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

December 23, 2016

Abstract

During perception, the brain combines information received from its senses with prior information about the world (von Helmholtz, 1867) – a process whose neural basis is still unclear. If sensory neurons represent posterior beliefs in a Bayesian inference process, then they, just like the beliefs themselves, must depend both on sensory inputs and on prior information. We derive predictions for how prior knowledge relates a neuron's stimulus tuning to its response covariability in a way specific to the psychophysical task performed by the brain, and for how this covariability arises from both feedforward and feedback signals. We show that our predictions are in agreement with existing measurements. Finally, we demonstrate how to use neurophysiological measurements to reverse-engineer information about the subject's internal beliefs about the structure of the task. Our results reinterpret neural covariability as signatures of Bayesian inference and provide new insights into their cause and their function.

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. 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 (von Helmholtz, 1867). Mathematically, this process can be formalized as probabilistic inference in which posterior beliefs about the outside world (our perception), are computed as the product of a likelihood function (based on sensory inputs) and prior expectations. While there is ample empirical evidence that human behavior is consistent with such probabilistic computations (reviewed in (Pouget et al., 2013; Ma and Jazayeri, 2014)), how these computations are implemented in the brain is far from clear. Our work builds on the previous observation that these Bayesian computations map naturally onto a cortical architecture in which feedforward (bottom-up) pathways communicate the information in the likelihood function about the sensory inputs, feedback (top-down) pathways communicate prior expectations, and cortical sensory neurons compute posterior beliefs about the variables that they represent (Mumford, 1992; Lee and Mumford, 2003). While it is conceptually straightforward to investigate the feedforward pathway by varying the external stimulus in a way controlled by the experimenter and recording neural responses and behavior (reviewed in (Parker and Newsome, 1998)), it is less obvious how to probe the feedback

*ralf.haefner@gmail.com

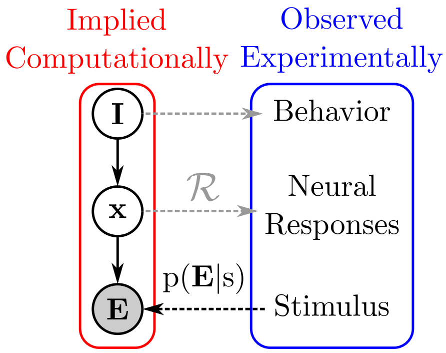

Figure 1: Illustration of the components of the probabilistic inference framework and how they relate to experimentally observed quantities. The experimenter varies the sensory inputs, **E**, (e.g. images on the retina) according to s (e.g. orientation). The brain computes $P(\mathbf{x}, \mathbf{I}|\mathbf{E})$, its beliefs about its variables of interest conditioned on those inputs. We have split the relevant internal variables into two groups: $\mathbf{x}$ which represents the variables encoded by the neurons that are being recorded from, and $\mathbf{I}$, other variables that are probabilistically related to $\mathbf{x}$. The recorded neurons are assumed to encode the brain's posterior beliefs about $\mathbf{x}$ through some representation scheme, $\mathcal{R}$. The observed behavior, which we assume to be related to $\mathbf{x}$, will therefore reflect some aspect of $\mathbf{I}$. The solid arrows represent statistical dependencies in the implicit generative model, *not* information flow (for that see Figure 5g).

influences on sensory neurons without control of internal representations (see Figure 1). There are two principal ways to overcome this challenge: correlational studies that rely on changes to internal representations over natural development (Berkes et al., 2011), and – as we describe below – causal studies that affect internal representations in an experimenter-controlled way.

In the first part of this paper, we describe a general hypothesis of 'posterior coding' that relates firing rates directly to Bayesian inference with few assumptions. From this hypothesis we derive relationships between sensory neurons' stimulus tuning and their (co-)variability while the experimenter keeps the external stimulus constant. Importantly, those relationships are specific to the task context defined by the experimenter and thereby allow interventional tests of the predictions. A comparison of our *task-specific* predictions with existing empirical studies confirms them. We further relate these predictions to the ongoing debate about the cause and interpretation of decision-related signals and response correlations in sensory cortex.

The functional implications of response variability and covariability for sensory coding has almost exclusively been analyzed and discussed in the context of classical feedforward encoding/decoding models (Zohary et al., 1994; Abbott and Dayan, 1999; Shamir and Sompolinsky, 2006; Ecker et al., 2011) (reviewed in (Kohn et al., 2016)), even when explicitly acknowledging that some of that variability may be induced by extrasensory common inputs (Ecker et al., 2014, 2016). While it enables one to compute the effect of covariability on the information contained in neural responses about the external stimulus, the classical framework makes no predictions about its structure or source.

Our results extend those in a recent numerical study (Haefner et al., 2016) based on specific assumptions about how exactly probabilities are represented in the brain, about the stimulus tuning of the sensory neurons, and about the structure of the internal model. Our results further expose the analytical relationships that drive the numerical observations in that study.

