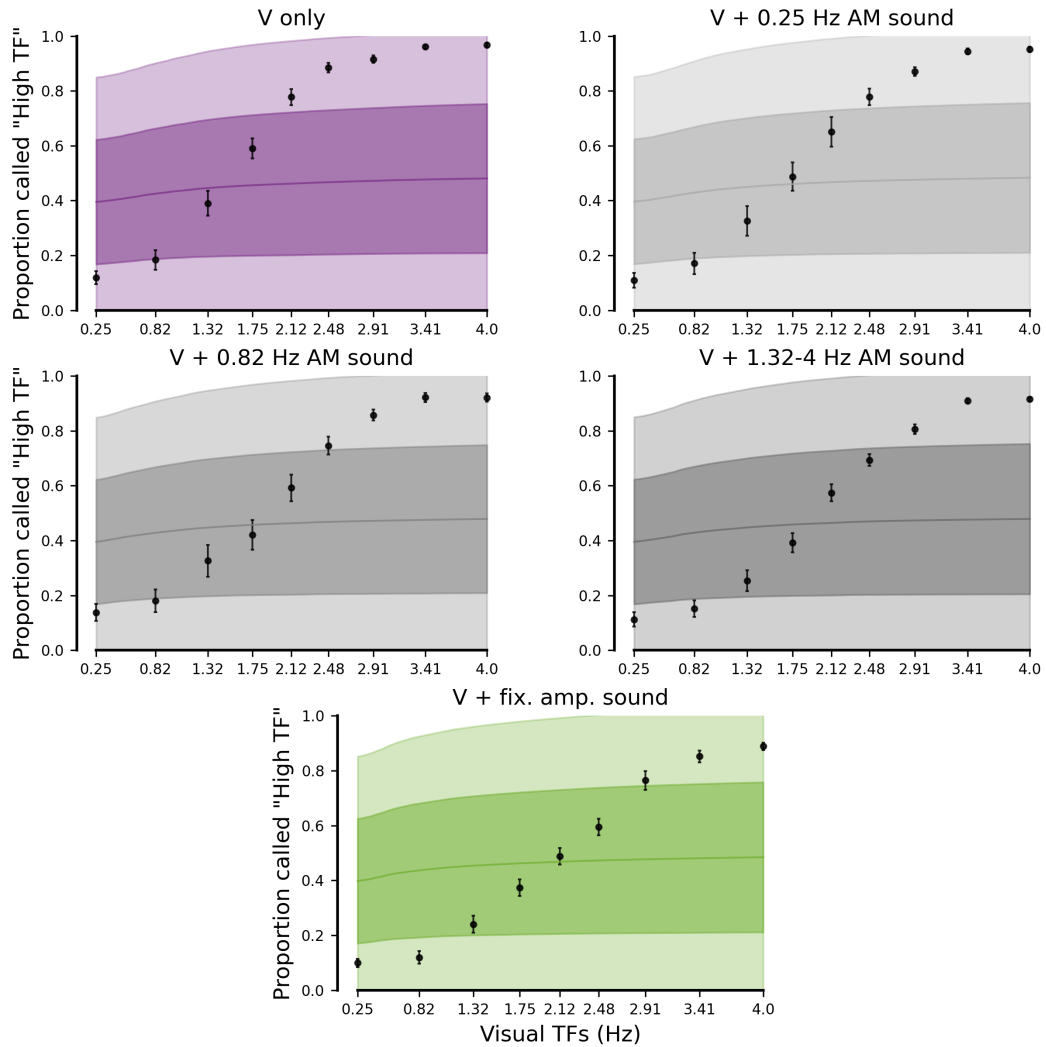


Fig. A.4 **Prior predictive simulation of the 5γ 1σ model.** In black, the average proportion of answers "High" of the rats. Error bars denote the s.e.m. across all rats. Each colored line is the average predicted psychometric curve obtained for one of the five noise conditions, generated by parameters sampled from its respective prior distributions. The γ and σ parameters are drawn from the population average parameters; the ε parameters, given the absence of a hierarchical structure, are averaged across all rats. The shaded areas represent one (dark area) and two (dim area) standard deviations from the mean. The five curves all look almost identical (up to sampling noise) as the priors for the noise conditions were the same.

