Inferring the brain's internal model from sensory responses in a probabilistic inference framework

Richard D. Lange & Ralf M. Haefner*
Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA

October 19, 2016

Abstract

During perception, the brain combines information received from its senses with prior information about the outside world (von Helmholtz, 1867). The mathematical concept of probabilistic inference has previously been suggested as a framework for understanding both perception (Lee and Mumford, 2003; Knill and Pouget, 2004; Yuille and Kersten, 2006) and cognition (Gershman and Beck, 2016). Whether this framework can explain not only behavior but also the underlying neural computations has been an open question. We propose that sensory neurons' activity represents a central quantity of Bayesian computations: posterior beliefs about the outside world. As a result, sensory responses, just like the beliefs themselves, should depend both on sensory inputs and on prior information represented in other parts of the brain. We show that this dependence on internal variables induces variability in sensory responses that – in the context of a psychophysical task – is related both to the structure of that task and to the neurons' stimulus tuning. We derive analytical predictions for the correlation between different neurons' responses, and for their correlation with behavior. Furthermore, we show that key neurophysiological observations from much studied perceptual discrimination and detection experiments agree with those predictions. Our work thereby provides a normative explanation for those observations, requiring a reinterpretation of the role of correlated variability for sensory coding. Finally, the fact that sensory responses (which we observe) are a product both of external inputs (which we control) and of internal beliefs, allows us to reverse-engineer information about the subject's internal beliefs by observing sensory neurons' responses alone. Population recordings of sensory neurons in animals performing a task can therefore be used to track changes in the internal beliefs with learning and attention.

*ralf.haefner@gmail.com

Introduction

At any moment in time, the sensory information entering the brain is insufficient to give rise to our rich perception of the outside world (von Helmholtz, 1867). To compute those rich percepts from incomplete and noisy inputs, the brain has to employ prior experience about which causes are most likely responsible for a given input. In the framework of Bayesian inference, our (posterior) beliefs of these causes are computed from a combination of (prior) expectations and incoming sensory information (likelihood). While there is increasing empirical evidence that behavior approximates optimal Bayesian inference in many situations(Pouget et al., 2013; Ma and Jazayeri, 2014), it is unclear whether behavior is simply the result of task-specific heuristics or whether neural activity can also be described in a Bayesian framework. In the first part of this paper we demonstrate that sensory responses change with detection and discrimination tasks as if they do indeed represent posterior beliefs. In the second part we show how this observation can be used to infer the structure of the internal beliefs held by a particular subject about an incoming stimulus.

Results

We start by testing the hypothesis that sensory neurons encode *posterior* beliefs over latent variables in the brain's internal model (Lee and Mumford, 2003; Hoyer and Hyvärinen, 2003; Fiser et al., 2010; Haefner et al., 2016). If they do then their responses will depend both on information from the sensory periphery (likelihood), and on relevant information in the rest of the brain (prior). In a hierarchical model, the former are communicated by feedforward connections from the periphery, and the latter are relayed by feedback connections from higher-level areas (Lee and Mumford, 2003) (Figure 1a).

We represent the directly observed variable – the sensory input – by $\mathbf{E}$ while we call the variable represented by the recorded neural population under consideration $\mathbf{x}$. $\mathbf{I}$ is a high-dimensional vector representing all other internal variables in the brain that are probabilistically related to $\mathbf{x}$. For instance, when considering the responses of a population of V1 neurons, $\mathbf{E}$ is the high-dimensional image projected onto the retina, and $\mathbf{x}$ has been hypothesized to represent the presence or absence of Gabor-like features at particular retinotopic locations (Bornschein et al., 2013) or the intensity of such features (Olshausen and Field, 1996; Schwartz and Simoncelli, 2001). In higher visual areas, on the other hand, variables are likely related to the identity of objects and faces (Kersten et al., 2004). $\mathbf{I}$ represents these higher-level variables, as well as knowledge about the visual surround, task-related knowledge about the probability of upcoming stimuli, etc.

In this framework, measuring tuning curves corresponds to changing the external inputs $\mathbf{E}$ along some experimenter-defined stimulus axis s, for example visual orientation or auditory frequency. If the variable $\mathbf{x}$

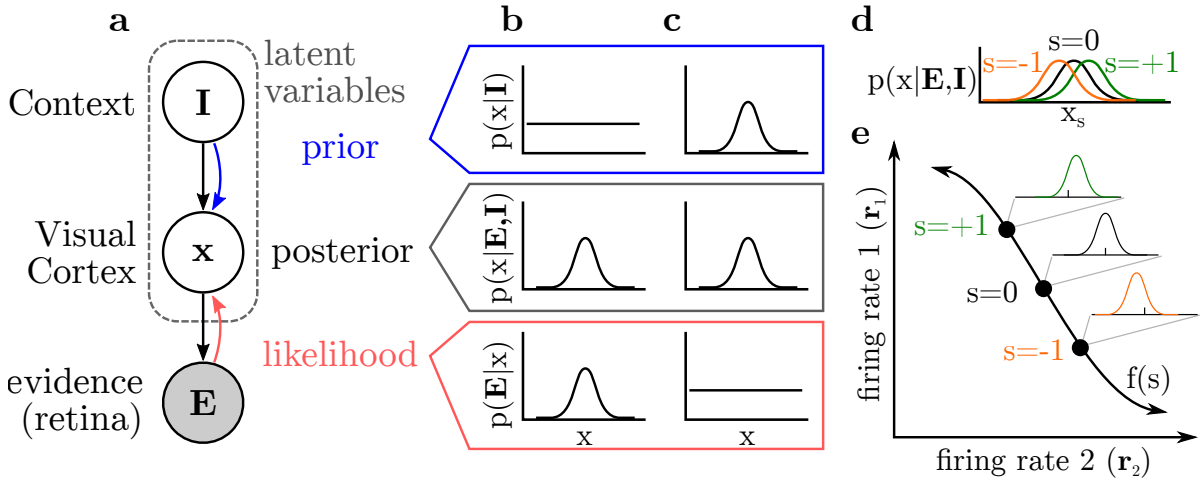

Figure 1: Illustration of 'posterior coding' for the visual system (also see Figure S1). **(a)** Neurons in visual cortex represent latent, unobserved variables in a hierarchical probabilistic model. Posterior beliefs about $\mathbf{x}$ depend both on the image on the retina, $\mathbf{E}$, and relevant higher-level components of the internal model $\mathbf{I}$. Black arrows depict the implicit generative model, while red and blue arrows indicate the actual information flow necessary to perform inference over $\mathbf{x}$ when $\mathbf{E}$ is "observed" (Lee and Mumford, 2003). **(b-c)** Top (blue): prior, middle (gray): posterior, bottom (red): likelihood. $\mathbf{x}$ may live in a high dimensional space, but only one dimension is illustrated here. In general, an informative likelihood and uninformative prior (b) can yield the same posterior as an informative prior and uninformative likelihood (c). Although we have illustrated the prior as flat in (b), in general it will not be, but will instead reflect the statistics of the natural world (e.g. differences in the occurrence of vertical and horizontal orientations). **(d)** In psychophysical tasks, the experimenter varies some parameter s to generate an image (e.g. changing the orientation of a grating pattern). This will cause changes to the distribution $p(\mathbf{x})$ if $\mathbf{x}$ depends on s. $\mathbf{x}_s$ represents a projection of the $\mathbf{x}$-space along which the posterior over $\mathbf{x}$ varies with s. **(e)** The "posterior coding" hypothesis: a neuron's firing rate, r, depends on some statistics of the posterior distribution over $\mathbf{x}$. Tuning curves $f(s)$ in this framework arise due to consistent changes in r as the posterior, $p(\mathbf{x}|\mathbf{E},\mathbf{I})$, changes as a function of s.

