# A Method for Testing Bayesian Models Using Neural Data

Bayesian models have been successful at accounting for human and animal behavior, yet to what degree they can also explain neural activity is still an open question. While decoding approaches that link neural variability to behavioral uncertainty provide some evidence, stronger tests have tried to link posterior beliefs about specific latent variables in a generative model to neural responses. On one hand, the specificity of the resulting predictions is desirable since it allows us to decide which of the infinitely many parameterizations of the task model (ideal observer) is more closely aligned with the brain's internal model. On the other hand, it is unclear under what conditions we can even expect a match of predictions and data given that current models are drastic simplifications of the rich internal model the brain uses. Furthermore, this approach so far has required strong assumptions about how probabilities are represented in neural responses. Here, we formalize and address both of these problems and derive predictions for when they can be overcome. In particular, we show how to meaningfully differentiate between Bayesian models using neural data with a weak assumption about the neural representation of probabilities, i.e. a kind of linearity that holds for a wide class of probabilistic representations including distributed distributional codes (DDCs) and neural sampling schemes. We demonstrate our method by using simulated V1 neural data to differentiate between two Bayesian models for an orientation discrimination task that are practically indistinguishable based on behavior. The first model contains orientation as an explicit variable to be inferred, while the second model assumes inference over a set of oriented gratings. Our results pave the way for strong and rigorous neural tests of Bayesian models of behavior using neural data, and give us deeper insights into how to correctly interpret neural data.

---

**Previous studies:** Testing the Bayesian Brain Hypothesis [Knill & Pouget 2004] using neural data has been of major interest in systems neuroscience and neural data has been used to argue in favor of different theories of how probabilistic beliefs are encoded in neural activity [e.g. Pouget et al. 2013, Orban et al. 2016, Ujfalussy & Orban 2022] and different structures of the generative model [e.g. Olshausen & Field 1996, Schwartz & Simoncelli 2001, Coen-Cagli et al. 2015]. Broadly, there are two principal ways in which Bayesian models have been constructed: 1) deriving the model from a task (ideal observer model) and assuming that the task-relevant variables can be decoded from neural responses [e.g. Ma et al. 2006, Tajima et al. 2016, Walker et al. 2020] or 2) deriving the model from natural input statistics [e.g. Olshausen & Field 1996, Schwartz & Simoncelli 2001] and linking posterior beliefs to neural activity making very specific assumptions like 'neural sampling' [e.g. Haefner et al. 2016, Orban et al. 2016]. However, models derived from tasks only consider a few task-relevant variables and specific stimuli (e.g., orientation). Even image-computable models derived from natural input statistics ignore many stimulus dimensions like color or binocular disparity.

**Method:** As commonly done, we propose to evaluate the quality of a Bayesian model based on its ability to predict neural responses to complex 'test' stimuli by an extrapolation from measured neural responses to simple 'baseline' stimuli. Instead of making a specific (and likely wrong) assumption about how the brain represents probabilities, we use the measured baseline responses together with the posteriors predicted by the Bayesian model for each baseline stimulus as a look-up table. We first approximate the model's (typically more complex) posterior for each test stimulus as a mixture of baseline posteriors with mixture weights $w$. Next, under the assumption that the brain's representation is linear[1], we use the mixture weights derived from matching the baseline posteriors to the test posterior in conjunction with the measured baseline responses to predict the response to each test stimulus. Importantly, the quality of the prediction will depend on the ability of the scientist's model to correctly capture the *relationship* between the brain's posteriors for the baseline stimuli and the brain's posteriors for the test stimuli. Critically, the relationships predicted by different models differ depending on the nature of their latents. This is what allows us to use neural data to differentiate between different models even if they make identical behavioral predictions.

**Demonstration of the method with known ground truth:** We describe our method using an example where we compare two Bayesian models for an orientation discrimination task: (1) a scalar orientation model derived from the orientation task, and (2) a Gaussian mixture model derived from the stimulus images (Fig.1B). These two models are practically indistinguishable using behavior, and we aim to show that we

[1]We call it 'linear' if the representation of a mixture is the weighted average of the representations of each mixture component which holds for neural sampling [Fiser et al. 2010] and DDCs [Vertes & Sahani 2019]

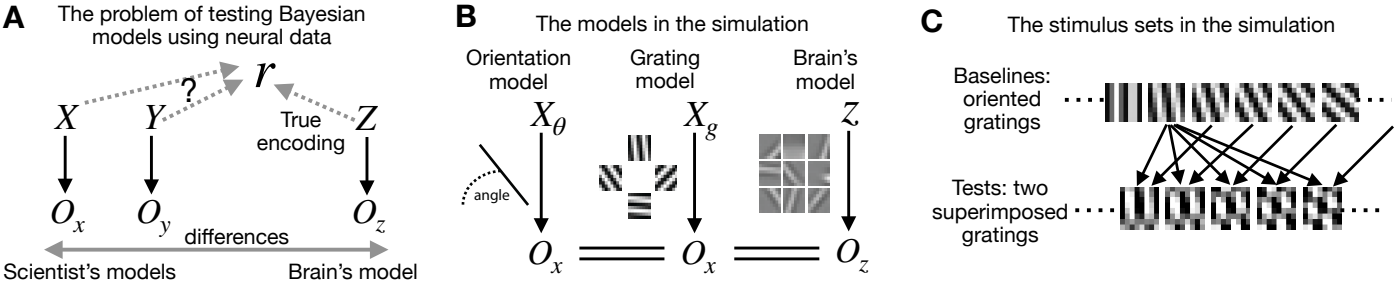


**Figure 1. A:** Solid arrows: generative directions. Dotted arrows: functional mappings. **B-D:** Double horizontal lines denotes that the observations are the same across the models. **E:** Examples of baseline and test stimuli.