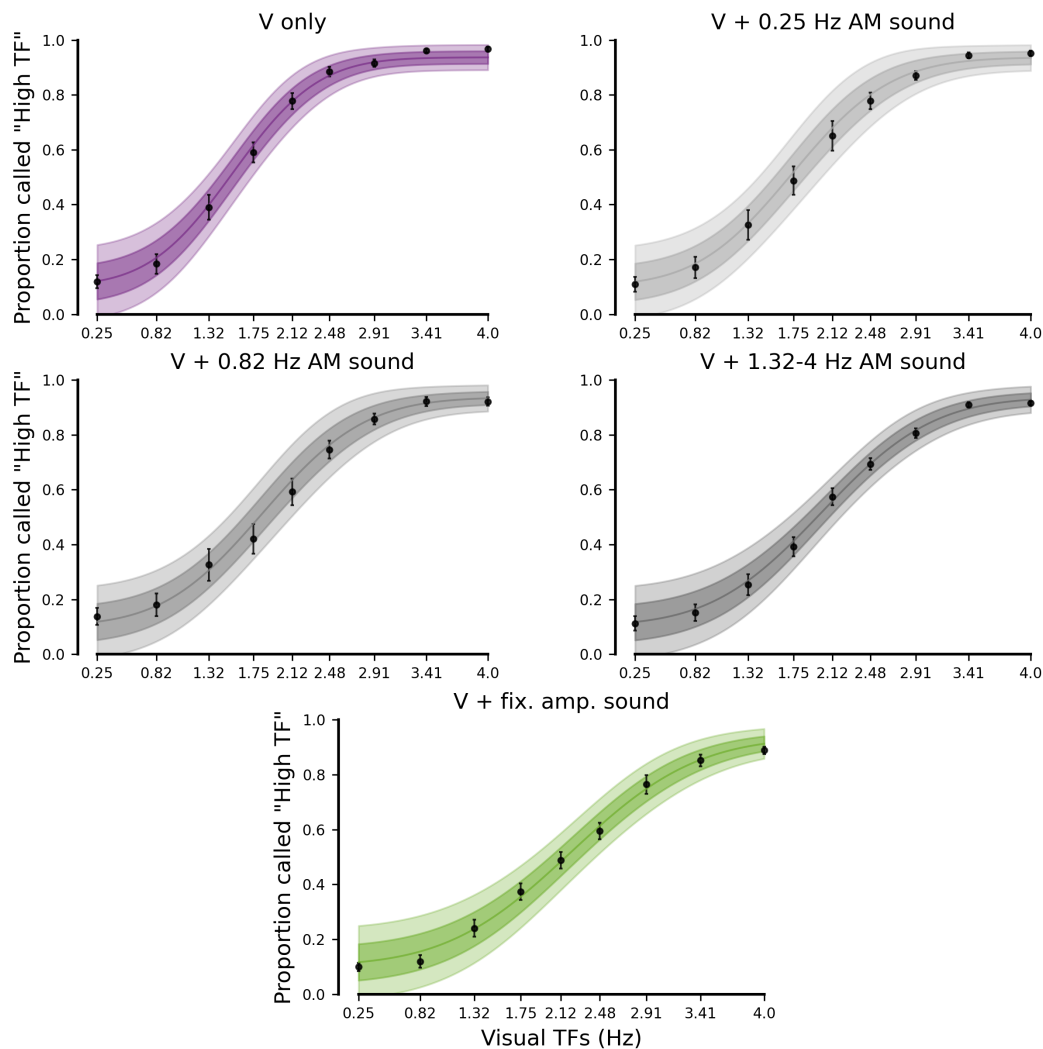


Fig. A.5 **Posterior predictive simulation of the $5\gamma 1\sigma$ model.** In black, the average proportion of answers "High" of the rats. Error bars denote the s.e.m. across all rats. Each colored line is the average predicted psychometric curve obtained for one of the five noise conditions, generated by the parameters sampled in the MCMC chain, following the respective posterior distributions (same as the solid lines displayed in Fig. 4.5A). The γ and σ parameters are drawn from the population average parameters; the ϵ parameters, given the absence of a hierarchical structure, are averaged across all rats. The shaded areas represent one (dark area) and two (dim area) standard deviations from the mean of the posterior.