represented by the recorded neurons depends on s, then the likelihood $p(\mathbf{E}|\mathbf{x})$ will vary as s is varied. As a result, the posterior $p(\mathbf{x})$ will vary (Figure 1d), and in turn so will the neural responses representing it. The dependence of the mean of those responses on s gives rise to tuning curves (Figure 1e). The very same posterior, however, can also arise as the result of no information about $\mathbf{x}$ in the sensory evidence, but prior information about it in the rest of the brain encoded by $p(\mathbf{x}|\mathbf{I})$ (Figure 1c), resulting in a dependence of sensory responses on internal variables even when the external stimulus is kept constant.

Training a subject on a particular psychophysical task, on the other hand, involves learning the sensory statistics defined by the task. Prior information relevant to the task, such as which stimuli are more likely to appear, will influence the posterior, especially when the visual input is uninformative (Figure 1a). If neural responses represent posterior beliefs, then they should be the same whether this belief is due to an informative stimulus on the screen (Figure 1b), or prior expectations about this stimulus in the rest of the brain (Figure 1c). Hence there is an "equivalence" between changes in the external world and changes in internal beliefs, and formalizing this equivalence for a particular experimental context allows us to make predictions for changes in neural responses due to changing internal beliefs.

To make this idea more concrete, consider a standard discrimination paradigm in which subjects make a categorical decision about a stimulus falling into one of two categories. Over time, the subject learns to expect a stimulus from one of two categories. Let us assume for the sake of exposition that the stimulus distribution across trials is bimodal, inducing a bimodal prior in the brain (Figure 2a-c). Many experiments contain a fraction of 'zero-signal' trials in which the stimulus is uninformative about the correct decision (Britten et al., 1996; Nienborg et al., 2012), that is the likelihood is symmetric with respect to the two categories. If both categories are equally likely *a priori*, then performing exact inference in these trials will yield a symmetric posterior (Figure 2a). However, inference in the brain is at best approximate, both in terms of computation and in terms of representation. On any one trial, the actual prior used by the brain deviates from the correct one, for example due to erroneously assumed serial dependencies between the trials (Fischer and Whitney, 2014) (Figure 2c). The likelihood also varies from trial to trial due to sensory noise, e.g. in photo receptors (Figure 2b). As a result, the posterior varies from trial-to-trial even in these zero-signal trials. Given our assumption that neural responses encode posteriors, this trial-by-trial variability in the brain's posterior induces variability and covariability in the responses of sensory neurons representing that posterior. Having completely learned the task implies that the brain only expects stimuli that vary along the experimenter-defined $s-$dimension and, hence, any variability in internal beliefs, or sensory noise, will translate into variability in the posterior along the $s-$dimension (Figure 2d).

Now consider the firing responses of two neurons as the external stimulus is changed along the task-relevant axis. Their mean responses change along a line in r_1-r_2-space as a result of the changing posterior

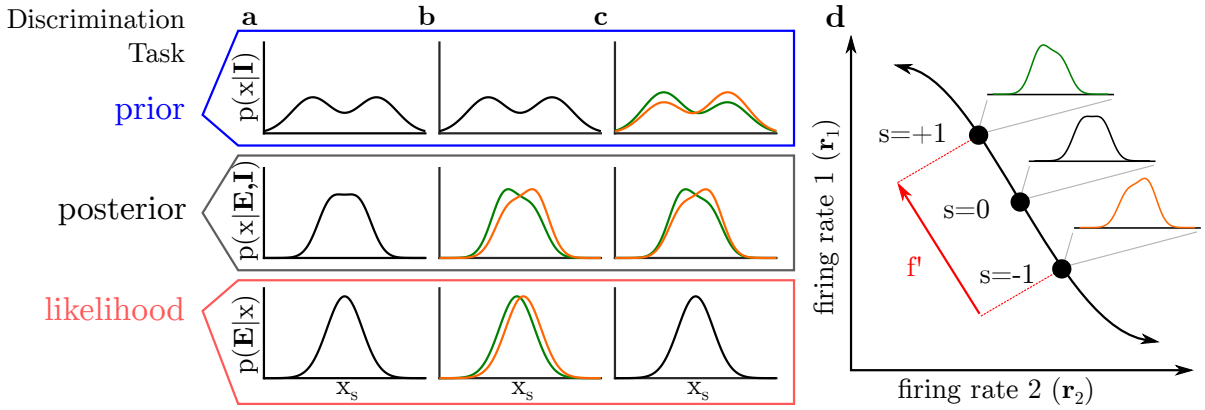

Figure 2: Posterior coding in a discrimination task. The x-axis in these plots illustrates a projection of the $\mathbf{x}$-space along which the posterior over $\mathbf{x}$ varies with s, $\mathbf{x}_s$. **(a-c)** as in Figure 1b-c. **(a)** The subject has learned to expect stimuli from either of the categories, increasing prior mass in $\mathbf{x}$-space along $\mathbf{x}_s$. "Zero-signal" trials in which the given stimulus contains no information about the correct category correspond to a likelihood with mass on either side of the decision boundary. Whether the prior is bimodal depends on the fraction of zero-signal and zero-signal trials in the experiment and is not important for our argument (see Figures S2 and S3). **(b)** Trial-by-trial changes in the likelihood, whether due to changes in the stimulus or due to noise in its representation, will shift mass in the posterior along the $\mathbf{x}_s$ direction. **(c)** Unequal prior expectations about the upcoming category at the beginning of the trial (e.g. due to serial dependencies) will shift the posterior along $\mathbf{x}_s$ similar to the changing likelihoods in (b). **(d)** Axes and $f(s)$ as in Figure 1e, with the change in mean firing rates around the decision boundary ($s = 0$) indicated by the derivative of the tuning curves, $\mathbf{f}'$. The equivalence of posteriors in (b) and (c) implies that firing rates will move along f' regardless of whether the stimulus itself changed or beliefs about it changed. We assume that $\mathbf{f}$ is measured during the task in order to account for the task-specific prior.

