



# On the Number of Neurons and Time Scale of Integration Underlying the Formation of Percepts in the Brain

Adrien Wohrer, Christian Machens

## ► To cite this version:

Adrien Wohrer, Christian Machens. On the Number of Neurons and Time Scale of Integration Underlying the Formation of Percepts in the Brain. PLoS Computational Biology, 2015, 11 (3), pp.e1004082. 10.1371/journal.pcbi.1004082 . hal-01905047

**HAL Id: hal-01905047**

**<https://uca.hal.science/hal-01905047>**

Submitted on 19 Dec 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

RESEARCH ARTICLE

# On the Number of Neurons and Time Scale of Integration Underlying the Formation of Percepts in the Brain

Adrien Wohrer<sup>1\*</sup> , Christian K. Machens<sup>1,2</sup>

**1** Group for Neural Theory, Laboratoire de Neurosciences Cognitives, INSERM U960, École Normale Supérieure, Paris, France, **2** Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, Lisbon, Portugal

 Current address: Image-Guided Clinical Neurosciences and Connectomics, EA 7282, Université d'Auvergne, Clermont-Ferrand, France

\* [adrien.wohrer@udamail.fr](mailto:adrien.wohrer@udamail.fr)



## Abstract

All of our perceptual experiences arise from the activity of neural populations. Here we study the formation of such percepts under the assumption that they emerge from a linear readout, i.e., a weighted sum of the neurons' firing rates. We show that this assumption constrains the trial-to-trial covariance structure of neural activities and animal behavior. The predicted covariance structure depends on the readout parameters, and in particular on the temporal integration window  $w$  and typical number of neurons  $K$  used in the formation of the percept. Using these predictions, we show how to infer the readout parameters from joint measurements of a subject's behavior and neural activities. We consider three such scenarios: (1) recordings from the complete neural population, (2) recordings of neuronal sub-ensembles whose size exceeds  $K$ , and (3) recordings of neuronal sub-ensembles that are smaller than  $K$ . Using theoretical arguments and artificially generated data, we show that the first two scenarios allow us to recover the typical spatial and temporal scales of the readout. In the third scenario, we show that the readout parameters can only be recovered by making additional assumptions about the structure of the full population activity. Our work provides the first thorough interpretation of (feed-forward) percept formation from a population of sensory neurons. We discuss applications to experimental recordings in classic sensory decision-making tasks, which will hopefully provide new insights into the nature of perceptual integration.

## OPEN ACCESS

**Citation:** Wohrer A, Machens CK (2015) On the Number of Neurons and Time Scale of Integration Underlying the Formation of Percepts in the Brain. *PLoS Comput Biol* 11(3): e1004082. doi:10.1371/journal.pcbi.1004082

**Editor:** Jonathan W. Pillow, The University of Texas at Austin, UNITED STATES

**Received:** August 27, 2013

**Accepted:** December 10, 2014

**Published:** March 20, 2015

**Copyright:** © 2015 Wohrer, Machens. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors acknowledge support from an "Emmy-Noether Grant" of the Deutsche Forschungsgemeinschaft (Germany) and a "Chaire d'excellence" of the Agence Nationale de la Recherche (France). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Author Summary

This article deals with the interpretation of neural activities during perceptual decision-making tasks, where animals must assess the value of a sensory stimulus and take a decision on the basis of their percept. A "standard model" for these tasks has progressively emerged, whence the animal's percept and subsequent choice on each trial are obtained from a linear integration of the activity of sensory neurons. However, up to date, there has

been no principled method to estimate the parameters of this model: mainly, the typical number of neurons  $K$  from the population involved in conveying the percept, and the typical time scale  $w$  during which these neurons' activities are integrated. In this article, we propose a novel method to estimate these quantities from experimental data, and thus assess the validity of the standard model of percept formation. In the process, we clarify the predictions of the standard model regarding two classic experimental measures in these tasks: *sensitivity*, which is the animal's ability to distinguish nearby stimulus values, and *choice signals*, which assess the amount of correlation between the activity of single neurons and the animal's ultimate choice on each trial.

## Introduction

Most cortical neurons are noisy, or at least appear so in experiments. When we record the responses of sensory neurons to well-controlled stimuli, their spike patterns vary from trial to trial. Does this variability reflect the uncertainties of the measurement process, or does it have a direct impact on behavior? These questions are central to our understanding of percept formation and decision-making in the brain and have been the focus of much previous work [1]. Many studies have sought to address these problems by studying animals that perform simple, perceptual decision-making tasks [2, 3]. In such tasks, an animal is typically presented with different stimuli  $s$  and trained to categorize them through a simple behavioral report. When this perceptual report is monitored simultaneously with the animal's neural activity, one can try to find a causal link between the two.

One particular hypothesis about this link—which we refer to as the “sensory noise” hypothesis—postulates that the accuracy of the animal's perceptual judgments is primarily limited by noise at the level of sensory neurons [4, 5]. In terms of signal detection theory, the hypothesis predicts a quantitative match between (1) the animal's ability to discriminate nearby stimulus values—known as *psychometric* sensitivity, and (2) an ideal observer's ability to discriminate nearby stimulus values based on the activities of the underlying neural population—known as *neurometric* sensitivity. Both types of sensitivities can be quantified as signal-to-noise ratios (SNR). With this idea in mind, several studies have compared the neurometric and psychometric sensitivities in various sensory systems and behavioral tasks (see [6, 7] for reference).

However, as was soon realized, any extrapolation from a few recorded cells to the entire population is fraught with implicit assumptions. For example, if neurons in a population behave independently one from another, then the SNR of the population is simply the sum of the individual SNRs. Consequently, any estimate of neurometric sensitivity will grow linearly with the number of recorded neurons  $K$ . However, if neurons in a population do not behave independently, the precise growth of neural sensitivity with  $K$  is determined by the correlation structure of noise in the population [8–10]. In addition, the neurometric sensitivities also depend on the time scale  $w$  that is used to integrate each neuron's spike train in a given trial [3, 11–13]. Indeed, the more spikes are incorporated in the readout, the more accurate that readout will be. Adding extra neurons by increasing  $K$ , or adding extra spikes by increasing  $w$ , are two dual ways of increasing the readout's overall SNR.

As there is no unique way of reading out information from a population of sensory neurons, the sensory noise hypothesis can only be tested if we understand how the organism itself “reads out” the relevant information. In other words, how many sensory neurons  $K$ , and what integration time scale  $w$ , provide a relevant description of the animal's percept formation? Given the

“ $K$ - $w$ ” duality mentioned above, we cannot answer that question based solely on sensitivity (SNR). Another experimental measure should also be included in the analysis.

A good candidate for such a measure are *choice signals*, i.e., measures of the trial-to-trial correlation between the activity of each recorded neuron and the animal’s ultimate choice on each trial. These signals, weak but often significant, arise from the unknown process by which each neuron’s activity influences—or is influenced by—the animal’s perceptual decision. In two-alternative forced choice (2AFC) discrimination tasks, they have generally been computed in the form of *choice probabilities* (CP) [14, 15]. The temporal evolution of CPs has been used to find the instants in time when a given population covaries with the animal’s percept [13, 16]. In a seminal study, Shadlen et al. (1996) proposed to jointly use sensitivity and choice signals, as two independent constraints characterizing the underlying neural code [17]. They derived a feed-forward model of perceptual integration in visual area MT, and studied numerically how the population’s sensitivity and CPs vary as a function of various model parameters. They acknowledged the existence of a link between CPs and pairwise noise correlations—both measures being (partial) reflections of how information is embedded in the neural population as a whole (see also [12, 18]). However, the quantitative nature of this link was only revealed recently, when Haefner et al. (2013) derived the analytical expression of CPs in the standard model of perceptual integration [19] (see [Methods](#)).

