bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Review: Characterizing and interpreting the influence of internal variables on sensory activity

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

July 5, 2017

Highlights

- Re-interpretation of neural correlations in terms of internal variables...
- ...can clarify whether they limit or enhance information
- Influence of internal variables can be captured by interpretable 'tuning functions'
- Estimation of both internal variables and tuning possible from population recordings

Abstract

The concept of a tuning curve has been central for our understanding of how the responses of cortical neurons depend on external stimuli. Here, we describe how the influence of unobserved internal variables on sensory responses, in particular correlated neural variability, can be understood in a similar framework. We suggest that this will lead to deeper insights into the relationship between stimulus, sensory responses, and behavior. We review related recent work and discuss its implication for distinguishing feedforward from feedback influences on sensory responses, and for the information contained in those responses.

Introduction

A large part of cortical function can be characterized as transforming sensory inputs to behavioral outputs. While this transformation is conceptually unidirectional, the anatomical structures implementing it are largely bidirectional [1]. In this review, we concentrate on discussing progress in understanding 'top-down,' or feedback signals (FB) that influence the responses of sensory neurons. While the correlated variability of neural responses has traditionally been interpreted primarily as noise and in terms of local network connectivity or limited input information (reviewed in [2]), it may also arise from variability in unobserved internal variables projecting back to sensory areas [3, 4••, 5••]. We summarize and compare experimental techniques for inferring those internal variables from sensory population recordings, and show that fluctuating internal variables offer simple

*{rlange,rhaefne2}@ur.rochester.edu

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

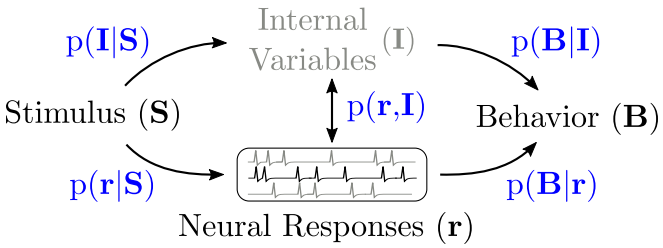

Figure 1: A computational-level description of the brain (upper path) may invoke abstract internal variables (**I**) that govern behavior (**B**) and are influenced by stimuli (**S**). On a neurophysiological level, we seek a similar description in terms of the responses (**r**) of populations of neurons (lower path) [7]. It is often useful to mix levels of abstraction, for example modeling the relationship between attention (an abstract state, or internal variable) and neural responses. Studying these relationships is limited by what quantities may be directly observed; **I** is not directly observable, and **r** is only partially observable (indicated in gray). Experiments in which an estimate of **I** is available from an observable correlate (**S** or **B**) lend themselves to *conditional* approaches, estimating p(**r**|**I**). *Unsupervised* approaches model the relationship between **r** and **I**, estimating both the instantaneous value of an internal variable and its influence on responses given only observations of neural responses p(**r**).

and intuitive accounts of many empirical results. While our focus here is on empirical studies of early visual areas, we expect the discussed techniques and computational framework to also be useful in understanding neural responses in other sensory or motor areas [6•].

The brain's representation of sensory inputs and transformation into behavioral outputs can be understood on at least two levels. On the first one, we would like to understand how a stimulus (**S**) influences neural responses (**r**), and how these responses influence behavior (**B**). On a more abstract level, one can try to understand how a stimulus affects abstract internal variables (**I**), and how these variables are linked to behavior (Figure 1). While these two levels are clearly related, the representation of abstract internal variables is generally unclear, and it is in principle only possible to directly observe the corresponding neural responses; internal variables are only accessible by their correlation with one or more of the observable quantities, **S**, **B**, and (a subset of) **r** (Figure 1).