(Figure 2d). The dependence of the mean of the neural responses on the stimulus s is given by each neuron's tuning function, $f_i(s)$, as measured while the subject is performing the task (Figure 2d). For small changes – as are typical during threshold psychophysics – this can be linearly approximated as: $f_i(s) \approx f_i(0) + s f_i'(0)$ where $f_i'(0)$ is the derivative of neuron i's tuning function, and where we have defined s to be zero for the stimulus at the decision boundary. As a result of the equivalence noted above, the change in mean responses (corresponding to changes in the posterior) lies along the same line in r_1-r_2−space *regardless* of the particular combination of likelihood and prior giving rise to it. Under the assumption that the behavioral decision of the subject is based on the posterior belief represented by the neurons under consideration, the average posterior preceding choice 1 will have more mass favoring choice 1, and the average posterior preceding choice 2 will have more mass favoring choice 2, even if the average posterior across all trials is symmetric with respect to the decision boundary. Since the difference in the corresponding mean responses is proportional to the slope vector $\mathbf{f}'(0)$, we derive as a first prediction that $\Delta_{\text{choice}} r_i \propto f_i'(0)$ where $\Delta_{\text{choice}} r_i$ is the difference between neuron i's mean response preceding choice 1 and the mean response preceding choice 2 (Methods). This prediction relates the dependence of a neuron's response on the external stimulus to the dependence of its response on the choice *given a fixed stimulus*. In fact, when dividing both sides of this proportionality by the standard deviation of the neuron's response, σ_i, one obtains a proportionality between choice probabilities and neural sensitivities(Britten et al., 1996; Nienborg et al., 2012; Haefner et al., 2013): $\text{CP}_i \propto d_i'$ where $d_i' = f_i'(0)/\sigma_i$ is the stimulus sensitivity of neuron i (measured as d-prime). Many empirical studies have found such a relationship (reviewed in (Nienborg et al., 2012)). Interestingly, the classic feedforward-only framework makes the same prediction when the decoding weights are linear optimal (Haefner et al., 2013). Therefore, this prediction alone cannot distinguish between the classic feedforward framework and the probabilistic inference framework.

However, our probabilistic inference framework goes beyond the classic feedforward model and also predicts a component of response (co)variance that is due to the shape of the prior and trial-to-trial fluctuations in internal beliefs. Since the prior learned in the task concentrates its mass along the task-relevant axis (where all the stimuli are shown), fluctuations in the subject's internal beliefs about the stimulus will lie along that axis. As a result, these fluctuations induce the same covariance between the sensory responses as fluctuations in the stimulus itself. Using the linear approximation from above, the covariability of the responses of two neurons i and j can be expressed as $\text{cov}(r_i, r_j) = C_{ij}^0 + f_i' f_j' \text{var}(s_{\mathbf{I}})$. Here, C_{ij}^0 is the intrinsic covariability of the neural responses in the absence of task-related variability in feedforward or feedback inputs (Methods). $s_{\mathbf{I}}$ denotes the difference between the internal estimate of s and the externally presented s due to prior expectations about it and fluctuates from trial to trial. Dividing both sides by the response variability, we obtain the prediction that task-dependent noise correlations are proportional to the product

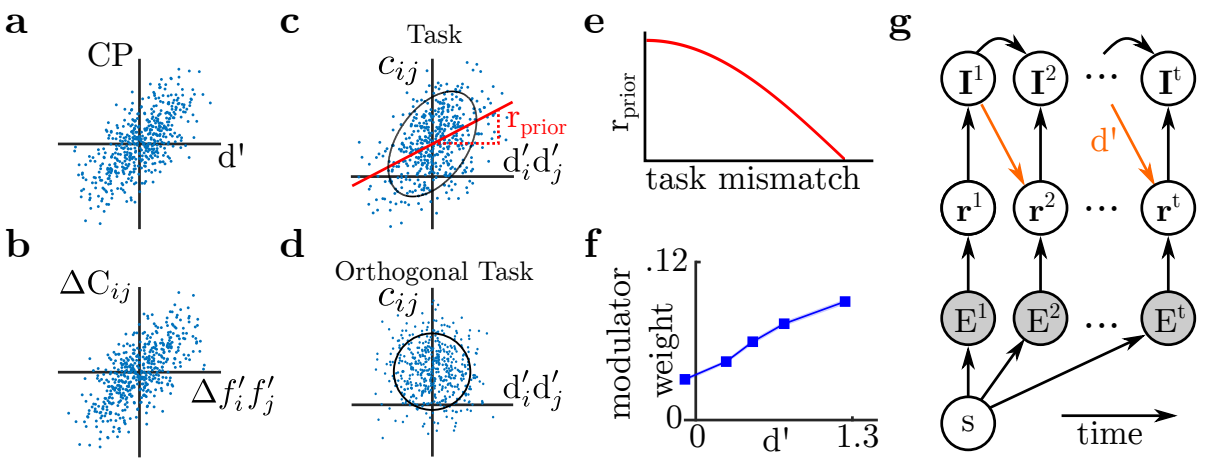

Figure 3: Predictions of the probabilistic inference framework. **(a)** first prediction, in agreement with classical feedforward encoding-decoding models with optimal linear readout: neurons' choice probabilities should be proportional to their sensitivity to the stimulus d'. **(b)** second prediction, requiring top-down signals: the difference in covariance structure between comparable tasks should be proportional to the difference in the product of tuning curve derivatives for each task. By subtracting out intrinsic covariability, this is a less noise-prone prediction than (c-e). **(c)** correlations induced by the prior should be proportional to $d'd'$. The strength of the prior should modulate the slope r_{prior} of this relationship. **(d)** the relationship in (c) should not hold for neural sensitivities d' measured with respect to other tasks' d' vectors. **(e)** summary of (c) and (d): r_{prior} should fall off with the "mismatch" between the task direction d' and the regressor direction. **(f)** Rabinowitz et al. (2015) results replotted, where it was found that the strength of top-down 'modulator' connections is linearly related to d'. **(f)** Emergence of differential correlations (Moreno-Bote et al., 2014) over the course of a trial. Here, arrows show information flow. The signal s is embedded in a sequence of noisy stimulus frames presented throughout the trial (Nienborg and Cumming, 2014; Bondy and Cumming, 2013). The developing posterior belief about the correct choice acts as a prior on subsequent responses within the same trial, inducing differential correlations. As a result, neural responses at any point throughout the trial will contain information not just about the current sensory input, but also stimuli presented earlier during the trial.

of the neural sensitivities: $c_{ij} \propto d'_i d'_j$. This predicted proportionality has two direct implications: first, performing a task should most change the noise correlation between neurons that are the most informative for this specific task, i.e. for whom d' is the largest (positive or negative). Second, this change should be positive for neurons with the same task-specific selectivity, i.e. should both increase or both decrease their activity in response to a stimulus predictive of a particular choice.