In the second part of this paper, we build on insights from the first part and demonstrate a way

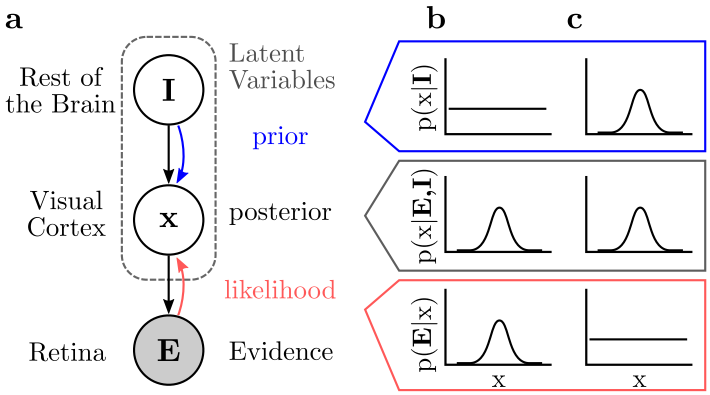

Figure 2: Illustration of hierarchical inference in the visual system. **(a)** Neurons in visual cortex represent latent, unobserved variables in a hierarchical probabilistic model. Posterior beliefs about $\mathbf{x}$ (the variables encoded by visual cortex) depend both on the image on the retina, $\mathbf{E}$, 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 $\mathbf{x}$ given $\mathbf{E}$ (Lee and Mumford, 2003). **(b-c)** Top (blue): prior; middle (gray): posterior; bottom (red): likelihood. $\mathbf{x}$ lives 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 learned statistics of the world.

to use recordings of sensory neurons' responses to infer aspects of a subject's internal, prior beliefs. In particular, we describe how to interpret them in terms of the stimulus to yield information about the subject-specific strategies in psychophysics tasks.

Results

Our central hypothesis is of 'posterior coding' – 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 2a). Many of our predictions below stem from the simple insight that any given posterior (Figure 2b-c, middle row) may arise from the combination of an uninformative prior with an informative likelihood (Figure 2b), or from the reverse (Figure 2c), implying that neurons that encode the *output* of the Bayesian computation (posteriors) will respond equivalently when they are informed by the stimulus or when they are informed by prior expectations about the stimulus.

We formalize these ideas in a hierarchical generative model (Figure 1, Figure 2a). $\mathbf{E}$ represents the directly observed variable – the sensory input, and $\mathbf{x}$ represents the variable corresponding to the recorded neural population under consideration. $\mathbf{I}$ 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-

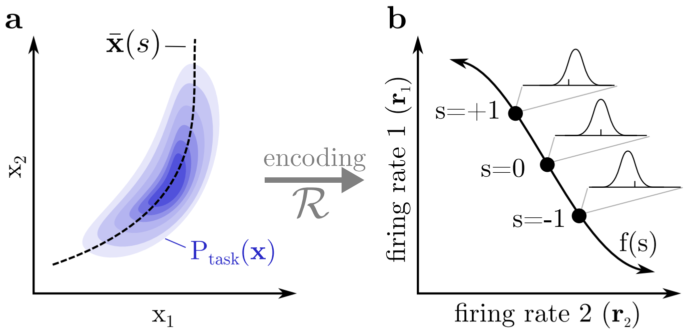

Figure 3: Illustration of 'posterior coding.' **(a)** In visual psychophysical tasks, the experimenter varies some parameter s to generate an image (e.g. changing the orientation of a grating pattern). This will cause changes to the distribution $P(\mathbf{x})$ if $\mathbf{x}$ depends on s. $\bar{\mathbf{x}}(s)$ represents the direction along which the posterior mean varies with s. **(b)** The 'posterior coding' hypothesis: a neuron's response, r, depends on some statistics of the posterior distribution over $\mathbf{x}$ through an unknown encoding $\mathcal{R}$. Tuning curves $f(s)$ in this framework reflect consistent changes in r as the posterior, $P(\mathbf{x}|\mathbf{E})$, changes as a function of s.

like features at particular retinotopic locations (Bornschein et al., 2013) or the intensity of such features (Olshausen and Field, 1996; Schwartz and Simoncelli, 2001), though the exact nature of these variables is not important for our results. In higher visual areas, 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. There is an important distinction between the *variables* in the brain's internal model (i.e. $\mathbf{x}$ and $\mathbf{I}$) and the responses of neurons that encode the *distributions* of these variables via some representation $\mathcal{R}$ (Figure 3).

In this framework, classical feedforward tuning curves (Dayan and Abbott, 2001) reflect probabilistic relationships between the variables represented by a neuron and the sensory inputs. Changes to the evidence $\mathbf{E}$ along an experimenter-defined direction s (e.g. rotating an image of a grating) affect the inferred probability of $P(\mathbf{x}|\mathbf{E})$. If the variable $\mathbf{x}$ represented by the recorded neurons is statistically dependent on s, then the likelihood $P(\mathbf{E}|\mathbf{x})$ will vary as s is varied. As a result, the posterior $P(\mathbf{x}|\mathbf{E})$ will also vary (Figure 3a), 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, denoted $\mathbf{f}(s)$ (Figure 3b, Methods). Furthermore, for small changes in s around some reference point, $s = 0$, we can linearly approximate the average neural responses: $\bar{\mathbf{r}} = \mathbf{f}(0) + \mathbf{f}'(0)s$. That is, the population response, $\mathbf{r}$, changes in the $\mathbf{f}' \equiv \mathrm{d}\mathbf{f}/\mathrm{d}s$-direction due a changing posterior belief about $\mathbf{x}$, which in turn is driven by changes in the external stimulus $\mathbf{E}(s)$ (Averbeck et al., 2006).