Mathematically, relationships between these variables may be framed either as conditional or as joint probability distributions. Conditional probability distributions are commonly used when one of the variables is known, or if a causal role of this variable is hypothesized, for example the stimulus-conditioned response distribution, the mean of which defines the tuning function with respect to that stimulus. Thinking in terms of joint probability distributions, on the other hand, is useful to characterize a wider range of relationships that are not necessarily causal, and where neither variables needs to be under the control of the experimenter, or even be observed. In the following, we review existing work in the context of these two approaches and relate them to attempts to distinguish feedforward from feedback signals in the brain. In particular, we describe how the concept of a 'tuning curve' can be extended to capture the relationship between neural responses and internal variables (Box 1), and further show that this yields simple and intuitive descriptions of correlated noise in sensory areas (Box 2).

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Inferring internal variables and their influence on neural responses

Since a subject's internal variables are not observable directly, the traditional approach to measuring their impact on neural responses has been to exploit their relationship to observed quantities like stimulus and behavior. For example, training an animal in an attention task induces a relationship between an internal variable ('attentional state') and an external sensory signal ('attention cue') [3], where improved performance on a task is taken as an indicator for successful manipulation of attentional state [8••].

The first limitation of this approach is the fact that there usually remains significant uncertainty about the relationship between observables and internal variables. For instance, even though an attention cue may only have two states, the animal's true attentional state is likely non-binary and more variable. This limitation can be overcome by bootstrapping from a binary cue to a continuous internal representation and by making additional assumptions about the relationship between the internal variable and the measured neural responses. Assuming linearity, for instance, allows one to define an 'axis' connecting the average neural responses conditioned on the two possible cues, along which neural responses will vary as the internal variable varies (Box 1). Projecting the neural activity on a single trial onto this axis, allows one to estimate an intermediate value of the internal variable [8••]. Cohen and Maunsell [8••, 9, 10, 11] indeed showed that the estimates obtained using this approach in a change-detection task while recording from neurons in area V4, predicted behavioral performance (accuracy) on single trials. An analogous approach can be taken to infer a 'choice axis' [12, 13] which is closely related to measuring choice probabilities across a population of sensory neurons [14, 15]. The projection of the population activity on a single trial onto this choice axis yields an estimate of a decision-related variable. If correct, this projection will correlate with behavioral measures of confidence or reaction time. Such an approach to inferring the state of an internal variable might be called *conditional* since it relies on an observable correlate of the variable of interest (Fig. 1).

The second limitation, which the conditional approach cannot overcome, is the fact that internal variables are surely richer than the behavioral output in perceptual decision-making tasks (especially when behavior is limited, as in discrimination tasks), or the number of trainable cues. For instance, 'attention' is likely to be a higher-dimensional variable incorporating different location(s), feature(s), etc. to which the animal is attending [9, 16•]. Equally, the brain is likely to form variable internal beliefs beyond those related to the behavioral report in a particular task. One way to overcome this limitation is to exploit the statistical structure induced in population responses by variability in those internal variables. Fluctuations in an (unobserved) internal variable increases (observable) response variability in a direction that can be thought of as the neural responses' 'tuning' to that variable (Box 1, Figure 2). Different internal variables all leave their own signatures that superpose (Figure 2b, Boxes 1 + 2). By fitting the *joint* probability distribution $p(\mathbf{r}, \mathbf{I})$, it is possible to infer both the internal ('latent') variables as well as their influence on the responses from observations of the responses alone alone, in an unsupervised way [6•, 17, 18]. A well-known example of this approach is Factor Analysis, with each 'factor' corresponding to an internal variable in this context. The 'factor loadings', i.e. each factor's influence of neural responses, corresponds to an 'axis' recovered in the conditional approach above, and can in turn be used to infer estimates of the corresponding internal variable over time [4••], or on a trial-by-trial basis [5••]. 'Demixed PCA' [19] presents a promising mixed approach, incorporating hypotheses about multiple internal variables into jointly fitting $p(\mathbf{r}, \mathbf{I})$. Knowledge of the inferred variables' values and their corresponding 'tuning curves' may allow a functional interpretation of them, e.g. as the

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

influence of anesthesia [4••], attentional state [8••, 20, 21, 22], motor activity [23], or in terms of their sensory representation, e.g. as beliefs about the outside world [15, 24, 25••, 26, 27, 28]. Even though this approach requires large neural data sets (but not prohibitively large [29]), we believe that it will be an extremely productive one (see Box 2 for an illustration).