Existing studies have used two primary strategies to isolate this type of extra-sensory response modulation experimentally. First, one could take advantage of the fact that d' is defined with respect to a particular task and *vary the task* a subject is performing, predictably altering their internal model. In such studies, the difference in neural responses to zero-signal stimuli will isolate the task-dependent component for which we make predictions. At least two studies have used this approach (Bondy and Cumming, 2013; Cohen and Newsome, 2008), and found changes in the correlation structure consistent with our predictions (discussed in (Haefner et al., 2016), Methods). The second experimental approach one could take is to *statistically* isolate the top-down component of neural variability within a single task. A recent study (Rabinowitz et al., 2015) inferred the main axis along which the responses of V4 neurons varied from trial to trial in a change-detection task (Cohen and Newsome, 2009), having accounted for feedforward sources of variability. The study found that the most important modulator affecting a neuron's response is proportional to its d'_i (Figure 3b), implying correlated variability in proportion to $d'_i d'_j$. Importantly, the predicted noise correlations are task-context-specific and therefore likely depend on top-down signals. For the same reason, our prediction is different from the often observed relationship between noise correlation and tuning curve/receptive field overlap (which are task-*in*dependent) (Kanitscheider et al., 2015).

In addition to making empirically testable predictions for the influence of top-down signals on neural responses, the probabilistic inference framework provides a normative explanation for their existence. While in the classic feedforward framework decision-related signals contaminate the sensory evidence and decrease behavioral performance (Wimmer et al., 2015), here they serve the function of communicating to a sensory neuron knowledge derived from stimuli at earlier points in time, or any other relevant information from the brain's complex internal model. Consider the case of a dynamic stimulus in which the noise obscuring the fixed signal is dynamically redrawn over the course of the trial. In that case the brain's posterior belief about the signal should integrate information over all stimulus frames presented up to that moment. At any point in time, this belief over the correct choice acts as a prior that is to be combined with the likelihood representing the next stimulus frame. Communicating that prior to sensory neurons allows them to take the information provided by previous stimulus frames into account and not just rely on the current inputs (Figure 3f). Interestingly, the $d'd'$-correlations induced through top-down signals in the probabilistic inference framework have the same shape as the information-limiting correlations previously described (Moreno-Bote et al., 2014).

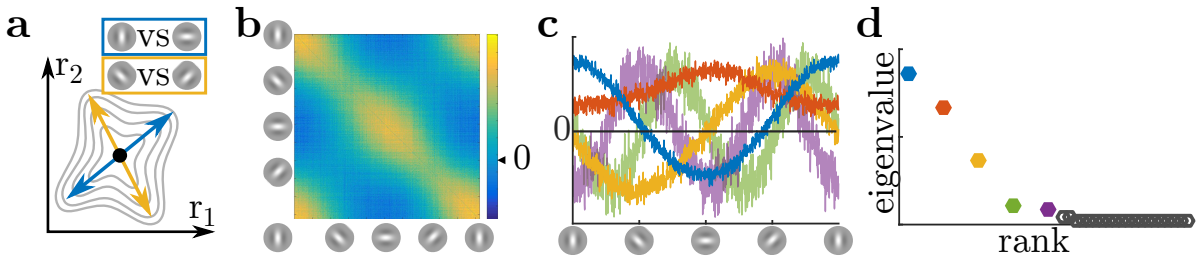

Figure 4: **(a)** Trial-to-trial fluctuations in the posterior beliefs about $\mathbf{x}$ imply trial-by-trial variability in the mean responses representing that posterior. **(b)** Correlation structure of simulated sensory responses during discrimination task. Neurons are sorted by their preferred orientation (based on (Haefner et al., 2016)). **(c)** Eigenvectors of correlation matrix (principal components) plotted as a function of neurons' preferred orientation. The blue vector corresponds to fluctuations in the belief that either a vertical or horizontal grating is present (task 1), and the yellow corresponds to fluctuations in the belief that an obliquely-oriented grating is present (task 2). See Methods for other colors. **(d)** Eigenspectrum of the correlation matrix showing five dominant subspace dimensions in responses corresponding to the five plotted eigenvectors, above.

However, unlike in the feedforward case where these correlations limit information (Moreno-Bote et al., 2014), here they are induced through feedback signals that may contain prior information about the stimulus, e.g. from earlier times in the trial (Figure 3f), or due to the subject's internal beliefs going into the trial. In general, differential correlations limit information only when they are induced by variability unrelated to the stimulus (i.e. actual noise), and not if they are induced by prior knowledge about the stimulus, e.g. due to temporal dependencies within a trial.

Reverse-engineering the internal model

From our analysis follows that variability in internal beliefs will induce correlated variability in the sensory responses of neurons related to these beliefs. Conversely, this means that the statistical structure in sensory responses can be used to infer properties of these beliefs. Importantly, this applies not just to the task-induced prior but also to priors corresponding to natural input statistics concentrated their mass in a low-dimensional subspace of $\mathbf{x}$ (Olshausen and Field, 2004). As a result, trial-by-trial variability in internal beliefs will lie within this subspace, and variability in the feedforward inputs will induce posterior variability that is larger within that subspace than in directions outside it. Hence, inferring the directions of largest variability in sensory responses can yield information about the structure of the brain's prior on $\mathbf{x}$, in particular its task-related component.

The task structure of a simple discrimination task as discussed above determines the only task-relevant belief (which of two target stimuli is the better explanation for the external inputs). However, more complicated tasks may involve inference over more than one binary variable, and therefore more than one task-relevant

belief. For instance, a task in which the target stimuli can vary from trial to trial involves inference both over the correct task and over the correct choice. Even if a pre-trial cue indicates the correct task, the cue may not be completely reliable, or the subject may not be completely certain about the cue (Cohen and Newsome, 2008; Sasaki and Uka, 2009). This uncertainty may be about the task parameters (e.g. the specific target orientation, or spatial frequency), or due to confusion with a previously learned task. If those task-related uncertainties are sufficiently large, trial-by-trial variability in the associated beliefs will lead to measurable changes in the statistical structure of sensory responses, as well as a decrease in behavioral performance. Importantly, since we know how the neural responses depend on the stimulus, we can gain an intuitive understanding of these statistical structures in terms of the stimulus.