We now derive predictions for the effect of the *prior* on sensory responses. When a subject performs a perceptual decision-making task, the experimenter defines a distribution of stimuli $P_{\text{task}}(\mathbf{E})$ used in that task. Learning a task implies an increase in the subject's prior for $P_{\text{task}}(\mathbf{E})$ as they begin to expect stimuli drawn from this distribution. In discrimination tasks, the stimulus is varied along the experimenter-defined axis s, and subjects must make decisions about the category of s by observing $\mathbf{E}$. For example, s could be the orientation coherence of a grating (Bondy and Cumming, 2016), dot motion coherence (Britten et al., 1992), or the frequency of tactile stimulation (Romo and Salinas, 2001). The experimenter also determines the distribution of stimuli at a particular value of s (e.g. by embedding the signal in noise), as well as the distribution of s, $P(s)$. Consequently, $P_{\text{task}}(\mathbf{E}) = \int P(\mathbf{E}|s)P(s)\mathrm{d}s$. If the subject has completely learnt the task, the prior over $\mathbf{x}$ will correspond to the average likelihood in the task (Berkes et al., 2011):

$$P(\mathbf{x}) = \underbrace{\int P(\mathbf{x}|\mathbf{I})P(\mathbf{I})\mathrm{d}\mathbf{I}}_{\text{avg. effect of prior}} = \underbrace{\int P(\mathbf{x}|\mathbf{E})P_{\text{task}}(\mathbf{E})\mathrm{d}\mathbf{E}}_{\text{avg. effect of stimulus}} \tag{1}$$

Intuitively, $P(\mathbf{x})$ defines a small volume of increased probability mass in $\mathbf{x}-$space, *elongated* along a line $\bar{\mathbf{x}}(s)$ given by the dependence of the mean of $P(\mathbf{x})$ on s (Figure 3a).

For illustration, consider a stimulus distribution, $P(s)$, that is symmetric with respect to the decision-boundary, $s = 0$, i.e. training with an equivalent number of trials for each signal level for both choices. This induces a symmetric prior along $\bar{\mathbf{x}}(s)$ in the brain. Many experiments contain a fraction of 'zero-signal' trials in which the average 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 4a for an example). However, inference in the brain is at best approximate, both in terms of computation and in terms of representation. Hence on any one trial, the actual likelihood and prior used by the brain deviates from the correct one. The likelihood varies as the result of noise in the stimulus and because of noise in the afferent pathway. The prior varies if the subject erroneously assumes serial dependencies between trials (Fischer and Whitney, 2014), or if the subject develops a belief about the value of s over the course of each trial (Haefner et al., 2016).

Trial-to-trial changes in the likelihood entail trial-to-trial changes in the posterior that lie primarily along $\bar{\mathbf{x}}(s)$ since that is the line along which most of the prior mass is concentrated (Figure 4b; see also Figure S1). Furthermore, changes in the subject's internal beliefs about s – both within and across trials – will by definition cause a shift in the posterior mass along $\bar{\mathbf{x}}(s)$, this time through the prior (Figure 4c; see also Figure S2). At the same time, any changes along $\bar{\mathbf{x}}(s)$ entail changes in the neural responses along the $\mathbf{f}'-$direction – at least to a linear approximation as explained above (Figure 4d). Intuitively, this means that both variation in the stimulus and variation in the subject's beliefs about the stimulus are reflected in changes in neural responses along $\mathbf{f}'$. The consequence is increased covariability proportional to $\mathbf{f}'\mathbf{f}'^{\top}$. Dividing both sides by the response variability, task-dependent noise correlations are predicted to be proportional to the product of the neural sensitivities: $c_{ij} \propto d'_i d'_j$ (using d-prime to measure sensitivity). 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 $d'_i \equiv f'_i/\sigma_i$ has the largest magnitude. 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, and negative for those with opposite preferences. This is exactly the correlation structure observed in the empirical data recorded from primary visual cortex while a monkey was performing a coarse orientation discrimination task (Bondy and Cumming, 2016). Furthermore, it explains and generalizes numerical results generated for the specific case of a neural sampling-based representation (Haefner et al., 2016) to a wide range of representations including neural sampling and probabilistic population codes (Tajima et al., 2016; Hoyer and Hyvärinen, 2003; Haefner et al., 2016; Buesing et al., 2012; Pecevski, 2011; Savin and Denève, 2014).

We emphasize that our predictions only describe how learning a task-specific prior *changes* response correlations, and makes no predictions about correlations induced by the prior that the brain has learnt for natural images (Olshausen and Field, 2004; Berkes et al., 2011), or those that are the result of specific connectivity patterns between neurons. One strategy to experimentally