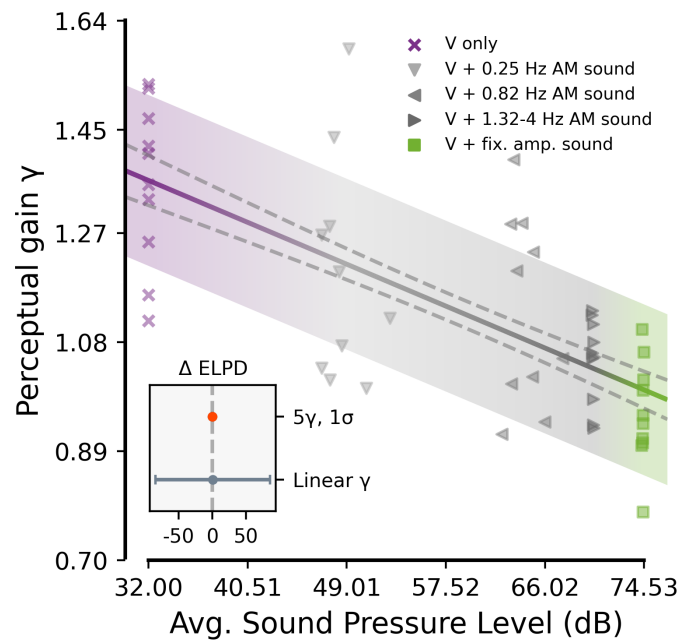


Fig. A.6 Posterior prediction of the linear relationship between sound intensity and perceptual gain factor in an ideal observer model. The solid line represents the population-level gain parameter μ_γ , modeled as a linear function of the average sound pressure level I . The slope of the line is $(-8.6 \pm 1.4) \times 10^{-3}$, the intercept 1.636 ± 0.083 (posterior mean $\pm$ st. dev.). The color gradient along the line maps the specific sound conditions: green for the fixed amplitude condition, grays for the V+AM conditions, and purple for the V-only condition. The dashed lines represent one standard deviation of μ_γ . The shaded area around the regression line represents the posterior mean of the between-subject standard deviation, σ_γ . This area illustrates the estimated population-level variability. The symbols (triangle, square, cross) denote the relative average perceived sound pressure (x axis) and the posterior means (y axis) for individual rats. The inset shows the relative distance in Expected Log-Predictive Density (ELPD) between the linear model and our best model (5γ , 1σ), error bars denote the standard deviation of the difference.

A.0.6 Supplementary Tables

Table A.15 Posterior summary statistics for the key population-level parameters of the model. The columns represent the posterior mean (**mean**), posterior standard deviation (**sd**), the 3% and 97% boundaries of the Highest Density Interval (**hdi_3%**, **hdi_97%**), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (**mcse_mean**, **mcse_sd**) (Gelman et al., 2014a), the bulk and tail Effective Sample Size (**ess_bulk**, **ess_tail**) (Vehtari et al., 2021), and the R-hat convergence diagnostic (**r_hat**) (Vehtari et al., 2021). All diagnostics were computed with ArviZ version 0.15.1 (Kumar et al., 2019). Note that the μ_ε values are not population parameters in the same hierarchical sense as the others; instead, they represent the simple average of the individual rat ε values estimated, as the ε parameters were modeled without a group-level distribution.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ_γ [V-only]	1.360	0.064	1.235	1.476	0.001	0.001	4046.0	2638.0	1.0
μ_γ [0.25 Hz]	1.209	0.083	1.051	1.361	0.001	0.001	4148.0	2599.0	1.0
μ_γ [0.82 Hz]	1.146	0.074	1.006	1.282	0.001	0.001	4764.0	2514.0	1.0
μ_γ [1.32-4 Hz]	1.048	0.031	0.990	1.106	0.000	0.000	3842.0	2210.0	1.0
μ_γ [Fixed sound]	0.955	0.038	0.883	1.026	0.001	0.000	3556.0	2558.0	1.0
σ_γ [V-only]	0.179	0.053	0.093	0.275	0.001	0.001	4440.0	2945.0	1.0
σ_γ [0.25 Hz]	0.255	0.080	0.131	0.395	0.001	0.001	4070.0	2756.0	1.0
σ_γ [0.82 Hz]	0.218	0.070	0.112	0.345	0.001	0.001	3984.0	2605.0	1.0
σ_γ [1.32-4 Hz]	0.090	0.028	0.046	0.139	0.001	0.000	3276.0	2714.0	1.0
σ_γ [Fixed sound]	0.106	0.035	0.054	0.168	0.001	0.001	2988.0	1973.0	1.0
μ_σ	0.876	0.093	0.714	1.045	0.002	0.001	3141.0	2399.0	1.0
σ_σ	0.263	0.092	0.137	0.424	0.002	0.001	2798.0	2162.0	1.0
μ_ε [High]	0.102	0.005	0.093	0.112	0.0	0.0	1219.0	1749.0	1.0
μ_ε [Task]	0.834	0.007	0.822	0.847	0.0	0.0	1475.0	2021.0	1.0
μ_ε [Low]	0.063	0.003	0.058	0.069	0.0	0.0	4128.0	3269.0	1.0