In order to demonstrate the usefulness of this approach, we used it to infer the structure of an existing neural-sampling-based probabilistic inference model for which the ground truth is known (Haefner et al., 2016). In the simulated task, subjects had to perform a coarse orientation discrimination task either between a vertical and a horizontal grating (cardinal context), or between a -45deg and $+45$deg grating (oblique context) (Figure 4b). The subject was cued to the correct context before each trial. In the model we assumed a remaining uncertainty about the correct task context corresponding to an $80\% - 20\%$ prior. The model simulates the responses of a population of primary visual cortex neurons with oriented receptive fields. Since the relevant stimulus dimension for this task is orientation, we sorted the neurons by preferred orientation. The resulting noise correlation matrix (Haefner et al., 2016) – computed for *zero-signal trials* – has a characteristic structure in agreement with empirical observations (Figure 4c)(Bondy and Cumming, 2013). The correlation matrix has five significant eigenvalues (Figure 4d) corresponding to five eigenvectors (Figure 4c). Each of these eigenvectors represents one direction in which the neural responses vary from trial-to-trial. Knowing the stimulus selectivity of each neuron, i.e. how the response of each neuron depends on variables in the external world, allows us to interpret this eigenvector in terms of variables in the external world. For instance, the elements of the eigenvector associated with the largest eigenvalue in our simulation (blue in Figure 4c) are largest for neurons with vertically oriented receptive fields, and negative for those neurons with preferred horizontal orientation. The means that on any one trial, the population response indicates the presence of a vertical orientation in the stimulus and not a horizontal orientation, or vice versa. Recall that the presented stimulus was fixed, i.e. that this variability is due to variability in the internal beliefs, not the external stimulus. Finding such an eigenvector in empirical data therefore indicates that there is trial-to-trial variability in the subject's internal belief (represented by the rest of the brain and communicated as a prior on the sensory responses) about whether "there is a vertical grating and not a horizontal grating" or vice versa in the stimulus. Knowing the stimulus-dependence of the neurons' responses allows us to interpret the abstract statistical structure in neural covariability in terms of the stimulus space defined by

the experimenter. Equally, one can interpret the eigenvector corresponding to the third-biggest eigenvalue (yellow in Figure 4c-d) as corresponding to the belief that a +45-degree grating is being presented, but not a −45-deg grating, or vice versa. This is the correct axis for the wrong (oblique) context, indicating that the subject maintained some uncertainty about which is the correct task context across trials. Maintaining this uncertainty is the optimal strategy from the subject's perspective given their imperfect knowledge of the world. However, when compared to certain (perfect knowledge), it decreases behavioral performance on the actual task defined by the experimenter. In the probabilistic inference framework, behavioral performance is optimal when the internal model learned by the subject exactly corresponds to the experimenter-defined one. An empirical prediction, therefore, is that eigenvalues corresponding to the correct task-defined stimulus dimension will increase with learning, while eigenvalues representing other dimensions should decrease (see Methods for interpretation of other eigenvectors shown in Figure 4c). While no study has analyzed data in this framework, we know that the first and third eigenvalue must initially be increasing during task learning simply because task-dependent correlations can by definition only emerge over the course of learning. At the same time, the third eigenvalue should decrease again at some point since it represents uncertainty over the correct task context, which is presumably decreasing with learning. Furthermore, a previous study reported a decrease in average noise correlations due to learning(Gu et al., 2011). In our analysis, this corresponds to a decrease in the 2nd eigenvalue, which happens to correspond to average noise correlations since the associated eigenvector is approximately constant (see Methods).

Much research has gone into inferring latent variables that contribute to the responses of neural responses (Cunningham and Yu, 2014; Archer et al., 2014; Kobak et al., 2016). Our predictions in the context of the probabilistic inference framework suggest that at least some of these latent variables can usefully be characterized as internal beliefs. Importantly, our framework suggests that the coefficients with which each latent variable influences each of the recorded sensory neurons can be interpreted in the stimulus space using knowledge of the stimulus-dependence of each neuron's tuning function (Figure 4c).

Discussion

In sum, we have derived task-specific, neurophysiologically testable, predictions from the mathematical framework of probabilistic inference (reviewed in (Ma and Jazayeri, 2014; Pouget et al., 2013; Fiser et al., 2010; Knill and Pouget, 2004; Kersten et al., 2004)). Our assumption that sensory neurons represent posterior beliefs, not likelihoods, means that sensory responses do not just represent information about the external stimulus but also include information about the brain's expectations about this stimulus. By treating task-training as an experimenter-controlled perturbation of the brain's expectations (part of the internal model),

we have derived predictions for how neural responses should change as a result of this perturbation. This approach has allowed us to sidestep two major challenges: that the brain's full internal model is currently unknown, and that there is currently no consensus on how neural responses represent probabilities(Pouget et al., 2013; Fiser et al., 2010). While the presented theoretical predictions are novel, they are in agreement with a range of previously published empirical findings(Cohen and Newsome, 2008; Law and Gold, 2008; Gu et al., 2011; Rabinowitz et al., 2015; Bondy and Cumming, 2013).

The nature of our predictions directly addresses several debates in the field. First, they provide a rationale for the apparent 'contamination' of sensory responses by top-down decision signals(Nienborg and Cumming, 2009; Wimmer et al., 2015; Ecker et al., 2016; Rabinowitz et al., 2015). In the context of our framework, top-down signals allow sensory responses to incorporate stimulus information from earlier in the trial, not reflecting the decision per se but integrating information about the outside world(Nienborg and Cumming, 2014). Second, this dynamic feedback of feedforward stimulus information from earlier in the trial induces choice probabilities that are the result of both feedforward and feedback components (Nienborg and Cumming, 2009, 2014; Haefner et al., 2016). Third, the same process introduces correlated sensory variability that appears to be information-limiting(Moreno-Bote et al., 2014) but is not. Whether $f'f'-$covariability increases or decreases information depends on its source: if the latent variable driving it contains information about the stimulus, as in our case, it adds information; if it is due to noise (Kanitscheider et al., 2015), then it reduces it. Furthermore, the assumption that sensory responses represent posterior beliefs formalizes previous ideas and agrees with empirical findings about the top-down influence of experience and beliefs on sensory responses(von der Heydt et al., 1984; Lee and Mumford, 2003; Nienborg and Cumming, 2014). In contrast, our predictions are at odds with traditional implementations of 'predictive coding'(Rao and Ballard, 1999) which postulate that sensory responses represent a prediction error and should decrease rather than increase when bottom-up and top-down information agree. During probabilistic inference, prior and likelihood 'reinforce' each other, which can lead to either an increase or decrease in activity.

It seems plausible that only a subset of sensory neurons actually represent the output of the hypothesized probabilistic computations (posterior), while others represent information about necessary 'ingredients' (likelihood, prior), or carry out other auxiliary functions. Since our work also shows how to generate task-dependent predictions for those ingredients, it can serve as a tool for a hypothesis-driven exploration of the functional and anatomical diversity of sensory neurons.