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Box 1: How fluctuations in internal variables induces correlated variability in neural responses

A neuron's response can be modeled as a function of both *internal* and *external* variables:

$$r = f(\underbrace{\mathrm{S}_1, \mathrm{S}_2, \ldots}_{\text{external}}, \underbrace{\mathrm{I}_1, \mathrm{I}_2, \ldots}_{\text{internal}}) + \text{noise}. \tag{1}$$

Locally, this *tuning function* can be approximated linearly by its slope with respect to each internal and external variable. Any variability in its arguments will induce variability in the neural response [3, 30]. Across a neural population, $\mathbf{r} = (r_1, r_2, ..)$, this variability will be correlated with the following response covariance†:

$$\mathbf{C} \approx \mathbf{C}^0 + \sum_i \underbrace{\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_i}\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_i}^\top \mathrm{var}(\mathrm{I}_i)}_{\text{'signature' of } \mathrm{I}_i} + \underbrace{\sum_{i \neq k} \left(\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_i}\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_k}^\top + \frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_k}\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_i}^\top \right) \mathrm{cov}(\mathrm{I}_i, \mathrm{I}_k)}_{\text{interaction terms}} \tag{2}$$

where $\mathbf{C}^0$ is 'intrinsic' covariance (i.e. the covariance of the 'noise' term that is not attributed to $\mathbf{S}$ or $\mathbf{I}$). The 'interaction terms' appear whenever two states are not independent, though independence might be assumed for simplicity. Terms due to variability in $\mathbf{S}$ have been omitted to focus on so-called 'noise correlations,' or correlated variability while keeping the stimulus constant.

By analyzing the trial-by-trial variability in the responses of a population of neurons (illustrated as dots in Figure 2), we may recover $\mathrm{d}\mathbf{f}/\mathrm{dI}$, and from this we learn about the internal variables modulating neural responses.

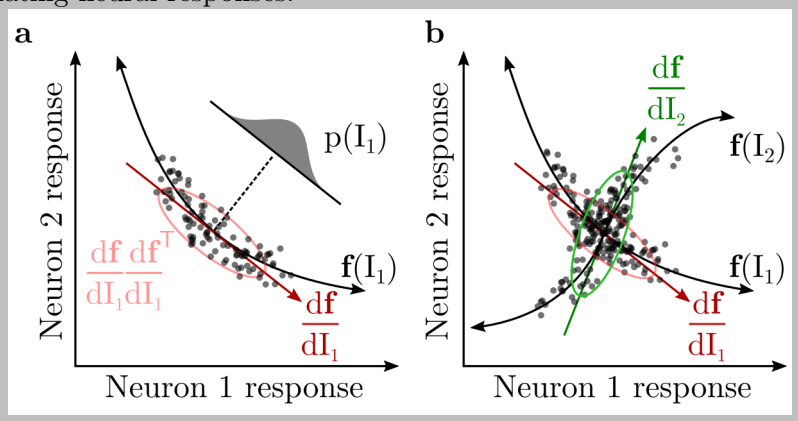

Figure 2: **a:** The population tuning of two neurons to some variable, $\mathbf{f}(\mathrm{I}_1)$, entails responses spread along $\mathbf{f}$ when I_1 varies according to $\mathrm{p}(\mathrm{I}_1)$ (each dot indicates the population response on a single trial) [31•]. The linear approximation to $\mathbf{f}$ models this as co-variability along $\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_1}\frac{\mathrm{d}\mathbf{f}}{\mathrm{dI}_1}^\top$ (ellipse), which we call the 'signature' of I_1 in this population. **b:** Variability in a second, independent internal variable, I_2, sums (equation (2)). Given certain assumptions and enough data, unsupervised methods are able to infer I_1, I_2, and $\mathbf{f}(\mathrm{I}_1, \mathrm{I}_2)$ from observations of responses alone.

†Correlations between $\mathbf{I}$ and $\mathbf{r}$ considered here may reflect causation from $\mathbf{I}$ to $\mathbf{r}$, or common input to both.