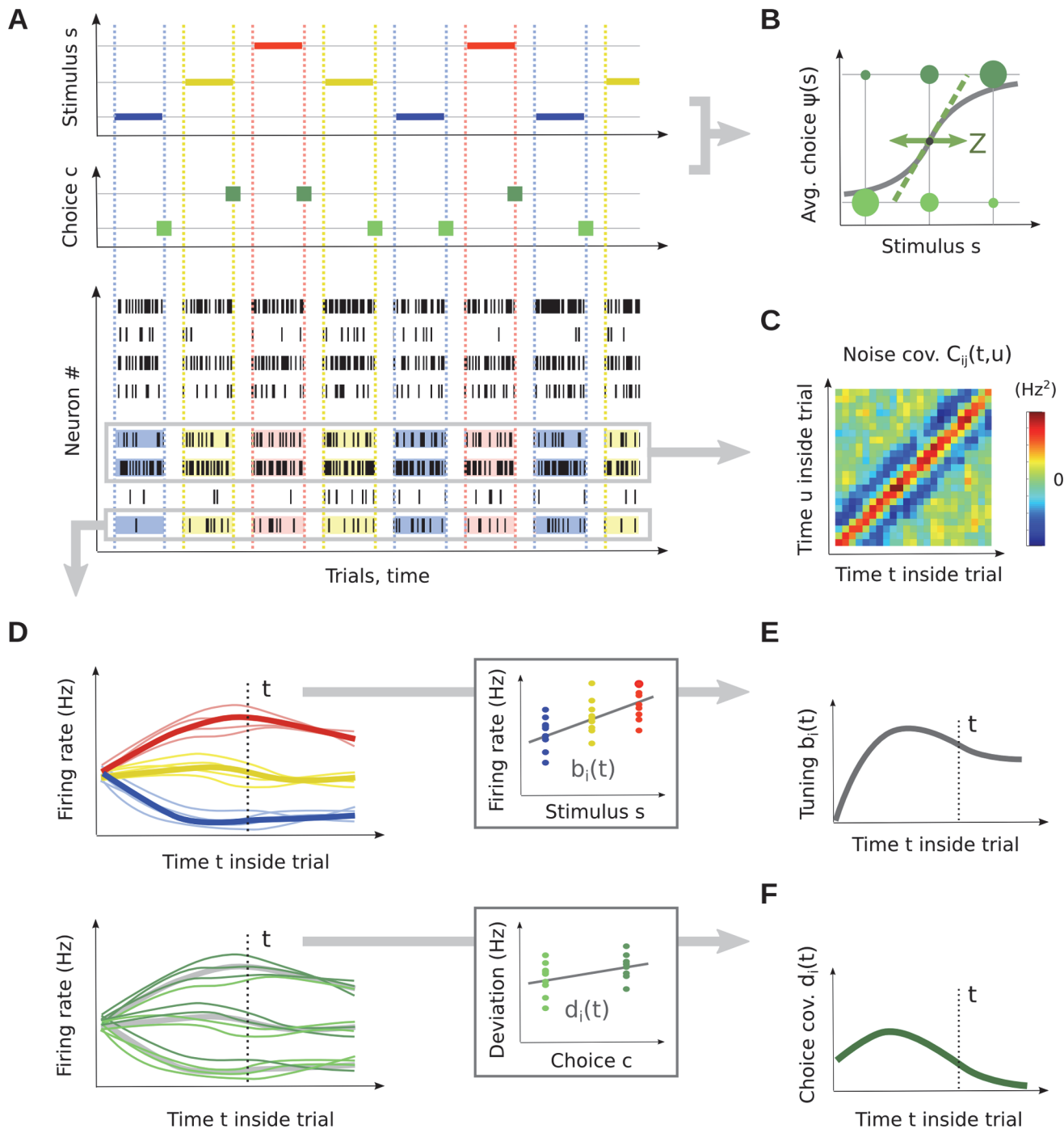
In this article, we show that the standard feed-forward model of percept formation gives rise to three *characteristic equations* that describe analytically the trial-to-trial covariance between neural activities and animal percept. These equations depend on the brain’s readout policy across neurons and time, and hold for any noise correlation structure in the neural population. In accordance with the intuition of Shadlen et al. (1996), we show that sensitivity and choice signals correspond to two distinct, characteristic properties of the readout. The equation describing choice signals is equivalent to the one derived by Haefner et al. (2013), but stripped from the non-linear complications inherent to the CP formulation. We use a linear formulation instead, which gives us a particularly simple prediction of choice signals at every instant in time.

We then show how these equations can be used in order to recover the time window and the number of neurons used in the formation of a percept. A quantitative analysis of choice signals allows us to overcome the “ $K$ - $w$  trade-off” inherent to neurometric sensitivity. We specifically focus on situations in which only a finite sample of neurons has been measured from a large, unknown population. We show how to recover the typical number of neurons  $K$ , provided that the experimenter could record at least  $K$  neurons simultaneously. Finally, we discuss the scope and the limitations of our method, and how it can be applied to real experimental data.

## Results

### Experimental measures of behavior and neural activities

We will study the formation of percepts in the context of perceptual decision-making experiments ([Fig. 1](#), see [Methods](#) or [Tables 1–3](#) for the corresponding formulas). In these experiments, an animal is typically confronted with a stimulus,  $s$ , and must then make a behavioral choice,  $c$ , according to the rules of the task. A specific example is the classic *discrimination task* in which the animal’s choice  $c$  is binary, and the animal must report whether it perceived  $s$  to be higher ( $c = 1$ ) or lower ( $c = 0$ ) than a fixed reference  $s_0$  ([Fig. 1A](#), top and middle panels). While the animal is performing the task, the neural activity in a given brain area can be monitored ([Fig. 1A](#), bottom panel). Typical examples from the literature include area MT in the context of a motion discrimination task [3], area MT or V2 in the context of a depth



**Fig 1. Framework and main experimental measures.** (A) Experimental setup. Top: A set of stimulus values  $s$  (color-coded as blue, yellow, red) are repeatedly presented to an animal. Middle: The animal's choice  $c$  on each trial (green) indicates whether the animal judged  $s$  to be larger or smaller than the fixed central value,  $s_0$ . Bottom: In each session, several task-relevant sensory neurons are recorded simultaneously with the behavior. (B) The psychometric curve  $\psi(s)$  quantifies the animal's sensory accuracy. Its inverse slope in  $s_0$  provides the just-noticeable-difference (JND),  $Z$ . (C) The noise covariance structure can be assessed in each pair of simultaneously recorded neurons, as their joint peri-stimulus histogram (JPSTH)  $C_{ij}(t, u)$ . (D) Responses of a particular neuron. Each thin line is the schematic (smoothed) representation of the spike train on one trial. Segregating trials according to stimulus (top), we access the neuron's peri-stimulus histogram (PSTH, thick lines) and its tuning signal  $b_i(t)$ —shown in panel (E). Fixing a stimulus value and segregating trials according to the animal's choice  $c$  (bottom), we access the neuron's choice covariance (CC) curve  $d_i(t)$ —shown in panel (F).

doi:10.1371/journal.pcbi.1004082.g001

**Table 1. Variables and notations: typography.**

Notation	Description	Examples
Bold	Lower case: vector notation, across neurons Upper case: matrix notation, across neurons	$\mathbf{r}(t)$ , $\mathbf{b}(t)$ $\mathbf{C}(t, u)$ , $\bar{\mathbf{C}}$
Overline	Temporal integration, using readout parameters ( $w$ , $t_R$ )	$\bar{r}_i$ , $\bar{C}_{ij}$ , $\bar{q}$
Starred	Pertaining to the animal's true behavior (as opposed to model-based predictions)	$Z^*$ , $d_i^*(t)$ , $t_R^*$
$E[\cdot]$	Expectation across trials (can also be conditional)	$E[s]$ , $E[r_i(t) s]$ , $E[sr_i(t)]$
$\text{Cov}[\cdot]$	Covariance across trials (can also be conditional)	$\text{Cov}[r_i(t), c^* s]$
$\mathbf{x}_r$	Vector (or matrix) $\mathbf{x}$ restricted to neurons in readout ensemble $\mathcal{E}$	$\mathbf{a}_r$ , $\bar{\mathbf{b}}_r$ , $\bar{\mathbf{C}}_r$ , $\bar{\mathbf{C}}_{rr}(t)$

doi:10.1371/journal.pcbi.1004082.t001

**Table 2. Variables and notations: experimental data.**

Raw experimental data			Ref in text
$s$	Stimulus—a varying scalar value on each trial		
$s_0$	Threshold stimulus value in the 2AFC task		
$c^*$	Animal choice—binary report on each trial		
$r_i(t)$	Spike train from neuron $i$ in a given trial		
$\sigma_s^2$	Stimulus variance across trials	$\sigma_s^2 := \text{Var}[s]$	after <a href="#">eq. 18</a>
Animal psychometry			
$\psi^*(s)$	Psychometric curve	$\psi^*(s) := E[c^* s]$	<a href="#">eq. 23</a>
$Z^*$	Just-noticeable difference	Best fit to $\psi^*(s) = \Phi\left(\frac{s + \mu_d^* - s_0}{Z^*}\right)$	<a href="#">eq. 27</a>
$\mu_d^*$	Decision bias		
Individual statistics for the neurons			
$m_i(t; s)$	PSTH for neuron $i$ with stimulus $s$	$m_i(t; s) := E[r_i(t) s]$	<a href="#">eq. 24</a>
$b_i(t)$	Tuning signal for neuron $i$ (variation of the PSTH wrt. stimulus)	$b_i(t) := \partial_s m_i(t; s)$	<a href="#">eq. 30, 31</a>
$C_{ij}(t, u)$	JPSTH for neurons $i$ and $j$ (pairwise noise correlations)	$C_{ij}(t, u) := E[\text{Cov}[r_i(t), r_j(u) s]]$	<a href="#">eq. 25, 32</a>
$\bar{\mathbf{C}}$	Noise covariance matrix (time average of $\mathbf{C}$ , with parameters ( $w$ , $t_R$ ))	$\bar{\mathbf{C}}_{ij} := E[\text{Cov}[\bar{r}_i, \bar{r}_j s]]$	<a href="#">eq. 40</a>
$d_i^*(t)$	Choice Covariance for neuron $i$ (linear equivalent of choice probabilities)	$d_i^*(t) := E[\text{Cov}[r_i(t), c^* s]]$	<a href="#">eq. 26, 33</a>

doi:10.1371/journal.pcbi.1004082.t002

discrimination task [11, 20], or area S1 in the context of a tactile discrimination task [21]. For concreteness, we will mostly focus on these discrimination tasks, although the general framework can be applied to arbitrary perceptual decision-making tasks.

The animal's behavior in a discrimination task can be quantified through the *psychometric curve*  $\psi(s)$ . This curve measures the animal's repartition of responses at each stimulus value  $s$  (Fig. 1B). If the animal is unbiased, it will choose randomly whenever the stimulus  $s$  is equal to the threshold value  $s_0$ , so that  $\psi(s_0) = 1/2$ . The slope of the psychometric curve at  $s = s_0$  determines the animal's ability to distinguish near-threshold values of the stimulus, i.e., its psychometric sensitivity. We assess this sensitivity through the *just noticeable difference* (JND) or *difference limen*, noted  $Z$ . The more sensitive the animal, the smaller  $Z$ , and the steeper its psychometric curve.