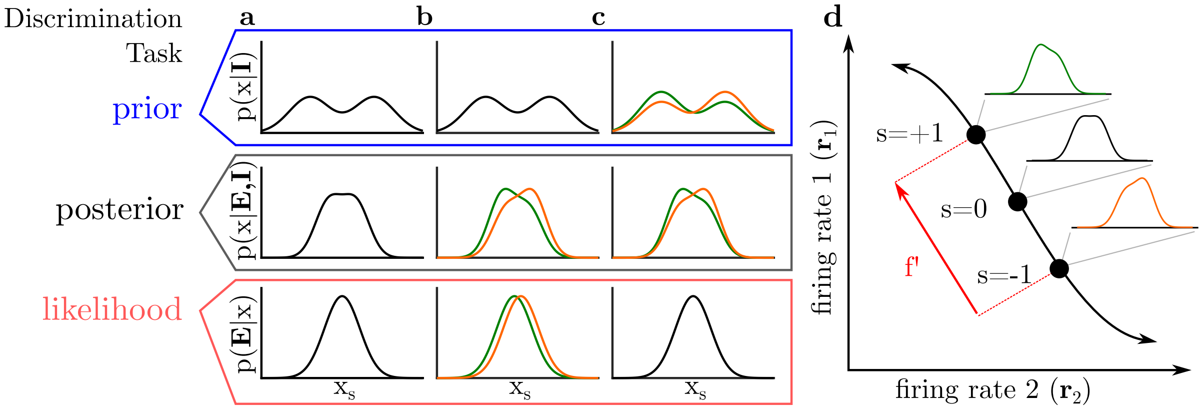

Figure 4: Posterior coding in a discrimination task, with $P(\mathbf{x})$ reduced to a single dimension along $\bar{\mathbf{x}}(s)$. **(a-c)** as in Figure 2b-c. **(a)** The subject has learned to expect stimuli from either of the categories, increasing prior mass in $\mathbf{x}$ along $\bar{\mathbf{x}}(s)$. '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 Figure S1). **(b)** Trial-to-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 $\bar{\mathbf{x}}(s)$ 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 $\bar{\mathbf{x}}(s)$ similar to the changing likelihoods in (b). **(d)** Axes and $f(s)$ as in Figure 3b, 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. $\mathbf{f}$ must be measured during the task in order to account for the task-specific prior.

test the task-specific predictions is to hold the stimulus constant while switching between two comparable tasks a subject is performing, predictably altering their task-specific prior (Methods). The difference in neural responses to zero-signal stimuli will isolate the task-dependent component to which the above predictions apply (Figure 5b). At least two existing studies have used a similar approach (Cohen and Newsome, 2008; Bondy and Cumming, 2016), and found changes in the correlation structure consistent with our predictions (discussed in (Haefner et al., 2016)). A related approach is to compare the amount of correlated variability in the current task's direction with other 'hypothetical' tasks, which is possible having measured the neurons' tuning curves in those other contexts (Figure 5c-e). A third strategy is to *statistically* isolate the top-down component of neural variability within a single task using a sufficiently powerful regression model. A recent study (Rabinowitz et al., 2015) used this type of approach to infer the primary top-down 'modulators' of V4 responses in a change-detection task (Cohen and Newsome, 2009), and found that the dominant modulator had projections to each neuron proportional to the neuron's d', implying correlated variability in the population proportional to $d'_i d'_j$ (their data replotted in Figure 5f).

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. Given the knowledge that the underlying signal has not changed, 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 previous stimulus frames 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 5f). Interestingly, the $d'd'$-correlations induced through top-down signals here have the same shape as the information-limiting correlations previously described (Moreno-Bote et al., 2014). However, unlike in the feedforward case where these correlations limit information (Moreno-Bote et al., 2014), here they are induced through feedback signals that reflect prior beliefs about the stimulus, e.g. from earlier frames in the trial (Figure 5g), or due to the subject's internal beliefs going into the trial. In general, differential correlations reduce 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.

We next ask what the implications of learning the task-specific sensory prior are for decision-related signals in sensory neurons (Parker and Newsome, 1998; Nienborg et al., 2012). 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 tuning curve slope vector $\mathbf{f}'$, it follows that $\mathrm{CTA_i} \propto \mathrm{f}'_\mathrm{i}(0)$ where $\mathrm{CTA_i}$ is the 'choice triggered average,' or difference between neuron i's mean response preceding choice 1 and its mean response preceding choice 2. This prediction relates the dependence of a neuron's response on the external stimulus at the category boundary 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 prediction for the relationship between a neuron's choice probability (CP) and its neural sensitivity: $\mathrm{CP}_i - \frac{1}{2} \propto d'_i$ (Figure 5a). Many empirical studies have found such a relationship between

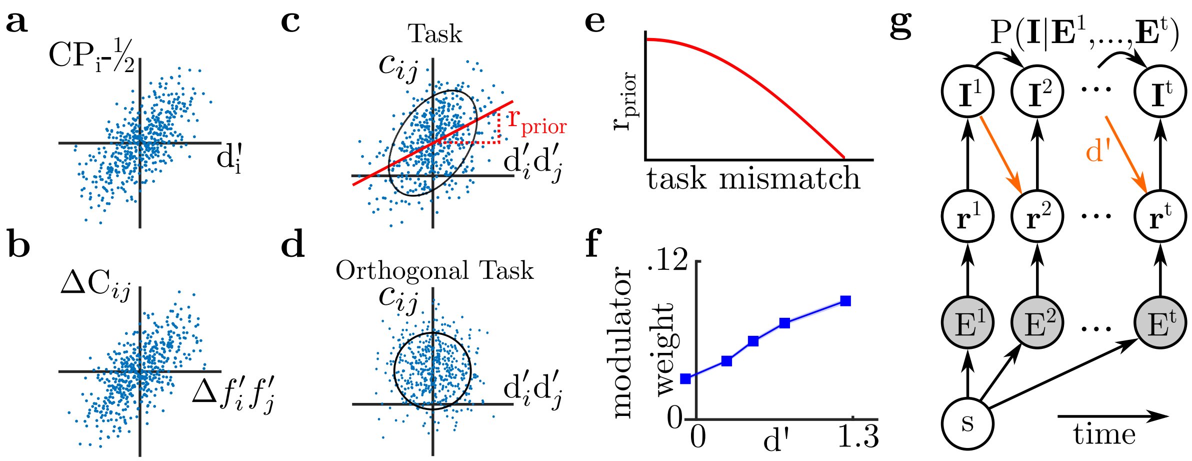

Figure 5: Predictions of the probabilistic inference framework. C denotes covariation, and c denotes correlation. **(a)** First prediction, in agreement with classical feedforward encoding-decoding models with optimal linear readout: neurons' choice probabilities should be proportional to their sensitivity 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. (Rabinowitz et al., 2015) results replotted, where it was found that the strength of top-down 'modulator' connections is linearly related to d'. **(g)** 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, 2016). 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.

a neuron's CP and its neurometric sensitivity (Nienborg et al., 2012). Interestingly, the classic feedforward-only framework makes the same prediction as the probabilistic inference framework 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.