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Box 2: Case Study: Cohen and Newsome (2008) [32••]

Following Box 1 we show how the empirical correlations reported in [32••] can be understood in terms of the (linear approximation of) tuning of the neural responses to two internal variables. The authors measured pairwise noise correlations between neurons in area MT during a motion discrimination task, where the discriminated directions changed from trial to trial, but were always separated by 180°. Comparing correlations for the same zero-signal external stimulus, the authors found that noise correlations changed depending on whether the two neurons either supported the same decision, or opposite decisions ('same pool' and 'different pool' respectively, Figure 3c).

The data are well-explained by two independently varying internal variables, I_1 and I_2 that influence neural responses according to the tuning functions, $d\mathbf{f}/dI_1$ and $d\mathbf{f}/dI_2$, shown in Figure 3a. I_1 increases the gain to neurons that prefer one of the task-relevant motion directions while suppressing responses of neurons that prefer the other direction, or vice versa. I_2 varies independently of I_1, increasing the gain of all neurons selective to either task-relevant motion direction together. While these factors were hypothesized post-hoc by the authors, they can in principle be inferred directly from population recordings using the *unsupervised* methods discussed in the text. Characterizing neural variability in these terms promises deeper insights into whether any of the internal variables are best understood as graded beliefs [26, 27, 28] (if they are correlated with choice and confidence), categorical actions [33] (if they are correlated with choice but not confidence), attentional states [8••, 20, 21] (if they are correlated with performance), or general network states [4••]. Here, the authors characterized the two variables as fluctuating 'attention' on one versus the other motion direction (I_1), and fluctuating 'attention' to the task in general (I_2) [32••].

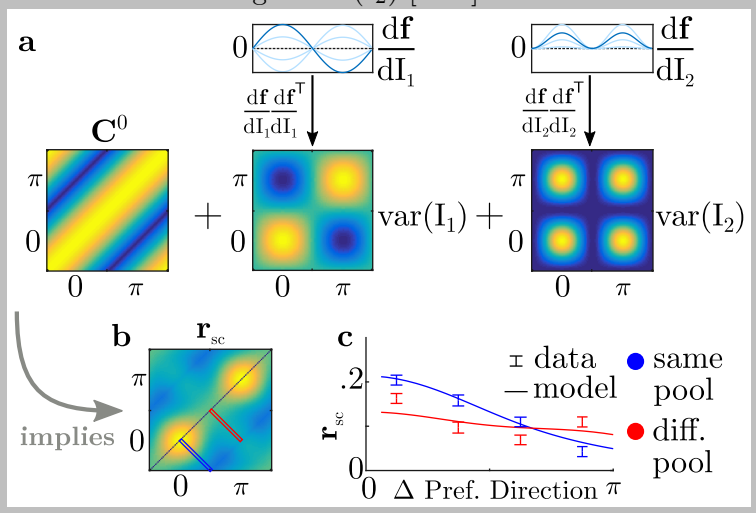

Figure 3: **a:** Visualization of equation (2) (Box 1), plots in **a-b** shown with respect to neurons' preferred motion direction relative to the task. Assuming limited range structure for $\mathbf{C^0}$ and two independently-fluctuating internal variables that act selectively on the task-relevant directions 0 and π. **b:** The sum of covariance matrices in **a** implies the correlation ($\mathbf{r}_{sc}$) structure shown here. Red and blue rectangles indicate slices shown in the next panel. **c:** Cohen and Newsome (2008) [32••] measured pairwise correlations as a function of the difference in the pair's preferred directions in two ('same pool', 'different pool') conditions. Data replotted as error bars and compared to model (lines). Note that in the absence of I_1 and I_2, the two conditions would have identical correlation structure.