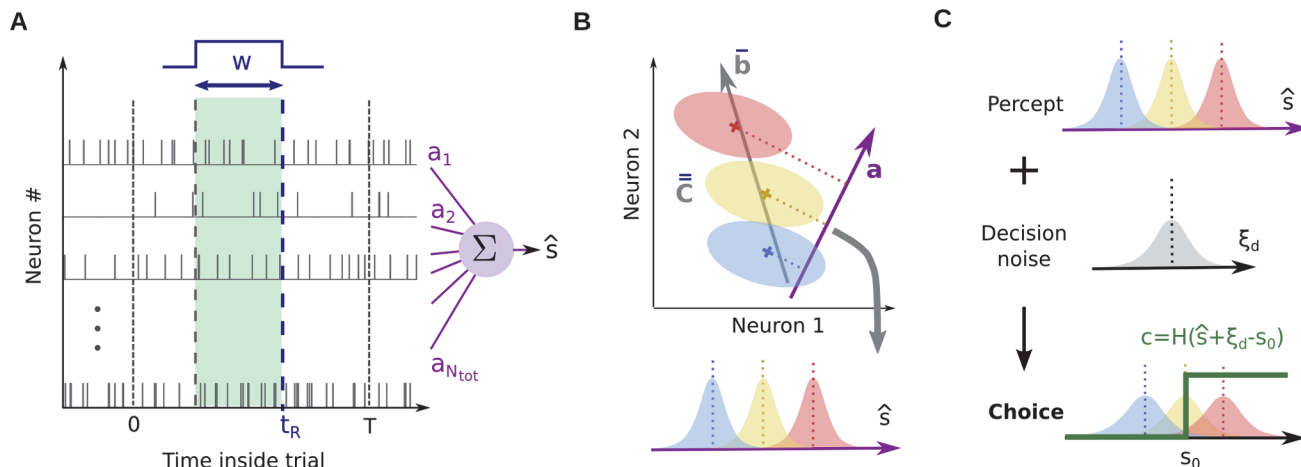
**Table 3. Variables and notations: model and methods.**

Linear readout and decision model			Ref in text
$w$	Readout window—duration of the temporal integration		
$t_R$	extraction time—time at which the percept is formed		
$a_i$	Readout weight—contribution of neuron $i$ to the percept		
$\sigma_d$	Decision noise—added to the percept at decision time		
$\hat{s}$	Readout—computed on every trial	$\hat{s} = a_0 + \sum_{i=1}^{N_{\text{tot}}} a_i \bar{r}_i$	eq. 2
$c$	Choice—computed on every trial	$c = H(\hat{s} + \xi_d - s_0)$	eq. 3
<b>Model predictions</b> (characteristic equations)			
$Z$	Just-noticeable difference	$Z^2 = \mathbf{a}^T \bar{\mathbf{C}} \mathbf{a} + \sigma_d^2$	eq. 8
$\mathbf{d}(t)$	Choice covariance for every neuron	$\mathbf{d}(t) = \kappa(Z) \bar{\mathbf{C}}(t) \mathbf{a}$	eq. 9
$\kappa(Z)$	Conversion factor from <i>Percept Covariance</i> to <i>Choice Covariance</i>		eq. 46
<b>Restricted optimality</b> hypothesis			
$\mathcal{E}$	Readout ensemble—neurons used for the readout		
$K$	Readout size—number of neurons in $\mathcal{E}$		
$\mathbf{H}$	Restriction matrix on $\mathcal{E}$ (of size $K \times N_{\text{tot}}$ )	$\mathbf{x}_r = \mathbf{H} \mathbf{x}$	eq. 54
$\mathbf{a}_r$	Optimal readout vector (over $\mathcal{E}$ )	$\mathbf{a}_r \sim (\bar{\mathbf{C}}_r)^{-1} \bar{\mathbf{b}}_r$	eq. 12
<b>Population-wide indicators for choice signals</b>			
$q(u, t)$	Population-wide link between tuning and CC	$q(u, t) := \langle b_i(u) d_i(t) \rangle_i$	eq. 14
$V$	Deviation from linearity between tuning and CC	$V := \langle \bar{b}_i^2 \rangle_i \langle \bar{d}_i^2 \rangle_i - \bar{q}^2$	eq. 15
<b>Rescaled indicators</b> (used in the SVD analysis)			
$\mathbf{A}$	Total covariance matrix	$\mathbf{A} = \bar{\mathbf{C}} + \sigma_s^2 \bar{\mathbf{b}} \bar{\mathbf{b}}^T$	eq. 50
$Y$	Sensitivity to stimulus	$Y := \sigma_s^2 / (\sigma_s^2 + Z^2)$	eq. 47
$\mathbf{e}$	Total percept covariance	$\mathbf{e} := \mathbf{A} \mathbf{a}$	eq. 76
$Q$	Rescaled version of $\bar{q}$	$Q := \langle e_i \bar{b}_i \rangle_i$	eq. 18, 79
$\boldsymbol{\eta}$	Tuning vector in the space of modes	$\bar{\mathbf{b}} = \mathbf{U} \boldsymbol{\Lambda} \boldsymbol{\eta}$	eq. 68

doi:10.1371/journal.pcbi.1004082.t003

We assume that the neural activity within the recorded brain area conveys the stimulus information that the animal uses to make its choice (Fig. 1A, bottom). We describe the activity of this neural population on every trial as a multivariate point process  $\mathbf{r}(t) = \{r_i(t)\}_{i=1 \dots N_{\text{tot}}}$ , where each  $r_i(t)$  is the spike train for neuron  $i$ , and  $N_{\text{tot}}$  denotes the full population size, a very large and unknown number. (The number of neurons actually recorded is generally much smaller.) As is common in electrophysiological recordings, we will quantify the raw spike trains by their first and second order statistics. First, neuron  $i$ 's trial-averaged activity in response to each tested stimulus  $s$  is given by the peri-stimulus time histogram (PSTH) or time-varying firing rate,  $m_i(t; s)$  (Fig. 1D). In so-called “fine” discrimination tasks, the stimuli  $s$  display only moderate variations around the central value  $s_0$ , so that the PSTH at each instant in time can often be approximated by a linear function of  $s$ :  $m_i(t; s) \simeq m_i^0(t) + b_i(t)s$ . The slope  $b_i(t)$ , defined at every instant in time, summarizes neuron  $i$ 's tuning properties (Fig. 1E). Second, we assume that several neurons can be recorded simultaneously, so that we can access samples from the trial-to-trial covariance structure of the population activity (Fig. 1C). For every pair of neurons  $(i, j)$  and instants in time  $(t, u)$ , the joint peri-stimulus time histogram (JPSTH, [22])  $C_{ij}(t, u)$  summarizes the pairwise noise correlations between the two neurons (eq. 25). For simplicity, we furthermore assume that the JPSTHs do not depend on the exact stimulus value  $s$ .

Finally, we can measure a choice signal for each neuron, which captures the trial-to-trial covariation of neuron activity  $r_i(t)$  with the animal's choice (Fig. 1F). Traditionally, this signal is



**Fig 2. Linear readout and its interpretation.** (A) We study a “standard” model of percept formation, with two parameters  $w$  and  $t_R$  defining integration in time, and a readout vector  $\mathbf{a}$  defining integration across neurons. (B) Geometric interpretation of the model. The temporal parameters  $w$  and  $t_R$  define the tuning vector  $\bar{\mathbf{b}}$  and noise covariance matrix  $\bar{\mathbf{C}}$  in the population. Colored ellipses represent the distribution of neural activities from trial to trial, for the three possible stimulus values. The readout  $\hat{s}$  can be viewed as an orthogonal projection of neural activities in the direction given by  $\mathbf{a}$ . (C) Behavioral part of the model. The percept  $\hat{s}$  can be corrupted by decision noise  $\xi_d$ . Then it is thresholded to produce a binary choice  $c$ .

doi:10.1371/journal.pcbi.1004082.g002

measured in the form of choice probability (CP) curves. We consider here a simpler linear equivalent, that we term *choice covariance* (CC) curves [3]. The CC curve for neuron  $i$ , denoted by  $d_i(t)$ , measures the difference in firing rate (at each instant in time) between trials where the animal chose  $c = 1$  and trials where it chose  $c = 0$ —all experimental features (including stimulus value) being fixed.

Unlike many characterizations of neural activity that rely only on spike counts, our framework requires an explicit temporal description of neural activity through PSTHs, JPSTHs, and CC curves. Exact formulas for these statistical measures are provided in the Methods. By keeping track of time, we will be able to predict *when*, and *how long*, perceptual integration takes place in an organism.

## From the neural activities to the animal’s choice

**Linear readout model.** Our goal is to quantify the mapping from the neural activities,  $\mathbf{r}(t)$ , to the animal’s choice,  $c$ . This can be done if we assume (1) how the stimulus information is extracted from the neural activities and (2) how the animal’s decision is formed. For (1) we assume the common linear readout model (Fig. 2A). Here, each neuron’s spike train  $r_i(t)$  is first integrated into a single number describing the neuron’s activity over the trial. We write,

$$\bar{r}_i = \frac{1}{w} \int_{t=0}^{t_R} dt h\left(\frac{t_R - t}{w}\right) r_i(t), \quad (1)$$

where the kernel  $h(\cdot)$  defines the shape of the integration window (e.g., square window, decreasing exponential, etc.), the parameter  $w$  controls the length of temporal integration, and the parameter  $t_R$  specifies the time at which the percept is built or read out. Second, the actual percept is given by a weighted sum over the neurons’ activities,

$$\hat{s} = a_0 + \sum_{i=1}^{N_{\text{tot}}} a_i \bar{r}_i, \quad (2)$$



where  $\mathbf{a} = (a_1, \dots, a_{N_{\text{tot}}})$  is a specific readout vector, or “perceptual policy”. This classic linear readout has sometimes been referred to as the “standard” model of perceptual integration [17, 19].

Previous studies have generally made ad hoc choices for the various constituents of this model. Most often,  $\bar{r}_i$  is taken to be the total spike count for neuron  $i$ , in which case  $t_R = w$  coincides with the end of the stimulation period, and  $h(\cdot)$  in eq. 1 is a square kernel. However, this readout is likely incorrect: the length of the integration window  $w$  influences the neurometric sensitivity, and experiments suggest that animals do not always use the full stimulation period to build their judgment [23]. Similarly, vector  $\mathbf{a}$  is often defined over an arbitrary set of neurons, typically those recorded by the experimenter. Again, this choice is arbitrary, and it has a direct influence on the predicted sensitivities.