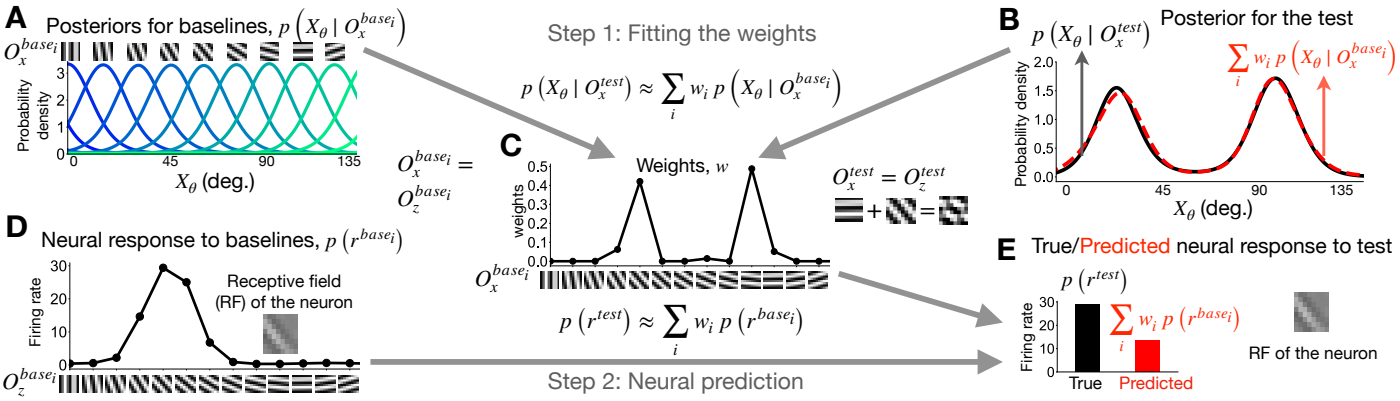


**Figure 2.** The two steps to predict the firing rate of a simulated neuron (receptive field shown in D & E, model learnt on natural images) in response to a test stimulus (shown beside, $O_x^{test}$ & $O_z^{test}$) from the orientation model.

can distinguish them with our method using neural activity. As ground truth, we assume a binary sparse coding image model [Bornschein et al. 2013] for V1, trained on natural images (Fig.1B). While the Gaussian mixture grating model appears closer to the ground truth, we note that it differs in major features: projective field, prior over latents, continuous latents rather than binary ones. For baseline stimuli we used 20 single gratings of varying orientation (Fig.1C). For test stimuli we used 80 plaids (Fig.1C). Following our method, we first computed the posteriors in the scientist's model (Fig.2A) and measured the neural responses to the baseline stimuli (Fig.2D). Then, we fit a set of weights to approximate the scientist's posterior for a single test stimulus as a mixture of baseline posteriors (step 1 in Fig.2A,B&C). Then, we used those weights to combine the measured baseline activities to predict the neural activity in response to the test stimulus (step 2 in Fig.2D,C&E). We repeated this procedure for every test stimulus. As expected, we found that the grating model provided a better match to our synthetic data (Fig.3A&B). These results imply that the grating model is a better approximation to the brain's model than the orientation model in the context of our stimuli (gratings and plaids).

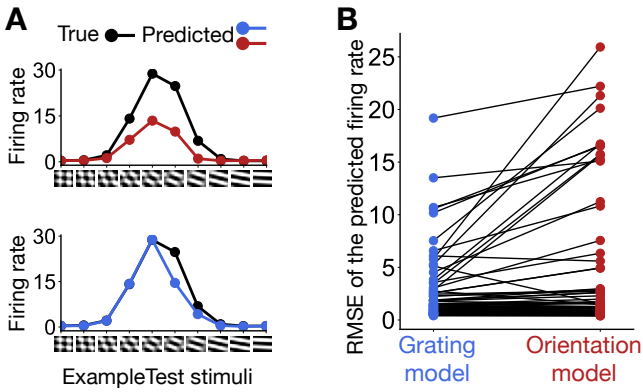


**Figure 3. A:** True and predicted firing rates for 15 example test stimuli. **B:** The average root mean squared error of the predicted firing rates (RMSE) for 80 test stimuli. Each line represents a different test stimulus.

Of course, when applied to real data, the ground truth is not known. This means that when the neural predictions from a model are accurate then the encoding was linear and the scientist's model correctly captured the brain's computations in the context of baseline and test stimuli. However, when the prediction does not agree with the data, then this means that either the model was a poor approximation to the brain's model, or that the probabilistic representation used by the brain was nonlinear. This degeneracy can in principle be broken across multiple studies: if some models correctly predict neural responses then this suggests that the representation is linear and, under the assumption that the representation is either always or never linear, that models with poorer predictions are simply poor approximations. In contrast, if no model correctly predicts neural responses then this would suggest that the neural representation of probabilities might be nonlinear which would itself be a major finding since it would falsify two of the three major classes of neural representations of probabilities.