Reverse-engineering the internal model

We have shown that internal beliefs about the stimulus induce corresponding structure in the correlated variability of sensory neurons' responses. Conversely, this means that the statistical structure in sensory responses can be used to infer properties of these beliefs.

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 variable, and therefore

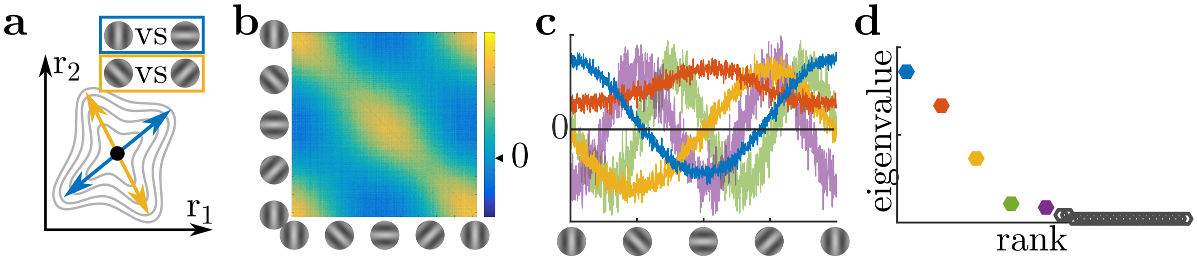

Figure 6: Inferring internal beliefs. **(a)** Trial-to-trial fluctuations in the posterior beliefs about $\mathbf{x}$ imply trial-to-trial variability in the mean responses representing that posterior. Each such 'belief' yields increased correlations in a different direction in $\mathbf{r}$. The model in b-d has uncertainty in each trial about whether the current task is vertical-horizontal orientation discrimination (task 1, blue) or oblique discrimination (task 2, yellow). **(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 colours. **(d)** Eigenspectrum of the correlation matrix showing five dominant subspace dimensions in responses corresponding to the five plotted eigenvectors in (c).

more than one task-relevant belief. For instance, a task in which the categories to be discriminated 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 learnt task. If those task-related uncertainties are sufficiently large, trial-to-trial variability in the associated beliefs will lead to measurable changes in the statistical structure of sensory responses (Figure 6a), as well as a decrease in behavioral performance.

Importantly, the probabilistic inference framework also suggests an intuitive method for interpreting top-down sources of covariability. As described above, tuning curves have a general probabilistic interpretation in terms of the statistical dependence between s and the variable(s) $\mathbf{x}$ represented by a population. As responses are assumed to encode the posterior over $\mathbf{x}$, it follows that variability in $\mathbf{x}$ – whether due top-down or bottom-up sources – may be understood in terms of the same stimulus parameters (i.e. s or $\mathbf{E}$) to which the neurons are tuned. For example, top-down modulators of neurons that are tuned to visual orientation may themselves be understood, in part, as varying prior beliefs about orientation.

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). The model was cued to the correct context before each trial, but had 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

– computed for *zero-signal trials* – has a characteristic structure in qualitative agreement with empirical observations (Figure 6b) (Bondy and Cumming, 2016). The correlation matrix has five significant eigenvalues (Figure 6d) corresponding to five eigenvectors (Figure 6c). Each of these eigenvectors (equivalent to the principal components of the population activity) represents one direction in which the trial-by-trial variability in the neural responses is larger than expected. 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 the eigenvectors in terms of variables in the external world. For instance, the elements of the eigenvector associated with the largest eigenvalue (blue in Figure 6c) are largest for neurons with vertically oriented receptive fields, and negative for those neurons with preferred horizontal orientation. 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. Recall that the external stimulus was fixed, i.e. that this variability is due to variability in the internal beliefs, not the external 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 6c-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. (see Methods for interpretation of other eigenvectors shown in Figure 6c).

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 learnt 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. 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. A previous study reported a decrease in average noise correlations due to learning (Gu et al., 2011). In our analysis, this would correspond to a decrease in the 2nd eigenvalue (average noise correlations are captured by the red eigenvector since it is approximately constant).

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 6c).

Discussion

We have derived task-specific, neurophysiologically testable, predictions within the mathematical framework of probabilistic inference (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 (Lee and Mumford, 2003). 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. Our derivation makes only minimal assumptions about the relationship between neural responses and posterior beliefs, making it applicable to a wide range of proposed neural implementations of probabilistic inference (Lee and Mumford, 2003; Tajima et al., 2016; Hoyer and Hyvärinen, 2003; Haefner et al., 2016; Buesing et al., 2012; Pecevski, 2011; Savin and Denève, 2014). Our approach has allowed us to sidestep two major challenges: that the brain's internal model is currently unknown, and that there is no consensus on how neurons 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 prior (Cohen and Newsome, 2008; Law and Gold, 2008; Gu et al., 2011; Rabinowitz et al., 2015) and new (Bondy and Cumming, 2016) empirical findings. Finally, we have used this framework to show how aspects of the low-dimensional structure in the observed covariability can be used to reverse engineer the structure of the internal beliefs that vary on a trial-to-trial basis.