Instead, we view the readout window  $w$  and extraction time  $t_R$  as free parameters, and we generically define  $\mathbf{a}$  over the full, unknown population of neurons. If a neuron does not contribute to the percept, it simply corresponds to a zero entry in  $\mathbf{a}$ . For conceptual and implementation simplicity, we take  $h(\cdot)$  to be a simple square window (see Discussion for a generalization). Our goal is now to understand whether the readout  $\hat{s}$  can be a good model for the animal’s true percept formation and if yes, for what set of parameters.

**Decision policy.** The linear model builds a continuous-valued, internal percept  $\hat{s}$  of stimulus value by the animal on each trial. To emulate the discrimination tasks, we also need to model the animal’s decision policy, which converts the continuous percept  $\hat{s}$  into a binary choice  $c$ . While the linear model is rather universal, the decision model will depend on the specifics of each experimental task. To ground our argumentation, we model here the required decision in a classic random dot motion discrimination task [3]. However, the ideas herein could also be transposed to other types of behavioral tasks (see Discussion).

On each trial, we assume that an extraneous source of noise  $\xi_d$  is added to the animal’s percept  $\hat{s}$  (Fig. 2C). Known in the literature as ‘decision noise’ or ‘pooling noise’,  $\xi_d$  encompasses all extra-sensory sources of variation which may influence the animal’s decision. We assume that  $\xi_d$  is a Gaussian variable with variance  $\sigma_d^2$ , which we take as an additional model parameter. Then, the animal’s choice on each trial is built deterministically, by comparing  $\hat{s} + \xi_d$  to the threshold value  $s_0$  (Fig. 2C), so that

$$c = H(\hat{s} + \xi_d - s_0), \quad (3)$$

where  $H(\cdot)$  is the Heaviside function. We note that the decision noise is negligible in the classic “sensory noise hypothesis”, in which case  $\sigma_d \rightarrow 0$ .

## The characteristic equations of the standard model

The linear readout model and the animal’s decision policy specify both how the animal’s percepts are formed from its neural activities and how its choices are generated from these percepts. If we had recorded the activities of the entire neural population together with the animal’s behavior, then the parameters of this model could be estimated from the data using any standard regression method. However, this is generally not a realistic experimental situation. Instead, we take here a statistical approach to the problem, which (1) allows us to deal with incomplete recordings and (2) relates the estimation problem to the standard experimental measures described above.

**Characteristic equations of the linear readout.** Thanks to its linear structure, the readout defined in eq. 2 induces a simple covariance between the neural activities,  $\mathbf{r}(t)$ , and the resulting percept,  $\hat{s}$  (Fig. 2B). Since the linear readout relies on the integrated spike trains, eq. 1, we need similarly integrated versions of the neural tuning and noise covariances in order to express the

respective covariance relations. In general, we will denote these time-integrated quantities by an overhead bar, and alert the reader that the respective quantities depend implicitly on the readout window  $w$  and the extraction time  $t_R$ . We will write  $\bar{b}_i$  for the integrated version of the neural tuning,  $b_i(t)$ , we will write  $\bar{C}_{ij}(t)$  for the once integrated noise covariances, and  $\bar{\bar{C}}_{ij}$  for the doubly integrated noise covariance. This latter quantity, known in the literature as the ‘noise covariance matrix’, measures how the spike counts of two neurons,  $\bar{r}_i$  and  $\bar{r}_j$ , covary due to shared random fluctuations across trials (stimulus  $s$  being held fixed). We can then summarize the covariances between neural activities and the resulting percepts by three characteristic equations (see [Methods](#)):

$$\partial_s E[\hat{s}|s] = \bar{\mathbf{b}}^\top \mathbf{a}, \quad (4)$$

$$\text{Var}[\hat{s}|s] = \mathbf{a}^\top \bar{\bar{\mathbf{C}}} \mathbf{a}, \quad (5)$$

$$\text{Cov}[\mathbf{r}(t), \hat{s}|s] = \bar{\mathbf{C}}(t) \mathbf{a}. \quad (6)$$

On the left-hand sides of [eq. 4–6](#), we find statistical quantities related to the percept  $\hat{s}$ . On the right-hand sides of these equations, we find the model’s predictions, which are based on the neurons’ (measurable) response statistics,  $b$  and  $C$ . More specifically, the first line describes the average dependency of  $\hat{s}$  on stimulus  $s$ , the second line expresses the resulting variance for the percept, and the third line expresses the linear covariance between each neuron’s spike train, and the animal’s percept  $\hat{s}$  on the trial.

**Characteristic equations of the decision policy.** To produce a binary choice, the continuous percept  $\hat{s}$  is fed into the decision model ([Fig. 2C](#)). From the output of this decision model, we obtain a second set of characteristic equations (see [Methods](#)),

$$1 = \partial_s E[\hat{s}|s],$$

$$Z^2 = \text{Var}[\hat{s}|s] + \sigma_d^2,$$

$$\mathbf{d}(t) = \kappa(Z) \text{Cov}[\mathbf{r}(t), \hat{s}|s].$$

Here the first equation simply expresses that both percept and decision are assumed to be unbiased. The second equation relates the JND,  $Z$ , extracted from the psychometric curve, to the variance in the percept,  $\hat{s}$ . The third equation restates the definition of choice covariance, except for the scaling factor,  $\kappa(Z)$ , which will be constant for most practical purposes, and is described in detail in the [Methods](#) ([eq. 46](#)). Hence, in our full model of the task, we are able to predict both the psychometric sensitivity and the individual neurons’ choice signals from the first and second-order statistics of the neural responses. Specifically, by combining the characteristic equations for the linear readout and the decision policy, we obtain

$$1 = \bar{\mathbf{b}}^\top \mathbf{a}, \quad (7)$$

$$Z^2 = \mathbf{a}^\top \bar{\bar{\mathbf{C}}} \mathbf{a} + \sigma_d^2, \quad (8)$$

$$\mathbf{d}(t) = \kappa(Z) \bar{\mathbf{C}}(t) \mathbf{a}. \quad (9)$$

Importantly, since these equations deal with integrated versions of the raw neural signals, they depend on both the readout time window,  $w$ , and the extraction time,  $t_R$ .

We note that the choice covariance equation (eq. 9) can also be derived in a simpler, time-averaged form. Let  $\bar{d}_i$  be the time-integrated version of  $d_i(t)$ , using the readout's temporal parameters ( $w$ ,  $t_R$ ). Then, eq. 9 becomes

$$\bar{\mathbf{d}} = \kappa(Z) \bar{\mathbf{C}} \mathbf{a}, \quad (10)$$

which provides the linear covariance between each neuron's spike count  $\bar{r}_i$  on the trial, and the animal's choice. This is essentially the relationship already revealed by Haefner et al. (2013) [19], that choice probabilities are related to readout weights through the noise covariance matrix. The simpler linear measure of choice covariance, used in this article, allows us (1) to get rid of some non-linearities inherent to the choice probability formulation, and (2) to easily extend the interpretation of choice signals in the time domain, with eq. 9.

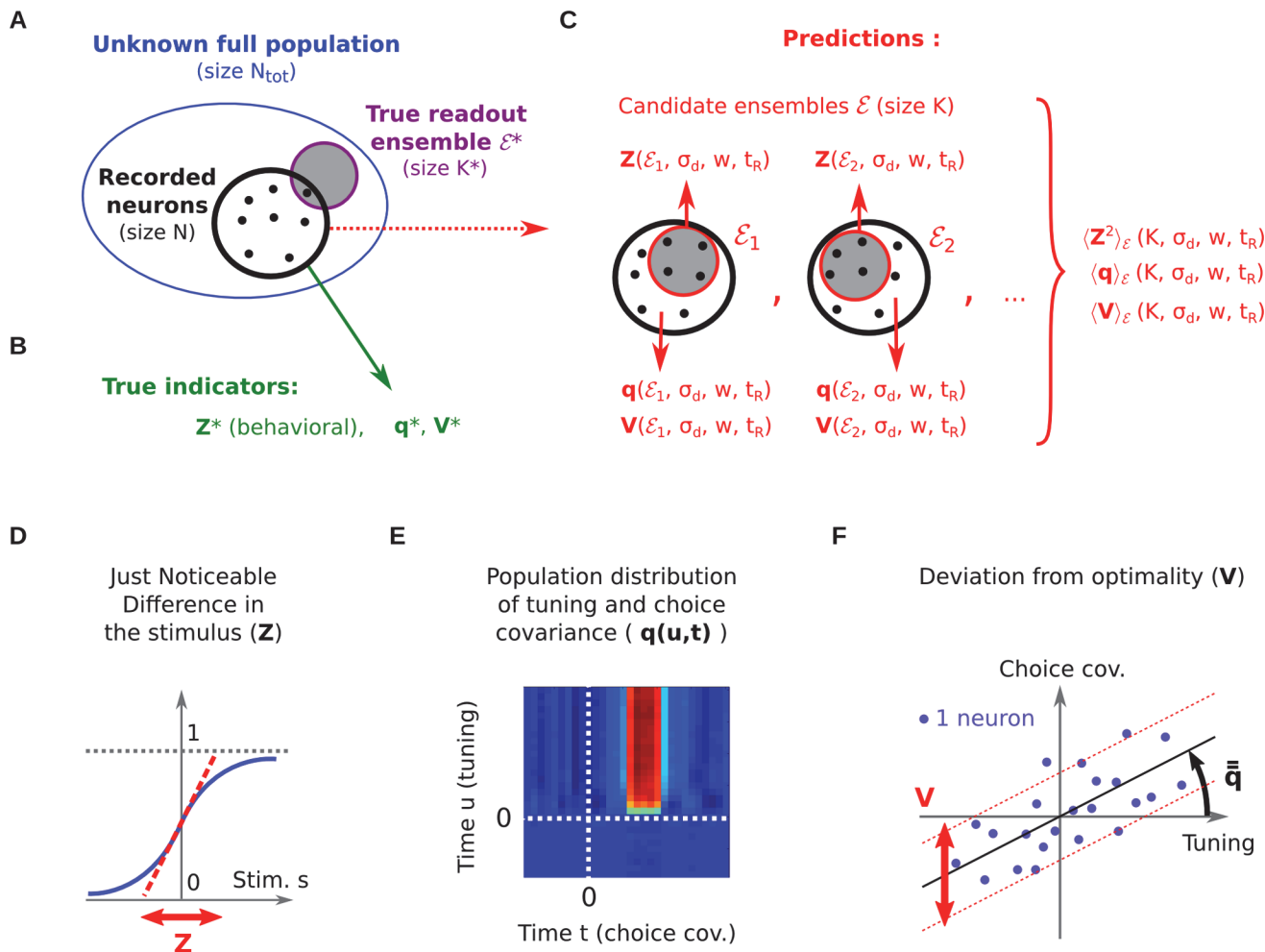
## Estimating the parameters of sensory integration