A.1 Chapter 5

A.1.1 Hilbert Space Gaussian Process

Gaussian Processes, while very powerful, become quickly intractable when the amount of data to fit increases. Indeed, to invert the covariance matrix, or to compute its Cholesky factorization (as we did in Chapter 2), we need $O(N^3)$ operations, where N is the number of points to fit.

The Hilbert Space Gaussian Process approximation (Riutort-Mayol et al., 2023; Solin and Särkkä, 2020) is a way to reduce the computational complexity of Gaussian Processes, reducing the number of operations to $O(MN + M)$, where M is the number of basis functions used in the approximation (a value that can be manipulated to control the computation time).

The goal of the approximation is to avoid computing the $N \times N$ covariance matrix entirely, approximating the GP as a sum of M basis functions. To do so, we start by approximating the kernel function $k(x, x')$. The function we used, the Matern 5/2 (Williams and Rasmussen, 2006), only depends on $r = x - x'$: we can therefore rewrite it as the Fourier transform of its spectral density $S(\omega)$:

$$k(x, x') = \frac{1}{2\pi} \int_{\mathbb{R}} S(\omega) e^{i\omega(x-x')} d\omega \quad (\text{A.32})$$

Where:

$$S(\omega) = \eta^2 \frac{16 l^5}{3(l^2 + \omega^2)^3} \quad (\text{A.33})$$

for Matern 5/2.

The spectral density can be expanded as a Taylor series. More specifically, we want to find a function $\psi(\omega^2) = S(\omega)$ to expand in terms of ω^2 : the link between ω^2 and the Laplacian operator (that in the one-dimensional space $\mathbb{R}$ corresponds to the second derivative) is the core of this approximation. The Taylor expansion gives us:

$$S(\omega) = \psi(\omega^2) = a_0 + a_1(\omega^2) + a_2(\omega^2)^2 + \dots \quad (\text{A.34})$$

Notice that the specific function ψ or its coefficients a_j are not relevant, as the approximation is only used as a bridge to introduce the Laplacian operator. To use this expansion, we define

an operator $\mathcal{K}$ that applies the kernel to a function $f(x')$:

$$\mathcal{K}f(x) = \int_{\mathbb{R}} k(x, x') f(x') dx' \quad (\text{A.35})$$

$$= \int_{\mathbb{R}} \left[\frac{1}{2\pi} \int_{\mathbb{R}} S(\omega) e^{i\omega(x-x')} d\omega \right] f(x') dx' \quad (\text{A.36})$$

$$= \mathcal{F}^{-1} \left[S(\omega) \int_{\mathbb{R}} f(x') e^{-i\omega x'} dx' \right] \quad (\text{A.37})$$

$$= \mathcal{F}^{-1} [S(\omega) \hat{f}(\omega)] \quad (\text{A.38})$$

$$= \mathcal{F}^{-1} \left[\left(\sum_j a_j \omega^{2j} \right) \hat{f}(\omega) \right] \quad (\text{A.39})$$

$$= \sum_j a_j \mathcal{F}^{-1} [\omega^{2j} \hat{f}(\omega)] \quad (\text{A.40})$$

Where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform and $\hat{f}(\omega)$ is the Fourier transform of $f(x')$. By taking advantage of the fact that $\mathcal{F} [\nabla^2 f] (\omega) = -\omega^2 \hat{f}(\omega)$, where ∇^2 is the Laplacian operator, we get:

$$\mathcal{K}f(x) = \sum_j a_j (-\nabla^2)^j f(x). \quad (\text{A.41})$$

Our goal is now to use the spectral decomposition of the Laplacian. To do so, we specify the operator in a space $\Omega \subset \mathbb{R}$ with Dirichlet boundary conditions: in one dimension, this corresponds to defining a region $[-L, L]$ of the real line such that $f(x)$ is only defined in the region and $f(L) = f(-L) = 0$. While this is not always true for the functions we want to fit with GPs, practically, we can take $[-L, L]$ to be slightly larger than the actual data range so boundary effects won't affect the inference.