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 Roelfsema, 2015). 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, 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). It also relates to a large literature on association learning and visual imagery (reviewed in (Albright, 2012)). In particular, the idea of 'perceptual equivalence' (Finke, 1989) reflects our starting point that the very same posterior belief (and hence the same percept) can be the result of different combinations of sensory inputs and prior expectations. In a discrimination task, for instance, there are three distinct associations inducing correlations. First, showing the same input many times induces positive correlations between sensory neurons responding to the same input. Second, presenting only one of two possible inputs induces negative correlations between neurons responding to different inputs. Third, keeping the input constant within a trial induces positive auto-correlations.

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 (Pecevski, 2011). 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.

In deriving the predictions for changes in the task-specific correlations we have implicitly assumed that the feedforward encoding of sensory information, i.e. the likelihood $P(\mathbf{E}|\mathbf{x})$, remains unchanged between the compared conditions. This is well-justified for lower sensory areas in adult subjects (Hensch, 2005), or when task contexts are switched on a trial-by-trial basis (Cohen and Newsome, 2008). However, it is not necessarily true for higher cortices (Li and DiCarlo, 2008), especially when conditions are compared separated by long periods of task (re)training. In those cases, changing sensory statistics may lead to changes in the feedforward encoding, and hence the nature of the represented variable $\mathbf{x}$ (Ganguli and Simoncelli, 2014; Wei and Stocker, 2015).

Previous work has demonstrated the possibility of using *behavioral* judgements to infer the shape of a subject's prior (Houlsby et al., 2013). Our results are complementary to behavioral methods, but have the advantage that the amount of information that can be collected in neurophysiology experiments far exceeds that in psychophysical studies.

The detail with which the internal beliefs can be recovered from the statistical structure in 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. For illustration, we used principal component analysis in Figure 6, implicitly assuming linear effects of varying beliefs on the sensory population (Methods) and orthogonality of their directions. With nonlinear effects of the prior and in order to infer non-orthogonal causes, more sophisticated tools will be required to infer latent structure in sensory responses (Cunningham and Yu, 2014). Importantly, our work suggests a way to interpret this structure, and makes predictions about how it should change with learning and attention.

Methods

Definition of tuning curves

Most generally, one can think of the process of encoding the posterior as a functional $\mathcal{R}$ that maps from a distribution over $\mathbf{x}$ to a distribution of neural responses: $P(\mathbf{r}) = \mathcal{R}\left[P(\mathbf{x})\right]$ (Figure 1). We require that $\bar{\mathcal{R}}$ is smooth as $\bar{\mathbf{x}}$ changes (where $\bar{y}$ denotes the mean of y across trials), which allows us to use linear approximations of tuning functions. We define the tuning function of neuron i as the neuron's mean response across trials within a specific task context as $\mathbf{E}$ is changed with s:

$$f_i(s) \equiv \bar{\mathcal{R}}_i[P(\mathbf{x}|s)] \tag{2}$$

where $P(\mathbf{x}|s) \equiv \int P(x|\mathbf{E},\mathbf{I})P(\mathbf{E}|s)P(\mathbf{I})\,\mathrm{d}\mathbf{E}\,\mathrm{d}\mathbf{I}$.

Prediction for the difference between comparable tasks

The magnitude of task-dependent response variability depends on the magnitude of the trial-to-trial changes in beliefs about s, and on strength and shape of the learned prior along $\bar{\mathbf{x}}(s)$. Two arbitrary tasks will in general differ in these aspects as well as in the intrinsic covariance of responses to the zero-signal stimulus. We call two tasks 'comparable' when they agree in both the magnitude of the prior and the intrinsic response covariance, as can reasonably be expected, for instance, in rotationally symmetric situations where all that changes between the tasks is the angle (Bondy and Cumming, 2016) or direction (Cohen and Newsome, 2008) of the discrimination boundary while the zero-signal stimulus stays the same. In that case the strength of the respective $f'f'-$component can be assumed to be the same and hence, the intrinsic covariability can be subtract out:

$$C_{ij}^{(1)} - C_{ij}^{(2)} \propto f_i^{(1)\prime} f_j^{(1)\prime} - f_i^{(2)\prime} f_j^{(2)\prime}$$

where superscripts denote the task. That is, $f_i^{(1)\prime}$ denotes the slope of neuron i's tuning curve with respect to the discrimination axis in task 1 measured at $s = 0$. Note that two fine discrimination tasks (e.g. orientation discrimination around the vertical and the horizontal axes, respectively) are not necessarily 'comparable' since the two tasks differ in their zero-signal stimulus (a vertical and a horizontal grating, respectively), which may yield different intrinsic covariability.

Inferring internal model

Complex tasks (e.g. those switching between different contexts), or incomplete learning (e.g. uncertainty about fixed task parameters), will often induce variability in multiple internal beliefs about the stimulus. Assuming that this variability is independent between the beliefs, we can write the observed covariance between two neurons as $\mathrm{cov}(r_i, r_j) = C_{ij}^0 + \sum_{k=1}^n \lambda^{(k)} b_i^{(k)} b_j^{(k)}$. Here, each vector $\mathbf{b}^{(k)} = (b_1^{(k)}, b_2^{(k)}, ..)$ corresponds to the change in the population response corresponding to a change in internal belief k. The coefficients $\lambda^{(k)}$ correspond to the variance of the trial-to-trial variability in belief k, and C_{ij}^0 represents the intrinsic covariance.