Equations 7–9 describe the analytical link between measures of neural response to the stimulus ( $b_i$  and  $C_{ij}$ ) and measures related to the animal's percept ( $Z$  and  $d_i$ ), based on the model's readout parameters ( $\mathbf{a}$ ,  $w$ ,  $t_R$ , and  $\sigma_d$ ). This naturally raises the reverse question: can we estimate the parameters of the standard model ( $\mathbf{a}$ ,  $w$ ,  $t_R$ , and  $\sigma_d$ ) from actual measurements? From here on, we will denote the *true* (and unknown) values of these parameters, i.e., the values used in the animal's actual percept formation, with a star ( $\mathbf{a}^*$ ,  $w^*$ ,  $t_R^*$ , and  $\sigma_d^*$ ).

As mentioned in the introduction, our primary interest concerns the trade-off between the time scale  $w^*$  of integration, and the size  $K^*$  of the functional population which conveys the animal's percept to downstream areas. Thus, we assume that the animal's percept is constructed from a specific sub-ensemble  $\mathcal{E}^*$  of neurons, of size  $K^*$  (Fig. 3A). Neurons inside  $\mathcal{E}^*$  correspond to nonzero entries in the readout vector  $\mathbf{a}^*$ , while neurons outside  $\mathcal{E}^*$  have zero entries. Since only a subset of neurons within a cortical area will project to a downstream area, we can generally assume that  $K^* < N_{\text{tot}}$ .

Naturally, all those parameters are not measurable experimentally. For any candidate set of parameters,  $\mathbf{a}$ ,  $w$ ,  $t_R$ , and  $\sigma_d$ , the characteristic equations 7–9 lead to *predictions* for  $Z$  and  $d_i(t)$  (note the absence of star when referring to predictions). In turn, the experimenter *can* measure the animal's actual choice  $c^*$  on each trial, from which they can estimate the JND  $Z^*$ , and the CC curves  $d_i^*(t)$  for all recorded neurons. In the next three sections, we study whether this information is sufficient to retrieve the true readout parameters, depending on the amount of data available.

In the ideal scenario where all neurons in the population are recorded simultaneously,  $N = N_{\text{tot}}$ , all parameters can be retrieved exactly (Case 1). In most experimental recordings, however, we only measure the activities of a small subset of that population (Fig. 3A). If this subset is representative of the full population, we may want to retrieve the readout parameters through extrapolation. Unfortunately, any such extrapolation is fraught with additional assumptions—whether implicit or explicit—as it requires to replace the missing data with some form of (generative) model. In Case 2, we impose a generative model for the readout vector  $\mathbf{a}$ . Coupled with a statistical principle, it allows us to estimate the true size  $K^*$  of the readout ensemble, provided that the number of neurons recorded simultaneously,  $N$ , is larger:  $N > K^*$ . In Case 3, we study the scenario in which  $N \leq K^*$ . Here, we need to assume a generative model for the neural activities themselves. Since the noise covariance structure assumed by that model exerts a strong influence on the predicted JND and CC curves, a direct inference of the readout scales becomes impossible.



**Fig 3. Statistical recovery of readout parameters: method.** (A) The full population (of size  $N_{\text{tot}}$ ) and the true readout ensemble  $\mathcal{E}^*$  (of size  $K^*$ ), are not fully measured. Only subsets of  $N$  neurons are recorded simultaneously from the population. (B) True measures for the three statistical indicators: the animal's psychometric JND  $\mathbf{Z}^*$ , plus indicators  $\mathbf{q}^*(u, t)$  and  $\mathbf{V}^*$  that summarize the distribution of the recorded neurons' CC curves  $d_i^*(t)$ . (C) A large number of neural ensembles  $\mathcal{E}$  of size  $K$  are randomly selected from the experimental pool and proposed as the candidate readout ensemble. This yields model-based predictions for the indicators as a function of the proposed readout parameters ( $K, \sigma_d, w, t_R$ ). (D-F) The three statistical indicators considered (see text for details).

doi:10.1371/journal.pcbi.1004082.g003

### Case 1: all cells recorded

If all neurons in the population have been recorded, with a sufficient amount of trials to estimate the complete covariance structure of the population, then the only unknown quantities in [eq. 7–9](#) are the readout parameters  $w, t_R$  and  $\mathbf{a}$ , and the decision noise  $\sigma_d$ . For fixed parameters  $w$  and  $t_R$ , [eq. 7](#) and [9](#) impose linear constraints on vector  $\mathbf{a}$ . These constraints are generally over-complete, since  $\mathbf{a}$  is  $N_{\text{tot}}$ -dimensional, while each time  $t$  in [eq. 9](#) provides  $N_{\text{tot}}$  additional linear constraints. Thus, in general, a solution  $\mathbf{a}$  will only exist if one has targeted the true parameters  $w^*$  and  $t_R^*$ , and it will then be unique. (If no choice of the readout parameters approximately fulfills the characteristic equations, we would have to conclude that the linear readout model is fundamentally wrong.) In practice, we can find the best solution to the characteristic

equations by simply combining them and then minimizing the following mean-square error:

$$L(w, t_R, \mathbf{a}, \sigma_d) = (1 - \bar{\mathbf{b}}^\top \mathbf{a})^2 + \lambda(Z^{*2} - \sigma_d^2 - \mathbf{a}^\top \bar{\mathbf{C}} \mathbf{a})^2 + \mu \int dt \|\mathbf{d}^*(t) - \kappa(Z^*) \bar{\mathbf{C}}(t) \mathbf{a}\|^2, \quad (11)$$

where the parameters  $\lambda$  and  $\mu$  trade off the importance of the errors in the different characteristic equations. Note that the loss function  $L$  depends not only on the readout weights  $\mathbf{a}$  and the decision noise  $\sigma_d$ , but also on the parameters  $w$  and  $t_R$ , both of which enter all the time integrations that are denoted by an overhead bar. Once vector  $\mathbf{a}^*$  is estimated, the readout ensemble  $\mathcal{E}^*$  will correspond to the set of neurons with nonzero readout weights.

## Case 2: more than $K^*$ cells recorded

Unfortunately, measuring the neural activity of a full population is essentially impossible, although optogenetic techniques are coming ever closer to this goal [24–26]. Nevertheless, if the activity patterns of the recorded cells are statistically similar to those of the readout ensemble, and if the number of simultaneously recorded cells exceeds the number of cells in the readout ensemble, we can still retrieve the readout parameters by making specific assumptions about the true readout vector  $\mathbf{a}^*$ .

**A statistical approach.** Our central assumption will be that the system uses the principle of *restricted optimality*: we assume that the readout vector  $\mathbf{a}^*$  extracts as much information as possible from the neurons within the readout ensemble,  $\mathcal{E}^*$ , and no information from all other neurons. Since most of the neurons contributing to the readout were probably not recorded, we cannot directly estimate the true readout vector,  $\mathbf{a}^*$ . However, we can form candidate ensembles from the recorded pool of neurons,  $\mathcal{E}$ , compute their optimal readout vector,  $\mathbf{a}_r(\mathcal{E})$ , and then test to what extent these candidate ensembles can predict the JND or the CC curves (Fig. 3C). By changing the size of the candidate ensembles,  $K$ , we can in turn infer the number of neurons involved in the readout.

For an arbitrary candidate ensemble  $\mathcal{E}$ , we can express its optimal readout vector,  $\mathbf{a}_r(\mathcal{E}) := \{a_i\}_{i \in \mathcal{E}}$ , on the basis of the neurons' tuning and noise covariance, through a formula known as Fisher's linear discriminant [27]:

$$\mathbf{a}_r = \frac{(\bar{\mathbf{C}}_r)^{-1} \bar{\mathbf{b}}_r}{\bar{\mathbf{b}}_r^\top (\bar{\mathbf{C}}_r)^{-1} \bar{\mathbf{b}}_r}. \quad (12)$$

Here, the subscript  $r$  indicates that all quantities are only evaluated for the neurons within the ensemble  $\mathcal{E}$ . The remaining neurons in the population do not participate in the readout. The resulting readout vector verifies eq. 7, and minimizes the just noticeable difference  $Z$  under the given constraints. Specifically, by entering the optimal readout into eq. 8, we obtain a prediction for the JND (Fig. 3D),

$$Z^2 = \frac{1}{\bar{\mathbf{b}}_r^\top (\bar{\mathbf{C}}_r)^{-1} \bar{\mathbf{b}}_r} + \sigma_d^2. \quad (13)$$

As for CC signals, the statistical description eliminates any reference to neuron identities, so we can no longer work directly with eq. 9. Instead, we re-express this equation in terms of two population-wide indicators, that summarize the CC signals of the individual neurons. The first indicator assesses the population-wide link between a neuron's tuning at each time  $u$ , and its choice covariance at each time  $t$ . The second indicator measures the average deviation from

this link (see also [Methods](#)):

$$q(u, t) := \langle b_i(u) d_i(t) \rangle_i \quad (14)$$

$$V := \langle \bar{b}_i^2 \rangle_i \langle \bar{d}_i^2 \rangle_i - \bar{q}^2. \quad (15)$$

Here, the angular brackets  $\langle \cdot \rangle_i$  denote averaging over the full neural population—or, in practice, over a representative ensemble of neurons (see [Methods](#) on how to construct this from actual data).