Inside the bounded space $[-L, L]$, the Laplacian operator is a self-adjoint operator defined on the Hilbert space $L^2(\Omega)$. In simpler terms, this means that we can use spectral decomposition on it: there are functions $\phi_k(x)$ defined in this space (i.e., orthonormal vectors of the Hilbert space) such that $-\nabla^2 \phi_k(x) = \lambda_k \phi_k(x)$, with $\lambda_k \in \mathbb{R}$, and therefore:

$$-\nabla^2 f(x) = \sum \lambda_k \phi_k(x) \langle \phi_k, f \rangle \quad (\text{A.42})$$

Where $\langle \phi_k, f \rangle = \int_{-L}^L \phi_k(x') f(x') dx'$ is the inner product. In one dimension, the explicit solutions are:

$$\lambda_k = \left(\frac{k\pi}{2L} \right)^2, \quad \phi_k(x) = \frac{1}{\sqrt{L}} \sin \left(\frac{k\pi(x+L)}{2L} \right) \quad (\text{A.43})$$

Moreover, functional calculus allows us to compute the powers of the operator: $-(\nabla^2)^j f(x) = \sum_k \lambda_k^j \phi_k(x) \langle \phi_k, f \rangle$. We can therefore rewrite Eq. A.41 as:

$$\mathcal{K} f(x) = \sum_j a_j (-\nabla^2)^j f(x) \quad (\text{A.44})$$

$$= \sum_j a_j \sum_k \lambda_k^j \phi_k(x) \langle \phi_k, f \rangle \quad (\text{A.45})$$

$$= \sum_j a_j \sum_k \lambda_k^j \phi_k(x) \int_{-L}^L \phi_k(x') f(x') dx' \quad (\text{A.46})$$

$$= \int_{-L}^L \sum_k \left[\sum_j a_j \lambda_k^j \right] \phi_k(x) \phi_k(x') f(x') dx' \quad (\text{A.47})$$

$$= \int_{-L}^L \sum_k S(\sqrt{\lambda_k}) \phi_k(x) \phi_k(x') f(x') dx' \quad (\text{A.48})$$

In the last passage, we used the Taylor expansion defined for $S(\omega) = \psi(\omega^2)$ to replace the sum of coefficients with the function $S(\cdot)$ directly. From Eq. A.35 and the fact that we assumed $f(x)$ to be only defined inside $[-L, L]$, we get:

$$k(x, x') \simeq \sum_k S(\sqrt{\lambda_k}) \phi_k(x) \phi_k(x') \quad (\text{A.49})$$

(which we write as an approximation due to the restriction of the domain and the boundary conditions).

The covariance matrix obtained from the kernel is therefore $\Sigma = \Phi \mathcal{D} \Phi^T$, with $\Phi_{ik} = \phi_k(x_i)$ and $\mathcal{D} = \text{diag}(S(\sqrt{\lambda_1}), \dots, S(\sqrt{\lambda_M}))$. Therefore, assuming mean 0, a function sampled from the approximated Gaussian process $f(x) \sim \text{MultivariateNormal}(0, \Sigma)$ will have the form:

$$f(x) = \sum_{k=1}^M \left(S(\sqrt{\lambda_k}) \right)^{1/2} \phi_k(x) z_k \quad (\text{A.50})$$

where we used the non-centered parametrization of the function with an independent prior distribution on $z_k \sim \mathcal{N}(0, 1)$. If we define $\tilde{S}(\theta) = \left(S(\sqrt{\lambda_k}; \theta) \right)^{1/2}$, we obtain equation 5.5. The fundamental property of this approximation is that the only dependence on the hyperparameters (η and l for the Matern 5/2 and the squared exponential kernels) is through the spectral density $S(\omega)$: therefore, when inferring the shape of the function, the problem is reduced to a linear regression in the eigenvector basis.