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

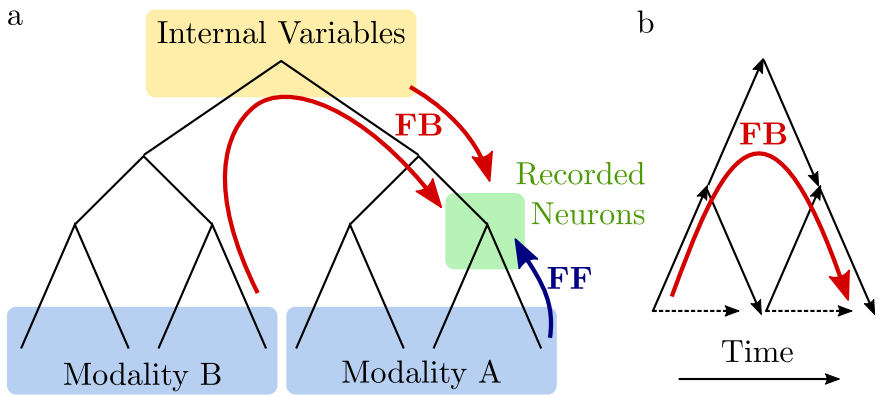

Figure 4: Cartoon of the functional relationships involved in common FB, or 'top-down' signals. **a:** FB through hierarchical structure: from the perspective of a recorded area (green), e.g. area V4 in the visual modality, lower areas within that modality are considered FF, and influences from other modalities like audition are considered FB. FB signals may also arise from intrinsically generated abstract internal variables (yellow), such as motivation, attention, or beliefs. **b:** FB through time: dependencies on external inputs at different times, sufficiently long in the past such that 'memory' about them within the FF pathway can be excluded, may also be considered FB.

2 Distinguishing feedforward from feedback influences

Two non-causal lines of evidence can help distinguish between FF and FB influences: (1) comparing measured responses to the predictions from a sufficiently powerful FF model and ascribing residuals to FB influences, and (2) comparing neural responses to identical stimuli for different induced states of internal variables under the *assumption* that they are represented by neurons downstream of those that are being observed (Figure 4). Note that we refer here only to the functional definition of FB as 'top-down' signals, and that FB signals from downstream neurons may themselves involve not just pure cortico-cortico FB pathways but also anatomically FF pathways [34], for instance via a thalamocortical loop [35, 36].

Nienborg and Cumming (2009) [37•] took the first of these approaches. Showing that the correlation of neural responses and behavior, $p(\mathbf{r}, \mathbf{B})$, had a different time course from the correlation between stimulus and behavior, $p(\mathbf{S}, \mathbf{B})$, they concluded that $p(\mathbf{r}, \mathbf{B})$ must at least in part be due to FB signals.

Most studies taking the second approach manipulate the attentional state of the subject, e.g. by presenting a pre-trial cue (Figure 4b), and have been extensively reviewed elsewhere [38]. A newer approach relying on population recordings is to analyze the co-variability of the neural responses and determine whether they contain task-dependent components. Under the critical assumption that task-learning, or task-switching, does not alter the FF connectivity or recurrent connectivity, one can then conclude that any such task-dependent component must be due to FB signals. The underlying assumption is more likely to be true for early sensory areas than higher levels of sensory processing where learning-induced plasticity has observed to be stronger [39]. The assumption is also more likely to be true for task-switches across shorter time scales, e.g. from trial to trial (seconds) rather than days or weeks over which relevant changes of the FF pathway become more likely. Both Cohen and Newsome (2008) [32••] (see Box 2) and Bondy and Cumming (2016) [25••] have taken this approach comparing changes across trials in area MT responses, and changes separated by several days in area V1 responses, respectively. Even though Rabinowitz et al (2015)

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[5••] did not explicitly vary the task, a similar logic can be applied: because the 'tuning curve' associated with one of the main latent variables driving trial-to-trial variability has a task-specific shape, it has to be due to an internal variable affecting neural responses through FB signals.

In all of these studies, training on a task induces a task-specific relationship between internal variables and sensory responses; this relationship is ascribed to FB under the assumption that those variables are represented by a 'downstream' area.

3 Implication for the information in sensory responses

As shown in Box 1, a fluctuating internal variable induces correlated variability in a population proportional to the population's 'tuning' to that variable. When 'tuning' to this internal variable, $d\mathbf{f}/dI$, is in the same direction as the population's tuning to the stimulus, $d\mathbf{f}/dS$, then variability in I has the *potential* to limit information about S, since it would result in changes in the population response that are indistinguishable from changes to the stimulus itself [31•]. However, if the internal variable covaries with the stimulus in a useful way, then it may enhance the information the population carries about S. For example, random trial-to-trial fluctuations in attention can only reduce information [16•] (because they are uncorrelated with the stimulus), while 'justified' beliefs about the stimulus that vary trial-to-trial do not [15, 27].