Experimentally,  $q(u, t)$  is expected to be globally positive, because the tuning of a neuron is often found to be somewhat correlated with its choice signal [[11](#), [15](#)] ([Fig. 3E](#))—likely due to the fact that positively-tuned neurons contribute positively to stimulus estimation, and negatively-tuned neurons negatively. This correlation can be quantified under the assumption of restricted optimality (see [Methods](#)). The indicator  $q(u, t)$  has a simple interpretation which we will illustrate by focusing on its doubly time-integrated version,  $\bar{q} = \langle \bar{b}_i \bar{d}_i \rangle_i$ . When we seek to predict a neuron’s choice covariance  $\bar{d}_i$  from its tuning  $\bar{b}_i$ , then  $\bar{q}$  is the best regression coefficient ([Fig. 3F](#)), so that

$$\bar{d}_i = \frac{\bar{q}}{\langle \bar{b}_j^2 \rangle_j} \bar{b}_i + \xi_i.$$

The deviations from this prediction are indicated by  $\xi_i$ , whose variance in turn is measured by the indicator  $V$  ([Fig. 3F](#)). A similar relation holds for the time-dependent indicator  $q(u, t)$ .

We now seek readout parameters which provide the best fit to the indicators introduced above. We set a number of potential values for parameters  $K, w, t_R, \sigma_d$ , and we explore routinely all their possible combinations. For each tested value of the readout ensemble size,  $K$ , we repeatedly pick a random neural ensemble  $\mathcal{E}$  of size  $K$  from the pool of neurons recorded by the experimenter, and propose it as the source of the animal’s percept ([Fig. 3C](#)). Then, we compute the average indicators across ensembles of similar size (see [Methods](#)), which we will denote by  $\langle Z^2 \rangle_{\mathcal{E}}$ ,  $\langle q \rangle_{\mathcal{E}}$ , and  $\langle V \rangle_{\mathcal{E}}$ . Note that all of these indicators depend on the parameters  $w, t_R, K$ , and  $\sigma_d$ . Finally, we replace the loss function of Case 1 ([eq. 11](#)) by the following “statistical” loss function:

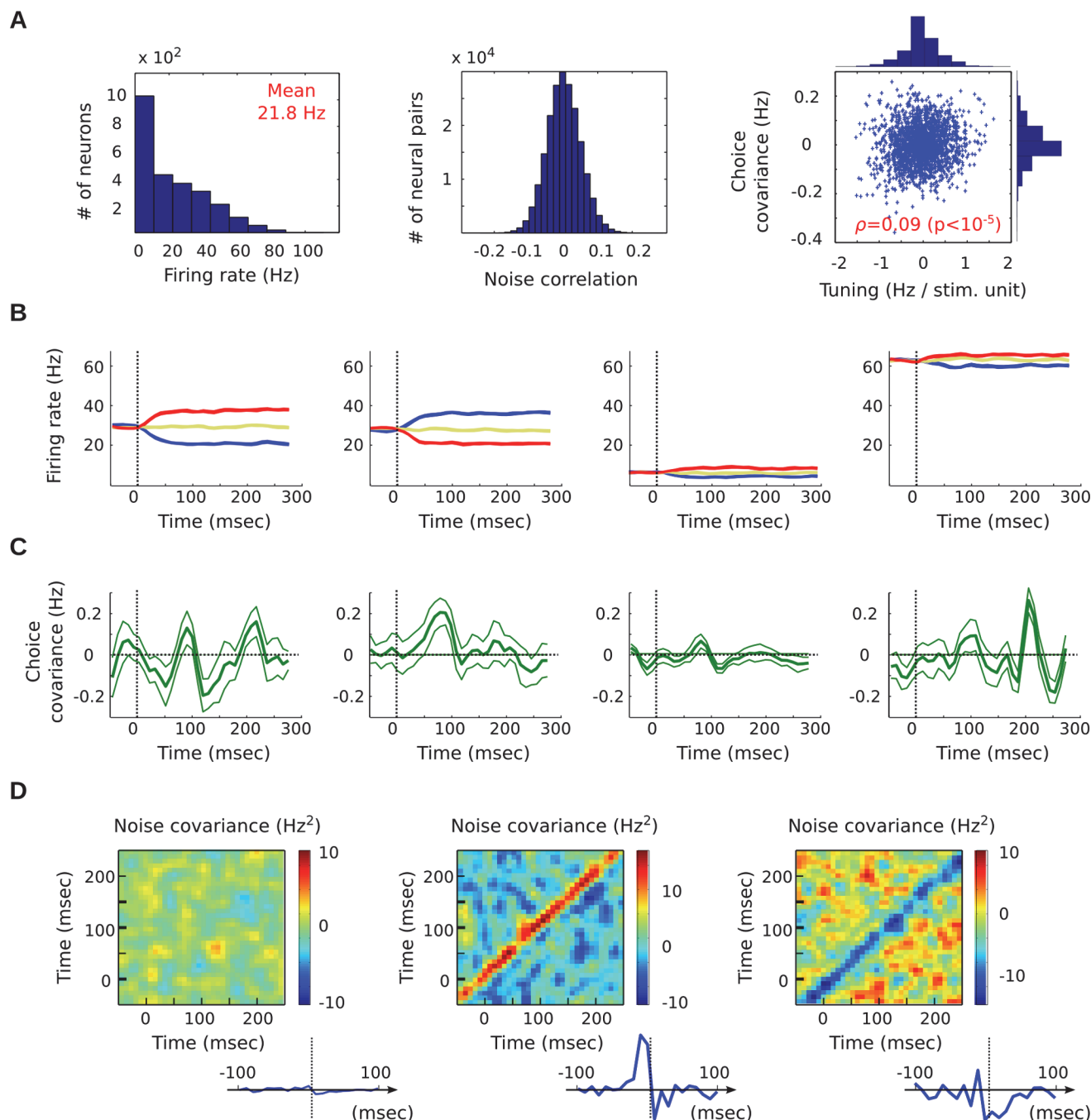
$$L(w, t_R, K, \sigma_d) = (Z^{*2} - \langle Z^2 \rangle_{\mathcal{E}})^2 + \lambda \int \int dt du (q^*(u, t) - \langle q(u, t) \rangle_{\mathcal{E}})^2 + \mu (V^* - \langle V \rangle_{\mathcal{E}})^2. \quad (16)$$

The minimum of the loss function then indicates what values of the readout parameters agree best with the recorded data.

**Network simulations and retrieval of readout parameters.** To validate these claims, we have tested our method on synthetic data—which are the only way to control the true parameters of integration, and thus to test our predictions. We implemented a recurrent neural network with  $N = 5000$  integrate-and-fire neurons that encodes some input stimulus  $s$  in the spiking activity of its neurons, and we built a perceptual readout from that network according to our model, with parameters  $K^* = 80$  neurons,  $w^* = 50$  ms,  $t_R^* = 100$  ms, and  $\sigma_d^* = 1$  stimulus units (see [Methods](#) for a description of the network, and supporting [S1 Text](#)).

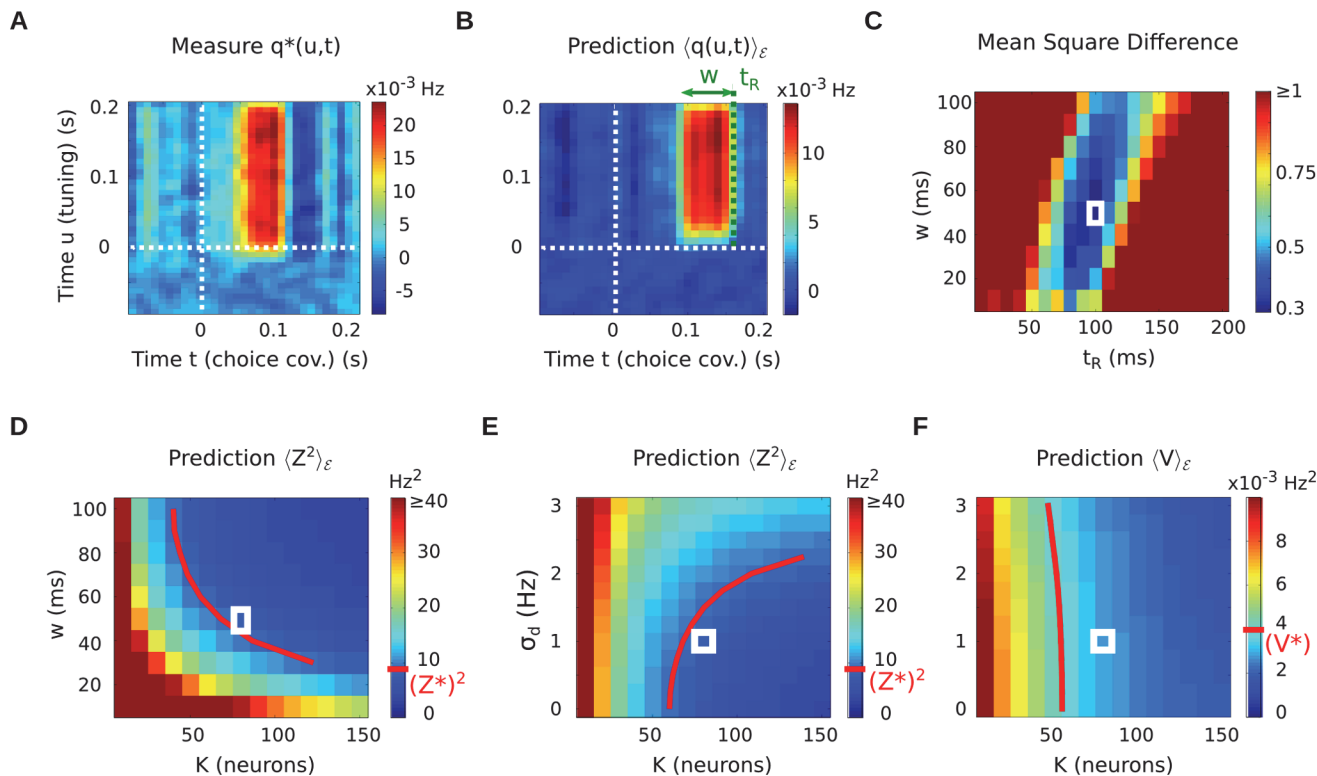
Then, as experimenters, we observed on every trial the perceptual report  $c^*$  and samples of network activity, from which we computed neural response statistics  $b_i(t)$  and  $C_{ij}(t, u)$ , the psychometric curve  $\psi^*(s)$ , and the neuron CC curves  $d_i^*(t)$  ([Fig. 4](#)). From these (partial) measures, we extracted the three population-wide indicators  $Z^*$ ,  $q^*(u, t)$  and  $V^*$ , and investigated whether the loss function from [eq. 16](#) allows us to recover the system’s true scales of perceptual integration ( $w^*, t_R^*, K^*, \sigma_d^*$ ).