A.1.2 Supplementary Tables

Table A.16 Posterior summary statistics for the key population-level parameters of the model with binary predictors only on the LTSL population. The columns represent the posterior mean (**mean**), posterior standard deviation (**sd**), the 3% and 97% boundaries of the Highest Density Interval (**hdi_3%**, **hdi_97%**), as well as a number of standard diagnostics for MCMC-based Bayesian inference, namely the Monte Carlo Standard Error for the mean and standard deviation (**mcse_mean**, **mcse_sd**) ([Gelman et al., 2014a](#)), the bulk and tail Effective Sample Size (**ess_bulk**, **ess_tail**) ([Vehtari et al., 2021](#)), and the R-hat convergence diagnostic (**r_hat**) ([Vehtari et al., 2021](#)). All diagnostics were computed with ArviZ version 0.15.1 ([Kumar et al., 2019](#)).

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$\eta_{\text{intercept}}^g$	0.110	0.014	0.085	0.136	0.000	0.000	1123.8	1848.2	1.000
$\eta_{\text{previousOutcome}}^g$	0.106	0.013	0.083	0.132	0.000	0.000	1146.7	1823.1	1.003
$\eta_{\text{currentOutcome}}^g$	0.131	0.017	0.100	0.162	0.000	0.000	1321.0	2074.4	1.005
$\eta_{\text{rewardDirection}}^g$	0.044	0.006	0.034	0.056	0.000	0.000	1463.8	2321.1	0.999
η_{pushPull}^g	0.113	0.014	0.089	0.140	0.000	0.000	1617.1	2189.6	1.003
$l_{\text{intercept}}^g$	11.622	1.116	9.513	13.667	0.029	0.021	1425.1	2929.2	1.000
$l_{\text{previousOutcome}}^g$	13.543	1.435	10.958	16.314	0.036	0.025	1606.4	2703.7	1.000
$l_{\text{currentOutcome}}^g$	12.354	1.216	9.992	14.548	0.028	0.020	1822.0	2702.1	1.001
$l_{\text{rewardDirection}}^g$	9.764	1.297	7.516	12.378	0.030	0.021	1886.0	2697.9	1.001
l_{pushPull}^g	11.040	0.983	9.202	12.877	0.021	0.015	2168.4	2774.8	1.000
σ_{σ}	0.704	0.083	0.553	0.861	0.001	0.001	6316.5	3106.7	1.001
ν	11.377	0.136	11.116	11.629	0.002	0.002	3669.3	2624.8	1.000

Table A.17 Posterior summary statistics for the key population-level parameters of the model with binary predictors only on the THIN population. Details are as in Table A.16.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$\eta^g_{\text{intercept}}$	0.185	0.024	0.143	0.227	0.001	0.000	1600.5	2436.5	1.004
$\eta^g_{\text{previousOutcome}}$	0.088	0.011	0.068	0.110	0.000	0.000	1851.0	2765.5	1.004
$\eta^g_{\text{currentOutcome}}$	0.221	0.027	0.170	0.269	0.001	0.000	1897.7	2319.8	1.001
$\eta^g_{\text{rewardDirection}}$	0.030	0.004	0.022	0.039	0.000	0.000	2691.7	3139.4	1.000
η^g_{pushPull}	0.092	0.012	0.071	0.114	0.000	0.000	1917.6	2224.7	1.002
$l^g_{\text{intercept}}$	7.768	0.681	6.493	9.011	0.015	0.010	2117.2	2814.9	1.003
$l^g_{\text{previousOutcome}}$	8.275	1.008	6.464	10.192	0.023	0.016	1973.3	2446.6	1.002
$l^g_{\text{currentOutcome}}$	8.783	0.791	7.307	10.247	0.018	0.012	2022.7	2414.5	1.000
$l^g_{\text{rewardDirection}}$	4.935	0.822	3.449	6.464	0.016	0.012	2567.4	2845.3	1.001
l^g_{pushPull}	7.705	0.831	6.179	9.255	0.017	0.012	2298.5	2902.5	1.002
σ	0.744	0.082	0.594	0.893	0.001	0.001	9047.5	2937.6	1.000
ν	11.326	0.137	11.059	11.576	0.002	0.001	5779.1	3109.8	1.003