Finally, we have shown how aspects of the low-dimensional structure in the observed covariability can be interpreted as internal beliefs that vary on a trial-by-trial basis. These variable beliefs represent the main sensory hypotheses entertained by the internal model when interpreting the sensory inputs. The detail with which these hypotheses can be recovered from neurophysiological recordings is primarily limited by

experimental techniques. Much current research is aimed developing those techniques and at extracting the latent structure in the resulting recordings. Our work suggests a way to interpret this structure, and makes predictions about how it should change with learning and attention.

Methods

Predictions

The central assumption needed to derive our predictions is that sensory responses represent posterior beliefs ('posterior coding'), such that $p(\mathbf{r})$, the response distribution of sensory neurons under consideration, is a function of the brain's posterior over the variables, $\mathbf{x}$, that those neurons represent: $p(\mathbf{r}) = \mathcal{R}\left[p(\mathbf{x})\right]$ (Figure S1). Here, try to make as little assumptions about the nature of $\mathcal{R}$ staying compatible with previous proposals from sampling-based to parametric (Hoyer and Hyvärinen, 2003; Ma et al., 2006; Fiser et al., 2010; Buesing et al., 2011; Savin and Denève, 2014; Tajima et al., 2016; Orbán et al., 2016). From trial to trial, the brain's approximation to the posterior $p(\mathbf{x}) \equiv p(\mathbf{x}|\mathbf{E}) \propto \int p(\mathbf{E}|\mathbf{x})p(\mathbf{x}|\mathbf{I})p(\mathbf{I})\mathrm{d}\mathbf{I}$ will vary since each of the terms under the integral varies due to noise and erroneously assumed serial dependencies between the trials.

We define the tuning function of neuron i as the neuron's mean response across many trials within a specific task context, corresponding to taking the integral above across all trials as $\mathbf{E}$ is changed with s: $f_i(s) \equiv \langle \mathcal{R}_i[p(\mathbf{x}|\mathbf{E}(s))]\rangle$. If the subject has completely learnt the task, their prior will correspond to the average likelihood in the task, $\int p(\mathbf{x}|\mathbf{I})p(\mathbf{I})\mathrm{d}\mathbf{I} = \int p(\mathbf{E}(s)|\mathbf{x})p(s)\mathrm{d}s$ (Berkes et al., 2011), concentrating its probability mass along the same $\mathbf{x}(s)$ line as defined by the external inputs $\mathbf{E}(s)$ As a result, prior expectations about the upcoming stimulus s, encoded by $\mathbf{I}$, shift the posterior over $\mathbf{x}$ in the same way that changes in the externally presented $\mathbf{E}(s)$ do. For sufficiently small deviations, the implied changes in neural responses can be approximated linearly as $r_i = f_i(0) + f_i'(0)s_{\mathbf{I}} + \nu_i$ where $f_i' \equiv \mathrm{d}f_i/\mathrm{d}s$ is the derivative of the tuning curve with respect to s, and $s_{\mathbf{I}}$ denotes the difference between the internal estimate of s and the externally presented s due to prior expectations about it. (For specific example illustrations see Figure S2 and S3.)c $\boldsymbol{\nu} \equiv (\nu_1, .., \nu_n)$ represents task-independent response variability due to feedforward or intrinsic sources with covariance structure $\mathbf{C}^0$. Hence, trial-by-trial variability in the brain's expectations about s, and hence, in $s_{\mathbf{I}}$, implies that response covariability is given by $\mathrm{cov}(r_i, r_j) = C^0_{ij} + f_i' f_j' \mathrm{var} s_{\mathbf{I}}$. Our prediction concerns the last term in this equation and can be tested by comparing empirical covariances in two different tasks (e.g. (Cohen and Newsome, 2008)) or by inferring common variability (e.g. (Rabinowitz et al., 2015)).

Inferring internal model

For the simple tasks considered above, complete learning implies top-down variability in only 1 direction. However, more complex tasks (e.g. those switching between different contexts), or incomplete learning (e.g. uncertainty about fixed task parameters), will generally induce variability along multiple dimensions. Making the assumption that neural responses to a fixed stimulus are locally well-approximated by a correlated Gaussian distribution, we can write the covariance between two neurons as: $\mathrm{cov}(r_i, r_j) = C^0_{ij} + \sum_{k=1}^{n} \lambda^{(k)} e_i^{(k)} e_j^{(k)}$. Each eigenvector, $\mathbf{e} = (e_1, ..e_n)$, corresponds to the change in the population response in a particular direction which, by way of the tuning functions, $f_i(s)$, can be interpreted in stimulus space (e.g. change in orientation or, or increase in contrast of a particular pattern). The eigenvalues, $\lambda^{(k)}$, quantify the magnitude of the associated trial-to-trial variability which is shared between all neurons with non-zero entries in $\mathbf{e}^{(k)}$.

The model in our proof-of-concept simulations has been described previously(Haefner et al., 2016). In brief, it performs inference by neural sampling in a linear sparse-coding model of primary visual cortex (Olshausen and Field, 1996; Hoyer and Hyvärinen, 2003; Fiser et al., 2010). The prior is derived from an orientation discrimination task with 2 contexts – oblique orientations, and cardinal orientations – that is modeled on an analog direction discrimination task (Cohen and Newsome, 2008). The responses of 1024 neurons in the lower level whose receptive fields uniformly tiled the orientation space. Each neuron's response corresponds to a sample from the posterior distribution over the variable that it represents in accordance with the neural sampling hypothesis(Hoyer and Hyvärinen, 2003; Fiser et al., 2010). We simulated zero-signal trials by presenting uniform gray images to the model. The elements of the eigenvector corresponding to the 2nd largest eigenvalue are all approximately the same indicating that variability corresponding to the associated latent variable adds response variability that does not depend on the neurons' orientations. Since the recovered eigenvectors are orthogonal to each other, the eigenvalue corresponding to a constant eigenvector determines the average correlations in the population. The eigenvectors not described in the main text correspond to stimulus-driven covariability, the eigenvectors of which are plotted in Figure S4 for comparison.

Acknowledgements

We thank the main colleagues with whom we have discussed this work and who have provided us with valuable feedback.

References

Archer, E. W., Koster, U., Pillow, J. W. and Macke, J. H. (2014). Low-dimensional models of neural population activity in sensory cortical circuits. Advances in Neural Information Processing Systems *27*, 343–351.

Berkes, P., Orban, G., Lengyel, M. and Fiser, J. (2011). Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. Science *331*, 83–87.

Bondy, A. and Cumming, B. (2013). Top down signals influence the distribution of noise correlations amongst sensory neurons. In SfN Meeting Planner (Society for Neuroscience).

Bornschein, J., Henniges, M. and Lücke, J. (2013). Are V1 simple cells optimized for visual occlusions? A comparative study. PLoS computational biology *9*, e1003062.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. and Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis Neurosci *13*, 87–100.