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). We simulated the responses of 1024 V1 neurons whose receptive fields uniformly tiled the orientation space. Each neuron's response corresponds to a sample from the posterior distribution over the intensity of its receptive field in the input image. We simulated zero-signal trials by presenting white noise 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, plotted in Figure S3 for comparison.

Acknowledgements

We thank the many colleagues with whom we have discussed this work and who have provided us with valuable feedback, in particular Alex Ecker, Matthias Bethge, Hendrikje Nienborg, Jakob Macke, Adrian Bondy, and Bruce Cumming.

References

Abbott, L. F. and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. Neural computation *11*, 91–101.

Albright, T. D. (2012). On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception. Neuron *74*, 227–245.

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.

Averbeck, B. B., Latham, P. E. and Pouget, A. (2006). Neural correlations, population coding and computation. Nature Reviews Neuroscience *7*, 358–66.

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. G. and Cumming, B. G. (2016). Feedback Dynamics Determine the Structure of Spike-Count Correlation in Visual Cortex. bioRxiv .

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

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.

Britten, K. H., Shadlen, M. N., Newsome, W. T. and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. The Journal of Neuroscience *12*, 4745–4765.

Buesing, L., Macke, J. H. and Sahani, M. (2012). Learning stable, regularised latent models of neural population dynamics. Network: Computation in Neural Systems *23*, 1–9.

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*, 1500–1509.

Dayan, P. and Abbott, L. F. (2001). Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press, London.

Ecker, A. S., Berens, P., Cotton, R. J., Subramaniyan, M., Denfield, G. H., Cadwell, C. R., Smirnakis, S. M., Bethge, M. and Tolias, A. S. (2014). State dependence of noise correlations in macaque primary visual cortex. Neuron *82*, 235–248.

Ecker, A. S., Berens, P., Tolias, A. S. and Bethge, M. (2011). The effect of noise correlations in populations of diversely tuned neurons. J Neurosci *31*, 14272–14283.

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.

Finke, R. A. (1989). Principles of mental imagery. MIT Press.

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.

Ganguli, D. and Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. Neural computation , 2103–2134.

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.

Hensch, T. K. (2005). Critical period plasticity in local cortical circuits. Nature reviews. Neuroscience *6*, 877–88.

Houlsby, N. M. T., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M. and Lengyel, M. (2013). Cognitive Tomography Reveals Complex, Task-Independent Mental Representations. Current Biology *23*, 2169–2175.

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.

Kohn, A., Coen-cagli, R., Kanitscheider, I. and Pouget, A. (2016). Correlations and Neuronal Population Information. Annual Review of Neuroscience *39*, 237–256.

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 *20*, 1434–1448.

Li, N. and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science (New York, N.Y.) *321*, 1502–1507.

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. Nature Neuroscience *17*, 1410–1417.

Mumford, D. (1992). On the computational architecture of the neocortex. Biological cybernetics *251*, 241–251.

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.

Nienborg, H. and Roelfsema, P. R. (2015). Belief states as a framework to explain extra-retinal influences in visual cortex. Current opinion in neurobiology *32*, 45–52.

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.

Parker, A. J. and Newsome, W. T. (1998). Sense and the single neuron: probing the physiology of perception. Annu Rev Neurosci *21*, 227–277.

Pecevski, D. (2011). Probabilistic inferences general graphical models through sampling in stochastic networks of spiking neurons. PLOS Computational Biology *7*.

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*.

Romo, R. and Salinas, E. (2001). Touch and go: decision-making mechanisms in somatosensation. Annual review of neuroscience *24*, 107–37.

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.

Shamir, M. and Sompolinsky, H. (2006). Implications of neuronal diversity on population coding. Neural computation *18*, 1951–86.

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.

Wei, X.-X. and Stocker, A. a. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. Nature neuroscience *18*, 1509–17.

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.

Zohary, E., Shadlen, M. N. and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. Nature *370*, 140–143.