Hence the question of whether top-down sources of correlated variability are helpful or harmful for population coding depends not just on the 'shape' of the induced correlations, but more generally on whether the internal variables underlying them carry information about the stimulus.

4 Conclusion

The influence of internal variables on sensory responses can be characterized by tuning functions in analogy to the influence of the external stimulus on neural responses. Uncontrolled fluctuations in those variables leaves characteristic signatures in the statistics of sensory responses that can in principle be exploited to infer the state of those variables and variable-specific tuning functions. Interpreting tuning to internal variables and inferring their states allows us to ask new questions (Box 2): Which internal variables contain information about the stimulus, and which about the choice [13]? Which variables are correlated with overall performance in the task? Do they correspond to variables in a computational model of the brain performing the task? Answers to these questions promise deeper insights into the relationship between correlated variability and stimulus information in neural populations, and may allow us to bridge Marr's level of hierarchy between computational models and their neural implementation [7].

Acknowlegdements

We thank Alexander Ecker, Ruben Moreno-Bote, Daniel Chicharro, Jacob Yates, Greg DeAngelis, and Aki Anzai for feedback on the manuscript.

References and recommended reading

[1] Felleman, D.J. and Van Essen, D.C. (1991). Distributed hierachical processing in the primate cerebral cortex. Cerebral Cortex *1*, 1–47.

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

[3] Renart, A. and Machens, C.K. (2014). Variability in neural activity and behavior. Current Opinion in Neurobiology *25*, 211–220.

[4••] 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.

••Fit a shared, slow-varying latent source of additive noise on a population during anesthesia. Finds that it explains a large amount of the difference in correlated variability of V1 populations between awake and anesthetized conditions.

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

••Modeled V4 neural population activity as a combination of known (attentional cue, stimulus) and unknown (global gain, shared latent modulators) factors. A small number of shared latent factors explained much of the residual variance not accounted for by other sources. Task-dependence of latent factors implies interesting computational role for them.

[6•] Cunningham, J.P. and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nature Neuroscience *17*, 1500–1509.

•Review dimensionality reduction techniques used for characterizing sensory responses.

[7] Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Phenomenology and the Cognitive Sciences *8*, 397.

[8••] Cohen, M.R. and Maunsell, J.H.R. (2010). A Neuronal Population Measure of Attention Predicts Behavioral Performance on Individual Trials. The Journal of Neuroscience *30*, 15241–15253.

••Measure change in population responses as attention changes, and use this to extract a continuous estimate of attention on a trial-by-trial basis. The magnitude of this inferred attention level is correlated with behavioral accuracy.

[9] Cohen, M.R. and Maunsell, J.H.R. (2011). Using neuronal populations to study the mechanisms underlying spatial and feature attention. Neuron *70*, 1192–204.

[10] Cohen, M.R. and Maunsell, J.H.R. (2011). When attention wanders: how uncontrolled fluctuations in attention affect performance. J Neurosci *31*, 15802–15806.

[11] Mayo, J.P., Cohen, M.R., and Maunsell, J.H.R. (2015). A refined neuronal population measure of visual attention. PLoS ONE *10*, 1–12.

[12] Ni, A.M., Ruff, D.A., Alberts, J.J., Symmonds, J., and Cohen, M.R. (2017). Learning and attention reveal a general relationship between neuronal variability and perception. bioRxiv , 1–28.

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[13] Panzeri, S., Harvey, C.D., Piasini, E., Latham, P.E., and Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention and behavior. Neuron Perspective *93*, 491–507.

[14] 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.

[15] 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.

[16•] Ecker, A.S., Denfield, G.H., Bethge, M., and Tolias, A.S. (2016). On the structure of population activity under fluctuations in attentional state. Journal of Neuroscience *0*, 1–21.