Buesing, L., Bill, J., Nessler, B. and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS computational biology *7*, e1002211.

Cohen, M. R. and Newsome, W. T. (2008). Context-dependent changes in functional circuitry in visual area MT. Neuron *60*, 162–73.

Cohen, M. R. and Newsome, W. T. (2009). Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. The Journal of Neuroscience *29*, 6635–48.

Cunningham, J. P. and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. Nature Neuroscience *17*.

Ecker, A. S., Denfield, G. H., Bethge, M. and Tolias, A. S. (2016). On the structure of population activity under fluctuations in attentional state. The Journal of Neuroscience *36*, 1775–1789.

Fischer, J. and Whitney, D. (2014). Serial dependence in visual perception. Nature Neuroscience *17*, 738–743.

Fiser, J., Berkes, P., Orbán, G. and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends in cognitive sciences *14*, 119–30.

Gershman, S. J. and Beck, J. M. (2016). Complex Probabilistic Inference: From Cognition to Neural Computation. In Computational Models of Brain and Behavior, (Moustafa, A., ed.), chapter Complex Pr, pp. 1–17. Wiley-Blackwell.

Gu, Y., Liu, S., Fetsch, C. R., Yang, Y., Fok, S., Sunkara, A., DeAngelis, G. C. and Angelaki, D. E. (2011). Perceptual learning reduces interneuronal correlations in macaque visual cortex. Neuron *71*, 750–761.

Haefner, R. M., Berkes, P. and Fiser, J. (2016). Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron *90*, 649–660.

Haefner, R. M., Gerwinn, S., Macke, J. H. and Bethge, M. (2013). Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nature Neuroscience *16*, 235–242.

Hoyer, P. O. and Hyvärinen, A. (2003). Interpreting neural response variability as monte carlo sampling of the posterior. Advances in neural information processing systems *17*, 293–300.

Kanitscheider, I., Coen-Cagli, R. and Pouget, A. (2015). Origin of information-limiting noise correlations. Proceedings of the National Academy of Sciences *112*, E6973–82.

Kersten, D., Mamassian, P. and Yuille, A. (2004). Object perception as Bayesian inference. Annual review of psychology *55*, 271–304.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences *27*, 712–9.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X. L., Romo, R., Uchida, N. and Machens, C. K. (2016). Demixed principal component analysis of neural population data. eLife *5*, 1–36.

Law, C.-T. and Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. Nature Neuroscience *11*, 505–513.

Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America. A, Optics, image science, and vision *20*, 1434–48.

Ma, W. J., Beck, J. M., Latham, P. E. and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nature Neuroscience *9*, 1432–8.

Ma, W. J. and Jazayeri, M. (2014). Neural Coding of Uncertainty and Probability. Annual review of neuroscience *37*, 205–220.

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P. and Pouget, A. (2014). Information-limiting correlations. Nat Neurosci *17*, 1410–1417.

Nienborg, H., Cohen, M. and Cumming, B. G. (2012). Decision-Related Activity in Sensory Neurons: Correlations Among Neurons and with Behavior. Annual Review of Neuroscience *35*, 463–483.

Nienborg, H. and Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature *459*, 89–92.

Nienborg, H. and Cumming, B. G. (2014). Decision-Related Activity in Sensory Neurons May Depend on the Columnar Architecture of Cerebral Cortex. Journal of Neuroscience *34*, 3579–3585.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. Current Opinion in Neurobiology *14*, 481–487.

Orbán, G., Berkes, P., Fiser, J. and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. Neuron *92*, 530–543.

Pouget, A., Beck, J. M., Ma, W. J. and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. Nature Reviews Neuroscience *16*, 1170–1178.

Rabinowitz, N. C., Goris, R. L., Cohen, M. and Simoncelli, E. P. (2015). Attention stabilizes the shared gain of V4 populations. eLife *4*.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience *2*, 79–87.

Sasaki, R. and Uka, T. (2009). Dynamic Readout of Behaviorally Relevant Signals from Area MT during Task Switching. Neuron *62*, 147–157.

Savin, C. and Denève, S. (2014). Spatio-temporal representations of uncertainty in spiking neural networks. Advances in Neural Information Processing Systems *27*, 1–9.

Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. Nature neuroscience *4*, 819–825.

Tajima, C. I., Tajima, S., Koida, K., Komatsu, H., Aihara, K. and Suzuki, H. (2016). Population code dynamics in categorical perception. Nature Scientific Reports *5*, 1–13.

von der Heydt, R., Peterhans, E. and Baumgartner, G. (1984). Illusory Contours and Cortical Neuron Responses. Science *224*, 1260–2.

von Helmholtz, H. (1867). Handbuch der physiologischen Optik. Verlag von Leopold Voss.

Wimmer, K., Compte, A., Roxin, A., Peixoto, D., Renart, A. and Rocha, J. D. (2015). The dynamics of sensory integration in a hierarchical network explains choice probabilities in MT. Nature Communications *6*, 1–13.

Yuille, A. and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends in cognitive sciences *10*, 301–8.

Supplemental Figures

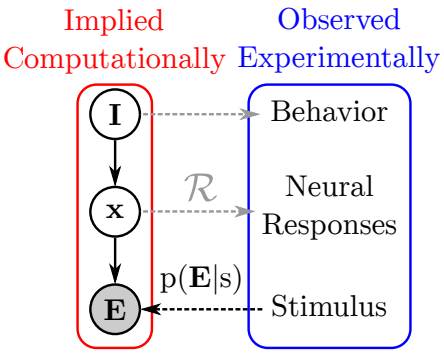

Figure S1: Schematic of posterior coding for an early sensory area that shows tuning to a stimulus parameter s. The experimenter defines $p(\mathbf{E}|s)$ (e.g. additive noise in an image), and the brain infers $p(\mathbf{x}|\mathbf{E}, \mathbf{I})$ for early sensory variables $\mathbf{x}$. Neurons representing this distribution through some encoding $\mathcal{R}$ will show 'tuning' to s by way of the posterior over $\mathbf{x}$ changing as s is changed. Our primary goal is not to infer a model relating stimulus, neural responses and behavior, but to infer the computations performed by the brain.