522 ## Supplemental Figures

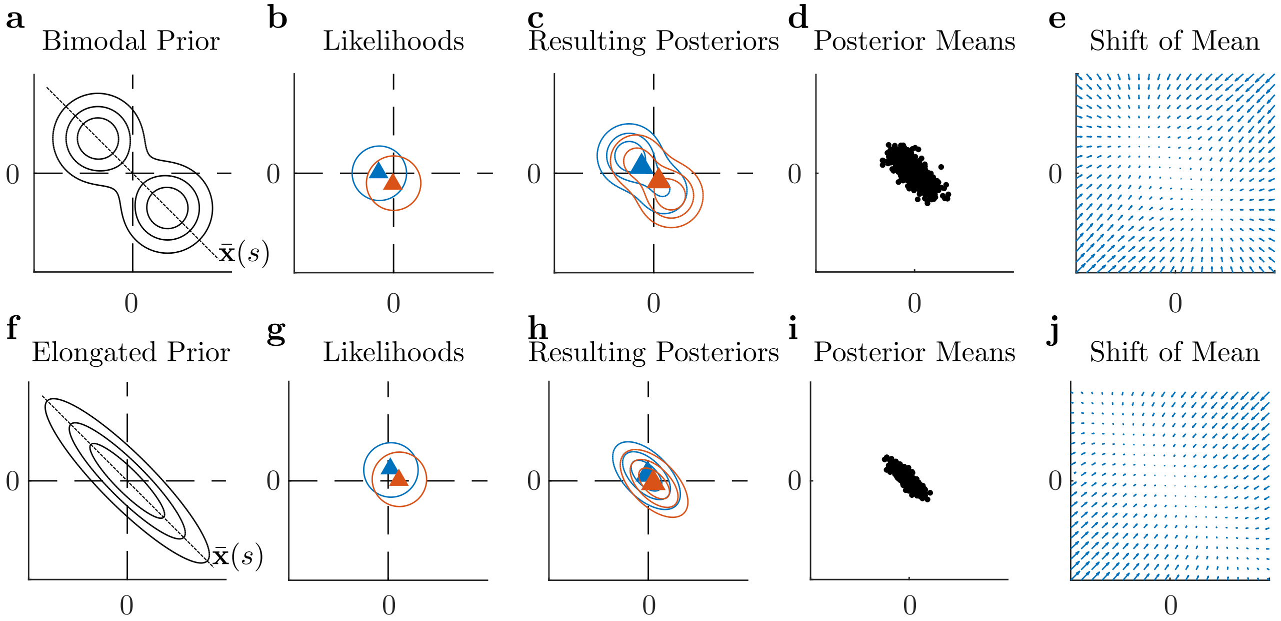

Figure S1: 2D simulation of the effect of a prior that is elongated/bimodal along $\bar{\mathbf{x}}(s)$ on the mean of the posterior. **a:** A bimodal prior, modeling the subject's expectations about the stimulus (in $\mathbf{x}$) during a coarse-discrimination task. **b:** On 'zero signal' trials, the stimulus is drawn from a distribution around $\mathbf{x}(s = 0)$, yielding likelihood functions that are shifted uniformly around $\mathbf{x} = \mathbf{0}$, shown here for two example trials. **c:** The resulting posteriors for each of these 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.

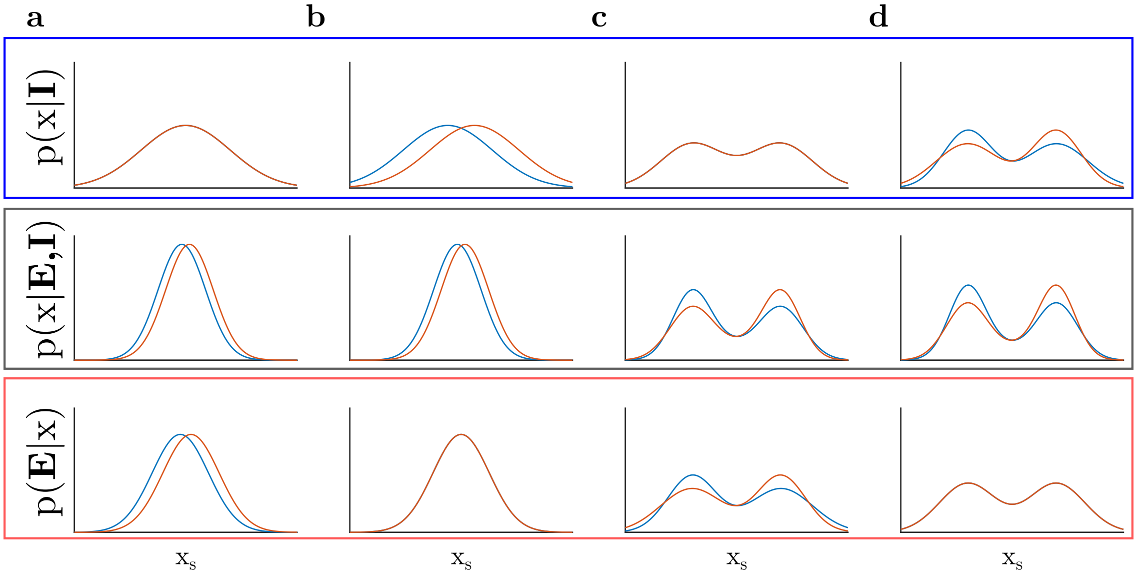

Figure S2: Equivalence of posterior for coarse and fine discrimination models. Fine discrimination (a-b) is modeled with a unimodal prior at the $s = 0$ boundary and a unimodal likelihood that shifts along $\bar{\mathbf{x}}(s)$. 2AFC coarse discrimination (i.e. categorical decisions) (c-d) is modeled 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. In each of these models, changes to the stimulus along $\bar{\mathbf{x}}(s)$ yields identical changes to the posterior as arise from changes in the prior. **(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).

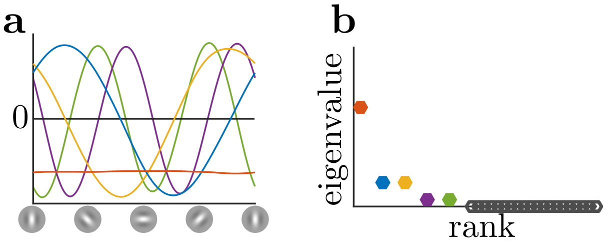

Figure S3: Principal components of model neurons due to only stimulus-driven correlations. Note that the sinusoidal eigenvectors at the same frequency have indistinguishable eigenvalues and hence form quadrature pairs, implying circular symmetry with respect to neurons' tuning. There is no more variance along the vertical-horizontal preferred orientation axis than then oblique axis.