Fitting normative neural sampling hypothesis models to neuronal response data

Summary

A prominent theory of sensory perception advocates that perception in the brain is implemented via probabilistic inference. The neural sampling hypothesis (NSH) posits that neuronal responses to a stimulus represent samples from the posterior distribution over latent world state variables (e.g., object identity) that underlie the stimulus. Existing work on NSH commonly evaluates qualitative agreement of experimental data with simple generative models of the stimulus and does not fit NSH models to experimentally observed sensory population responses.

We propose a novel formulation for NSH that allows us to directly fit NSH models to recorded stimulus-response pairs, and to formulate more flexible generative models. We formalize NSH as an equivalence between the distribution over stimulus-conditioned responses and the posterior distribution over stimulus-conditioned latent world-state variables. This enables us to fit generative models under NSH to responses and stimuli, including existing NSH models. Furthermore, we use a normalizing flow-based neural-network model and *learn* the generative model directly on image-response pairs. Our formulation allows us to directly compare NSH models to existing DNN-based encoding models of stimulus-conditioned responses. We fitted a Hoyer-Hyvärinen NSH model directly on macaque V1 responses to natural images, and compared its performance to a state-of-the-art deep neural-network system identification models. We found that the NSH model was outperformed by even a simple linear-nonlinear model. While this is somewhat expected, the size of the performance gap clearly indicates that current NSH models are too simple for real responses and motivates the development of more complex generative models. Overall, our work is an important first step toward a more quantitative evaluation of NSH models and provides a novel framework that will let us learn the generative model directly on data, paving the way for a better understanding of probabilistic computational principles that underlie perception and behavior.

Additional Detail

Theory NSH posits that the neuronal responses r elicited by the stimulus x can be interpreted as stochastic samples from the posterior distribution $p\left(z|x\right)$ computed from latent variable-based generative model over the data, giving rise to the relationship $z \longrightarrow x \longrightarrow r$ (Fig. 1A). Existing NSH models start by assuming a specific form of the generative model $p\left(x|z\right)$ and $p(z)$, giving rise to the posterior $p\left(z|x\right)$ which is then used to simulate responses r and qualitatively assess the similarity to neural data. Consequently, there has been little room for the generative model to be informed or learned from the data, and it has thus been difficult to quantitatively compare the quality of the fit of NSH models to other models, such as deep system-identification models. To enable a direct quantitative evaluation of NSH models on data, we formalize NSH as a functional equivalence between stimulus-conditioned neuronal response distribution and the posterior distribution over latents, i.e, $p_{r|x}\left(r|x\right) \stackrel{\mathrm{d}}{=} p_{z|x}\left(r|x\right)$ ($\stackrel{\mathrm{d}}{=}$ denotes equality in density functions). Consequently, we can re-express the joint distribution over the stimulus-response pair $p\left(x,r\right) = p\left(x|r\right)p\left(r\right) = p_{x|z}\left(x|r\right)p_z\left(r\right)$, allowing us to model the stimulus-response distribution in terms of the underlying generative model specified by $p_{x|z}$ and p_z. Importantly, we can now learn the generative model directly on the recorded stimulus-response pairs $\{x_i, r_i\}_{i=1}^N$ by maximizing the log-likelihood of observing the data: $\theta^* = \mathrm{argmax}_\theta \sum_i^N \left\{\log p_{x|z}\left(x_i|r_i;\theta\right) + \log p_z\left(r_i;\theta\right)\right\}$. Once trained, we can use the resulting posterior $p_{z|x}(r|x)$ to compare the NSH model's performance to that of system identification models by evaluating how well each model predicts the neural responses to test stimuli.

Simulations We simulated 10,000 pairs of image stimuli and neuronal responses from existing NSH models: (1) Hoyer & Hyvärinen model (HNH), where $p_z\left(r_i\right) = \lambda_i \exp\left(-\lambda_i r_i\right) H\left(r_i\right)$, where H is the heavyside function (2) Olshausen & Field (ONF) model where $p_z\left(r_i\right) = \frac{1}{2b_i}\exp\left(-\frac{|r_i|}{b_i}\right)$ and (3) a Gaus-