•Studies the theoretical effects of fluctuations in attention on response correlations and their implication for Fisher Information.

[17] Putzky, P., Franzen, F., Bassetto, G., and Macke, J.H. (2014). A Bayesian model for identifying hierarchically organised states in neural population activity. Advances in Neural Information Processing Systems 27 , 3095–3103.

[18] Pandarinath, C., O'Shea, D.J., Collins, J., Jozefowicz, R., Stavisky, S.D., Kao, J.C., Trautmann, E.M., Kaufman, M.T., Ryu, S.I., Hochberg, L.R., et al. (2017). Inferring single-trial neural population dynamics using sequential auto-encoders. bioRxiv , 1–22.

[19] 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.

[20] Mitchell, J.F., Sundberg, K.A., and Reynolds, J.H. (2009). Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. Neuron *63*, 879–888.

[21] Cohen, M.R. and Maunsell, J.H.R. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nature Neuroscience *12*, 1594–600.

[22] Stănişor, L., Togt, C.v.d., Pennartz, C.M.A., and Roelfsema, P.R. (2013). A unified selection signal for attention and reward in primary visual cortex. PNAS *110*, 9136–9141.

[23] Niell, C.M. and Stryker, M.P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. Neuron *65*, 472–479.

[24] Nienborg, H. and Cumming, B.G. (2007). Psychophysically measured task strategy for disparity discrimination is reflected in V2 neurons. Nature neuroscience *10*, 1608–14.

[25••] Bondy, A.G. and Cumming, B.G. (2016). Feedback Dynamics Determine the Structure of Spike-Count Correlation in Visual Cortex. bioRxiv , 1–41.

••Fit full noise correlation matrices to macaque V1 neurons in changing task contexts. The full matrix is dominated by a task-dependent component that is itself predominantly rank-1, consistent with a single (task-dependent) varying internal state.

bioRxiv preprint doi: https://doi.org/10.1101/114439; this version posted July 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

[27] Lange, R.D. and Haefner, R.M. (2016). Inferring the brain's internal model from sensory responses in a probabilistic inference framework. bioRxiv .

[28] Tajima, C.I., Tajima, S., Koida, K., Komatsu, H., Aihara, K., and Suzuki, H. (2016). Population code dynamics in categorical perception. Nature Scientific Reports *6*, 1–13.

[29] Gao, P. and Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. Current Opinion in Neurobiology *32*, 148–155.

[30] Goris, R.L.T., Movshon, J.A., and Simoncelli, E.P. (2014). Partitioning neuronal variability. Nature Neuroscience *17*, 858–865.

[31•] Moreno-Bote, R., Beck, J.M., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. Nature Neuroscience *17*, 1410–1417.

•Emphasize the point that response covariability in the $\frac{d\mathbf{f}}{dS}\frac{d\mathbf{f}}{dS}^\top$-direction limits information, when induced by the feedforward pathway.

[32••] Cohen, M.R. and Newsome, W.T. (2008). Context-Dependent Changes in Functional Circuitry in Visual Area MT. Neuron *60*, 162–173.

••First to show task-dependence of response correlations in MT neurons.

[33] Sommer, M.A. and Wurtz, R.H. (2008). Visual Perception and Corollary Discharge. Perception *37*, 408–418.

[34] Markov, N.T., Ercsey-Ravasz, M., Van Essen, D.C., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). Cortical High-Density Counterstream Architectures. Science *342*.

[35] Sherman, S.M. (2016). Thalamus plays a central role in ongoing cortical functioning. Nature neuroscience *16*, 533–41.

[36] Briggs, F., Mangun, G.R., and Usrey, W.M. (2013). Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature *499*, 476–80.

[37•] 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.

•Show that choice probabilities cannot be accounted for purely by a feedforward readout and constant stimulus co-variability, implying choice-related activity in V2 contains a feedback component.

[38] Moore, T. and Zirnsak, M. (2017). Neural Mechanisms of Selective Visual Attention. Annual Review of Psychology *68*, 47–72.

[39] Ahissar, M. and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. Trends in Cognitive Sciences *8*, 457–464.