**Fig 4. Simulated neural network used for testing the method.** (A) Spike count statistics amongst the population of 5000 neurons (spike counts over 400 msec, on  $3 \times 180$  stimulus repetitions). Note the weak, but significant, correlation between tuning ( $b_i$ ) and choice covariance ( $d_i$ ). (B) Sample PSTHs: the model neurons display varied firing rates, and tunings of different polarities. (C) Sample choice covariance curves for the same neurons as panel B (thin lines: bootstrap-based error bars). (D) Sample JPSTHs (noise correlations) for pairs of model neurons. Inset: corresponding cross-correlograms, obtained by projection along the diagonal. For better visibility, the curves in panels B-D were computed from a larger number of trials ( $3 \times 3000$ ) than used for the study itself ( $3 \times 180$ ), and time-averaged with a 10 msec Gaussian kernel.

doi:10.1371/journal.pcbi.1004082.g004



**Fig 5. Statistical recovery of readout parameters: structure of the results.** (A) Experimental indicator  $q^*(u, t)$ . Note the noisiness due to limited amounts of data. (B) Prediction  $\langle q(u, t) \rangle_{\varepsilon}$  for a given set of readout parameters ( $w, t_R, K, \sigma_d$ ). The temporal location of the CC signal is mostly governed by parameters  $w$  and  $t_R$ . (C) (Normalized) mean square error between measured and predicted  $q(u, t)$ , as a function of readout parameters ( $w, t_R$ ). The true values ( $w^*, t_R^*$ ) are indicated by a white square. (D) Predicted JND  $\langle Z^2 \rangle_{\varepsilon}$  as a function of  $K$  and  $w$ . (E) Predicted JND  $\langle Z^2 \rangle_{\varepsilon}$  as a function of  $K$  and  $\sigma_d$ . (F) Predicted deviation  $\langle V \rangle_{\varepsilon}$  as a function of  $K$  and  $\sigma_d$ . In panels C–F, the white square indicates the true (starred) value for the parameters being represented. The parameters not represented are always fixed at their true (starred) value. In panels D–F, the red curve marks the intersection of the predicted indicator with its measured value. All indicators have units derived from Hz, owing to stimulus  $s$  being itself a frequency (see [Methods](#)).

doi:10.1371/journal.pcbi.1004082.g005

The results, summarized in [Fig 5](#), show that the recovery is indeed possible. Each indicator plays a specific role in recovering some of the parameters. First, indicator  $q(u, t)$  allows us to recover the temporal parameters of integration ( $w, t_R$ ). Indeed, it characterizes the time interval during which the population—as a whole—shows the strongest choice covariance ([Fig 5A](#)), and the bounds of this interval are essentially governed by parameters ( $w, t_R$ ) ([Fig 5B](#)). As a result, the match between true measure and prediction—second term in [eq. 16](#)—shows a clear optimum near the true values ( $w^*, t_R^*$ ) ([Fig 5C](#)). The bi-temporal structure of  $q(u, t)$  in [eq. 14](#), with time index  $u$  corresponding to the neurons' tuning  $b_i(u)$ , stabilizes the results by insuring that  $q(u, t)$  is globally positive.

Second, indicator  $Z$  allows us to target readouts with the correct 'overall amount' of integration, resulting in a JND compatible with the data. [Fig 5D](#) depicts the predicted value for  $Z$  as a function of  $w$  and  $K$ . The mark of the 'K- $w$  trade-off' is visible: higher sensitivity to stimulus can be achieved either through longer temporal integration ( $w$ ), or through larger readout ensembles ( $K$ ). Analytically, the JND  $Z$  depends on  $w$  because the covariance matrix  $\bar{C}$  will generally scale with  $w^{-1}$  (under mild assumptions, supporting [S1 Text](#)). The red curve marks the pairs ( $K, w$ ) for which the prediction matches the measured JND  $Z^*$ —thereby minimizing the first loss term in [eq. 16](#). The true parameters ( $K^*, w^*$ ) lie along that curve (white square in

Fig. 5D). Since  $w^*$  is recovered independently thanks to indicator  $q(u, t)$ , this in turn allows us to recover parameter  $K^*$ .

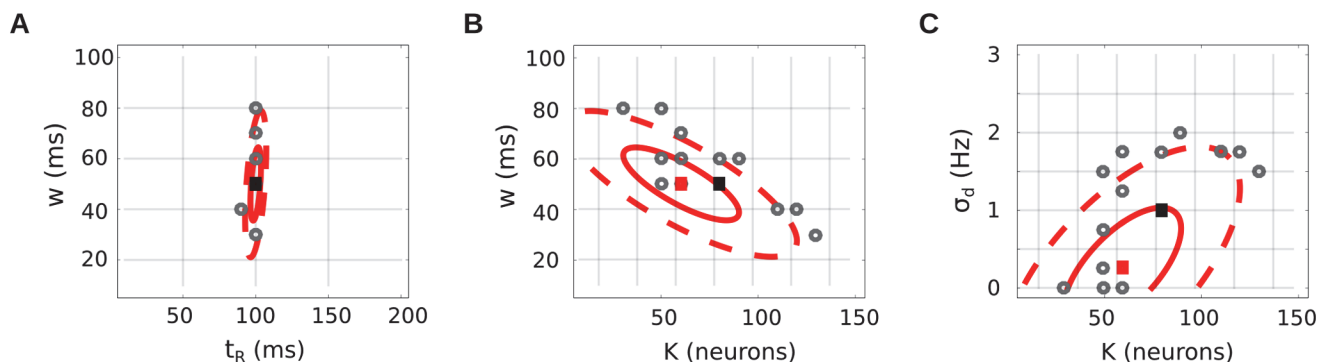
If sensory noise is the main source of error in the animal's judgments (meaning  $\sigma_d \simeq 0$  in the model), the two indicators  $q(u, t)$  and  $Z$  suffice to characterize the readout parameters. But in the general case, the observed JND  $Z^*$  can also be influenced by extraneous sources of noise in the animal's decision, and bias the comparison between  $Z^*$  and its prediction. To account for this potential effect, our model includes the decision-noise term  $\sigma_d$ . For a fixed value of  $w$ , the JND  $Z$  is influenced both by parameters  $K$  and  $\sigma_d$  (eq. 13, Fig. 5E). However, both parameters can be disentangled thanks to the third indicator  $V$ , which depends mostly on  $K$  (Fig. 5F).

The signification of  $V$  hinges on the following result, that was first shown in [19]: when the readout is *truly* optimal over the full population ( $K = N_{\text{tot}}$ ), then each neuron's choice covariance  $\bar{d}_i$  is simply proportional to its tuning  $\bar{b}_i$  (see Methods). Since the indicator  $V$  quantifies the deviations from perfect proportionality between  $\bar{b}_i$  and  $\bar{d}_i$  (eq. 15, Fig. 3F), it becomes a marker of the readout's *global* optimality, and decreases to zero as  $K$  grows to large populations. At the same time, the dependency of  $V$  on parameter  $\sigma_d$  is minimal, and limited to the influence of the scaling factor  $\kappa(Z)$  in eq. 9 (see Methods).

When minimizing the loss function in eq. 16, we impose the joint fit of the three indicators  $Z$ ,  $q(u, t)$  and  $V$ . Following the explanations above, this will be obtained for parameters close to the true values ( $w^*, t_R^*, K^*, \sigma_d^*$ ). In our simulation, the minimum was achieved for the following values:  $w = 50$  msec,  $t_R = 100$  msec,  $K = 60$  neurons,  $\sigma_d = 0.25$  stimulus units (with the following levels of discretization: 10 neurons for  $K$ , 0.25 stimulus units for  $\sigma_d$ , 10 msec for  $w$  and  $t_R$ ).

The best fit parameters are represented in Fig. 6, along with bootstrap confidence intervals derived from 14 resamplings of our original data. The temporal parameters ( $w, t_R$ ) are recovered with good precision (panel A). Conversely, parameters  $K$  and  $\sigma_d$  are somewhat underestimated (panels B and C) compared to their true values (black square). Indeed, the values of  $K$  and  $\sigma_d$  are disentangled thanks to indicator  $V$  which, of the three indicators introduced, is the most subject to measurement noise. As a result, the match between  $V^*$  and its prediction  $V$  is not as precise as the other two: see Fig. 5F. Nevertheless, the true values are rather close to the final estimates, lying within the 1-standard deviation confidence region (Fig. 6C).