sian model (Gauss), where $p_z(r) = \mathcal{N}(r|0, \sigma_r I)$. All three models shared a common linear Gaussian likelihood function $p_{x|z}(x|r) = \mathcal{N}(x|Ar, \sigma_x I)$, where importantly, A is learned via standard ICA model with a complete basis on natural image patches. Each simulated pair consists of an 8x8 image and a vector of 64 neuronal responses. Our flexible deep normalizing flow-based neural network model (Flex) (4) uses $p_z(r) = \mathcal{N}(r|0, I) \cdot |\det \frac{\partial \mathcal{F}_\phi(r)}{\partial r}|$, and $p_{x|z}(x|r) = \mathcal{N}(x|\mu_\theta(r), \sigma^2_\theta(r))$, where $\mathcal{F}_\phi(\cdot)$, $\mu_\theta(\cdot)$ and $\sigma^2_\theta(\cdot)$ are neural networks. We fitted all models on the simulated datasets and computed exact log-likelihoods (Fig 1B). We observe that the Flex model fits neuronal responses simulated under other NSH models well, i.e., learns $p_z(r)$ and $p_{x|z}(x|r)$ and outperforms other NHS models with mismatched generative distributions. This demonstrates that our framework allows for NSH model fitting and that Flex model has the ability to flexibily capture the data distribution across widely varying generative models.

Experiments on recorded neuronal responses To demonstrate our approach on real data, we used V1 population responses to natural images (ImageNet dataset) recorded from awake Macaques using 32-channel (NeuroNexus) arrays. Each image was presented for 120 ms and we extracted spike counts from 40ms to 160ms after the image onset. We fitted three models on our dataset: (a) an HNH model (an NSH model), (b) a linear-nonlinear (Linear-SI) and (c) a DNN-based (Deep-SI) neuronal encoding model or system identification model. We first fitted HNH generative model, i.e., $p_{x|z}(x|r)$ and $p_z(r)$ and then employed variational inference to compute the posterior $p_{z|x}(r|x)$ using Gamma distribution and a 2-layer fully-connected neural network as the amortized inference function. We then compared the predicted responses from the approximate posterior from the NSH with the predictions of the linear-nonlinear and DNN-based neuronal encoding models, by computing the average correlation between predicted mean response and the average neuronal response across repeated test stimulus presentations on a set of 28 well-isolated single units (1C). We conclude that: (1) as far as we know, this is the first time that the normative theory of NSH has been quantitatively evaluated by fitting and predicting neuronal responses to arbitrary stimuli, which we believe is an important step towards testing the normative theory, and (2) that the HNH model performs quite poorly which is expected given its fairly restricted form. This motivates the development of more flexible models such as the one proposed here to be fit to real data to learn the generative model harbored by the brain and better understand probabilistic computational principles that underlie perception and behavior.

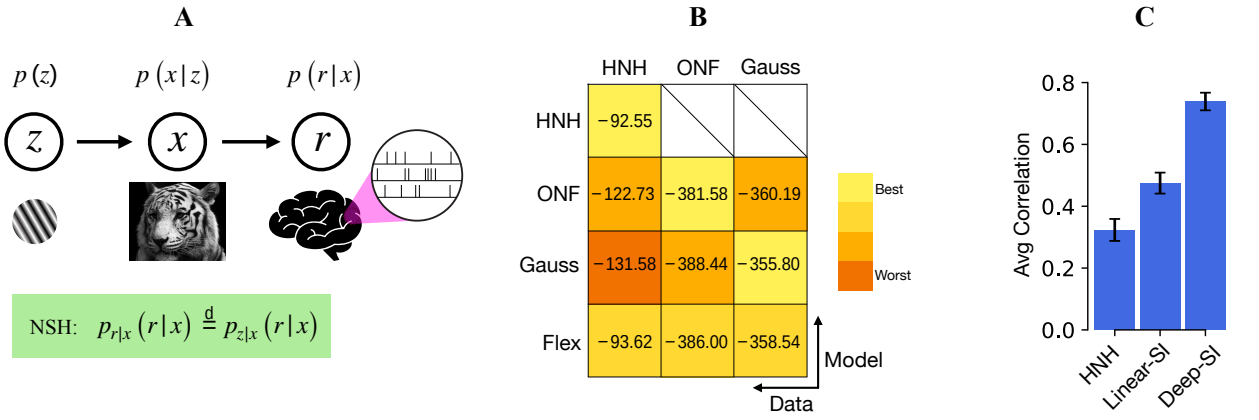

Figure 1: Fitting NSH directly to data. **A.** Our formulation of NSH as a latent variable probabilistic model, where z is the world state variable such as an oriented grating, x is the observable stimulus such as an image and r is the neuronal response, e.g. from V1, elicited by the stimulus. **B.** Log-likelihood scores (in bits, higher the better) of NSH models on simulated data. Each column corresponds to one simulated dataset and each row corresponds to log-likelihood of one NSH model on held-out test data. Our model (Flex) outperforms the fit of mismatched generative models, in that its log-likelihood is closest to ground-truth model in each column. We did not compute the scores for HNH under ONF and Gauss data since the exponential distribution in HNH has positive-only support. **C.** Average correlation of predicted means of NSH-based HNH model, and linear-nonlinear- and DNN-based neuronal response prediction models with the data mean.