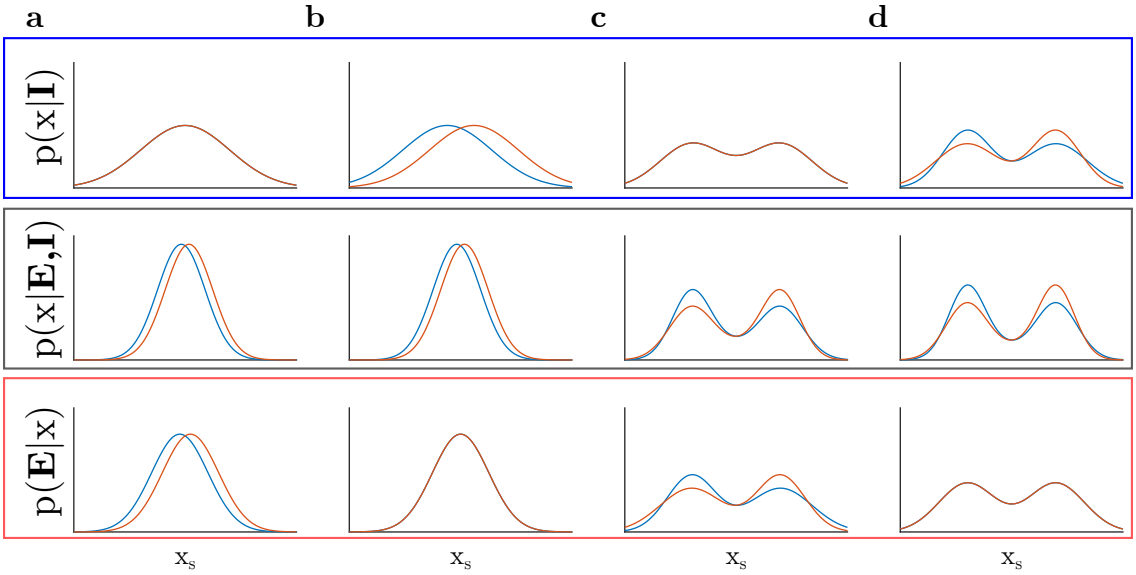

Figure S2: Equivalence of posterior for coarse and fine discrimination models. Fine discrimination (a-b) is modeled with a unimodal prior at the $s = 0$ boundary and a unimodal likelihood that shifts along x_s. 2AFC coarse discrimination (i.e. categorical decisions) (c-d) is modeled as a bimodal prior symmetric around $s = 0$ with bimodal likelihoods, where both prior expectations and evidence are modeled as a sharpening of one of the category modes. **(a)** feedforward (informative likelihood) case for fine discrimination. **(b)** feedback (informative prior) case for fine discrimination. Note equivalence of posterior with (a). **(c)** feedforward (informative likelihood) case for coarse discrimination. **(d)** feedback (informative prior) case for coarse discrimination. Note equivalence of posterior with (c).

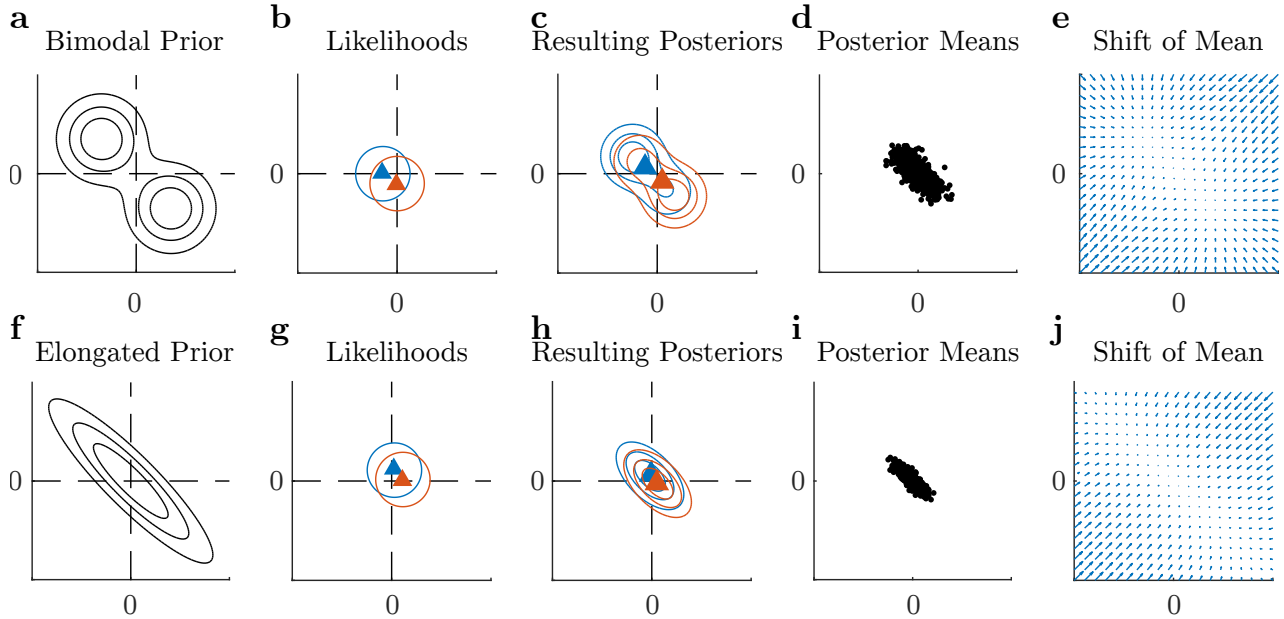

Figure S3: 2D illustration of the effect of an elongated/bimodal prior on the mean of the posterior. **a:** a bimodal prior, modeling the subject's expectations about the stimulus in $\mathbf{x}$-space during a coarse-discrimination task. **b:** on 'zero signal' trials, the stimulus is drawn from a distribution around $\mathbf{x}(s = 0)$, yielding likelihood functions that are shifted in $\mathbf{x}$-space uniformly around $\mathbf{x} = \mathbf{0}$, shown here for two example trials. **c:** the resulting posteriors for each of these trial-by-trial likelihoods are themselves bimodal. **d:** the means of these posteriors (triangles in **c**, dots here) tend to lie along the higher-probability region between the prior modes, despite an isotropic distribution of likelihood means. **e:** displacement of the mean of the likelihood to the mean of the posterior under the prior in **a**. Thus, even in the absence of serial dependencies, 'uniform' trial-to-trial variability in the stimulus yields variability in the posterior means primarily along the axis with the most mass in the prior. **f-j** same as a-e but for a unimodal but elongated prior, as might be expected in a fine discrimination task.

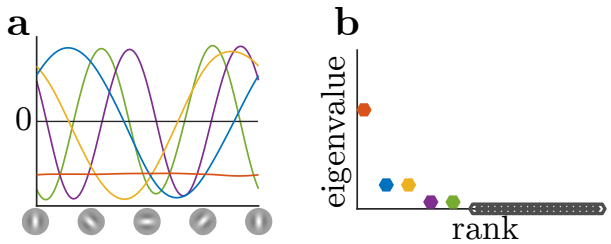

Figure S4: Principal components of model neurons due to only stimulus-driven correlations. Note that the sinusoidal eigenvectors at the same frequency have indistinguishable eigenvalues and hence form quadrature pairs, implying circular symmetry with respect to neurons' tuning. There is no more variance along the vertical-horizontal preferred orientation axis than then oblique axis.