Table A.18 Posterior summary statistics for the key population-level parameters of the model with binary and continuous predictors only on the LTSI population. Details are as in Table A.16.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$\eta_{\text{intercept}}^g$	0.137	0.018	0.103	0.169	0.001	0.000	850.6	1479.5	1.002
$\eta_{\text{previousOutcome}}^g$	0.068	0.009	0.052	0.085	0.000	0.000	1397.4	2191.4	1.001
$\eta_{\text{currentOutcome}}^g$	0.148	0.018	0.113	0.180	0.001	0.000	1160.3	1848.0	1.001
η_{pushPull}^g	0.127	0.016	0.099	0.158	0.000	0.000	1282.7	2246.5	1.002
η_{RPE}^g	0.032	0.004	0.023	0.039	0.000	0.000	1457.9	1979.2	1.000
$\eta_{\text{rewardRate}}^g$	0.049	0.006	0.038	0.061	0.000	0.000	1233.6	2245.7	1.003
$\eta_{\text{choiceLatency}}^g$	0.025	0.003	0.019	0.031	0.000	0.000	1319.4	1960.9	1.004
$l_{\text{intercept}}^g$	12.931	1.252	10.615	15.250	0.034	0.024	1325.8	2014.1	1.003
$l_{\text{previousOutcome}}^g$	11.219	1.355	8.756	13.739	0.029	0.021	2133.0	2718.1	1.001
$l_{\text{currentOutcome}}^g$	12.430	1.233	10.235	14.842	0.029	0.021	1785.9	2656.0	1.002
l_{pushPull}^g	11.688	1.071	9.744	13.750	0.026	0.018	1715.7	2190.5	1.002
l_{RPE}^g	9.367	1.220	7.014	11.543	0.028	0.020	1957.2	2550.4	1.001
$l_{\text{rewardRate}}^g$	12.179	1.388	9.684	14.976	0.030	0.021	2160.8	2401.3	1.000
$l_{\text{choiceLatency}}^g$	9.680	1.218	7.453	11.976	0.028	0.020	1887.1	2624.0	1.001
σ	0.693	0.082	0.545	0.844	0.001	0.001	4721.6	2832.1	1.001
ν	10.861	0.128	10.627	11.112	0.002	0.002	3313.1	2447.0	1.001

Table A.19 Posterior summary statistics for the key population-level parameters of the model with binary and continuous predictors only on the THIN population. Details are as in Table A.16.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$\eta^g_{\text{intercept}}$	0.174	0.022	0.136	0.216	0.000	0.000	4737.3	4529.8	1.000
$\eta^g_{\text{previousOutcome}}$	0.094	0.012	0.073	0.117	0.000	0.000	5825.9	4866.3	1.000
$\eta^g_{\text{currentOutcome}}$	0.217	0.027	0.166	0.267	0.000	0.000	5508.7	4440.6	1.000
η^g_{pushPull}	0.085	0.010	0.066	0.104	0.000	0.000	4796.7	5206.9	1.000
η^g_{RPE}	0.053	0.007	0.041	0.067	0.000	0.000	5312.2	4558.6	1.000
$\eta^g_{\text{rewardRate}}$	0.027	0.004	0.021	0.034	0.000	0.000	5696.1	4690.2	1.000
$\eta^g_{\text{choiceLatency}}$	0.079	0.010	0.062	0.099	0.000	0.000	4393.0	4250.2	1.000
$l^g_{\text{intercept}}$	6.553	0.705	5.236	7.855	0.011	0.008	4201.8	4792.1	1.001
$l^g_{\text{previousOutcome}}$	7.662	0.924	5.975	9.392	0.014	0.010	4606.7	4709.4	1.001
$l^g_{\text{currentOutcome}}$	8.291	0.849	6.678	9.842	0.012	0.009	4609.0	4709.3	1.000
l^g_{pushPull}	6.805	0.787	5.424	8.322	0.012	0.009	4141.9	4923.3	1.000
l^g_{RPE}	7.198	0.897	5.550	8.891	0.013	0.009	4581.4	4873.8	1.000
$l^g_{\text{rewardRate}}$	6.691	0.959	4.986	8.510	0.013	0.009	5148.7	5351.6	1.001
$l^g_{\text{choiceLatency}}$	11.275	1.154	9.153	13.475	0.018	0.013	4160.2	4037.9	1.001
σ	0.728	0.084	0.574	0.879	0.001	0.001	13344.1	4205.4	1.000
ν	10.380	0.111	10.166	10.582	0.001	0.001	11232.3	5084.5	1.001