Importantly, only a reasonable amount of data is required to produce these estimates. Network activity was monitored on 15 independent runs, each run consisting of 180 repetitions



**Fig 6. Statistical recovery of readout parameters: best fit parameters.** Efficiency of the inference method, applied to our simulated LIF network. The three panels show different 2d projections of the underlying 4d parameter space: ( $t_R, w$ ) plane (A), ( $K, w$ ) plane (B), ( $K, \sigma_d$ ) plane (C). Black square: true parameters ( $K^*, w^*, t_R^*, \sigma_d^*$ ) used to produce the data. Red square: best fit parameters ( $K, w, t_R, \sigma_d$ ), achieving the minimum of the loss function in eq. 16. Gray points: best fit parameters for 14 (bootstrap) resamplings of the original trials (some points are superimposed, due to the finite grid of tested parameters). Red ellipses: corresponding confidence intervals, as the 1- and 2- standard deviation of the bootstrap resamplings.

doi:10.1371/journal.pcbi.1004082.g006

for each of the 3 stimuli. On each run, a different set of  $N = 170$  random neurons were simultaneously recorded—out of a total population of  $N_{\text{tot}} = 5000$ . As a result, (i) individual neuron statistics such as  $C_{ij}(t, u)$  or  $d_i^*(t)$  display an important amount of measurement noise, (ii) population statistics such as indicator  $V$  are computed from relatively few neurons  $i$ . Numerically, this noisiness introduces a number of biases in the above indicators, such as overfitting, which require counteracting with specific corrections (see [Methods](#) and supplementary material for details). Naturally, the width of the confidence intervals in [Fig. 6](#) is directly related to the amount of data available.

In conclusion, if the data conform to a number of hypotheses (optimal linear readout from a neural ensemble typical of the full population, and smaller than the recording pool size), then it is possible to estimate the underlying readout's parameters, from a plausible amount of experimental samples.

### Case 3: less than $K^*$ cells recorded

By construction, the method presented in Case 2 can only test ensemble sizes  $K$  smaller than  $N$ , the number of neurons recorded simultaneously by the experimenter. If  $N$  is smaller than the true size  $K^*$ , the method will provide biased estimates. In current-day experiments,  $N$  can range from a few tens to a few hundred neurons. While it is not excluded that typical readout sizes  $K^*$  be of that magnitude in real neural populations (as suggested, e.g., by [\[8\]](#)), it is also possible that they are larger. In this case, the only way to estimate the readout parameters is to make specific assumptions about the nature of the full population activity. In turn, the extrapolated results will depend on these assumptions.

**Singular value analysis of the linear readout.** To investigate the underlying issues, and to explain why there is no “natural” extrapolation, we will study how the indicators  $Z$ ,  $\bar{q}$ , and  $V$  defined above evolve as a function of the number of neurons  $K$  used for the readout. For simplicity, we assume a fixed choice of  $(w, t_R)$  and focus on the time-integrated neural activities  $\bar{r}_i$  ([eq. 1](#)). We also suppose that the decision noise  $\sigma_d \simeq 0$  is negligible. Finally, we consider alternative definitions for the indicators  $Z$  and  $\bar{q}$  that simplify the following analysis. We define

$$Y := \frac{\sigma_s^2}{Z^2 + \sigma_s^2}, \quad (17)$$

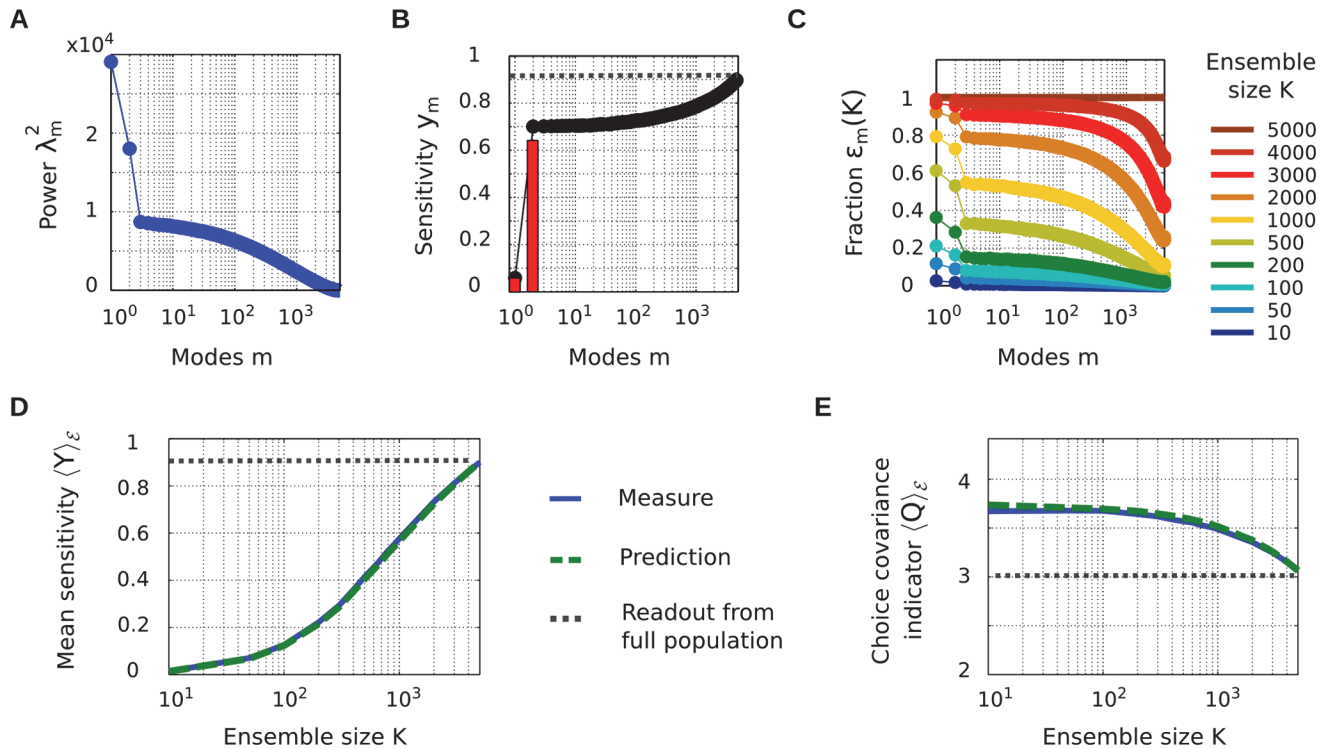
$$Q := \frac{1}{\kappa(Z)} \bar{q} + \sigma_s^2 \langle \bar{b}_i^2 \rangle_i, \quad (18)$$

where  $\sigma_s^2$  is the variance of the tested stimuli, i.e.,  $\sigma_s^2 := E[s^2] - E[s]^2$ . The *sensitivity*  $Y$  is simply an inverse reparametrization of the JND,  $Z$ . More specifically,  $Y$  is the ratio between the signal-related variance and the total variance (see [Methods](#)), which grows from zero (if  $Z = \infty$ ) to one (if  $Z = 0$ ) as the readout's sensitivity to the stimulus increases. As for  $Q$ , it is simply a convenient linear rescaling of  $\bar{q}$ .

Then, we re-express the population activity through a singular value decomposition (SVD) (see [Methods](#) for details). Specifically, we write the time-averaged activity of neuron  $i$  for the  $q$ -th presentation of stimulus  $s$  as

$$\bar{r}_i^{sq} = \bar{r}_i^0 + \sum_{m=1}^M \lambda_m u_i^m v_m^{sq}, \quad (19)$$

where  $\bar{r}_i^0$  is the trial-averaged activity of each neuron. This decomposition is best interpreted as a change of variables, which re-expresses the neural activities  $\{\bar{r}_i\}_{i=1 \dots N_{\text{tot}}}$  in terms of a new set



**Fig 7. Readout properties as a function of ensemble size  $K$ .** (A) The SVD decomposes population activity into a number of modes  $m$  with decreasing powers  $\lambda_m^2$ . (B) Each mode  $m$  has a sensitivity to stimulus,  $y_m$ . Red bars: individual sensitivities, black dots: cumulative distribution. (C) The fractions  $\epsilon_m(K)$  describe the “proportion” of each mode which can be observed, on average, in random neural ensembles of size  $K$ . They are a function of the SVD decomposition, but bear no analytical expression in general. (D) Mean sensitivity from neural ensembles of size  $K$ , empirically measured (blue) and predicted with [eq. 21](#) (green). The dashed black line indicates the optimal sensitivity, for a readout from the full population. (E) Same for the CC indicator  $Q$  ([eq. 22](#)). All panels computed from the spike counts  $\bar{r}$  of the 5000 simulated neurons, over the first 10 msec after stimulus presentation, on  $3 \times 3000$  recording trials (without correcting for measurement errors owing to the large dimensionality and limited number of trials).

doi:10.1371/journal.pcbi.1004082.g007

of variables,  $\{v_m\}_{m=1 \dots M}$ , which we will call the activations of the population’s *modes*. These modes can be viewed as the underlying “patterns of activity” that shape the population on each trial. Each mode  $m$  has a strength  $\lambda_m > 0$  which describes the mode’s overall impact on population activity. We assume  $\lambda_1 \geq \dots \geq \lambda_M$ , so we progressively include modes with lower strengths ([Fig. 7A](#)). The vector  $\mathbf{u}^m$  is the “shape” of mode  $m$  and describes how the mode affects the individual neurons. Finally,  $v_m^{sq}$  is the mode’s activation variable, which takes a different (random) value on every trial  $q$  for a given stimulus  $s$ . The number of modes  $M$  is the intrinsic dimensionality of the neural population’s activity. In real populations we may expect  $M < N_{\text{tot}}$  because neural activities are largely correlated.

Since the singular value decomposition is simply a linear coordinate transform, we can redefine all quantities with respect to the activity modes. Of particular interest is the *sensitivity* of each mode, which is the square of its respective tuning parameter, or (see [Methods](#))

$$y_m = \left[ \sigma_s \sum_{i=1}^{N_{\text{tot}}} \lambda_m^{-1} u_i^m \bar{b}_i \right]^2. \quad (20)$$

If the readout vector  $\mathbf{a}$  is chosen optimally over the full population, the resulting percept’s sensitivity will be a simple sum over the modes:  $Y^{\text{tot}} = \sum_m y_m$ .



The mode sensitivities and their cumulative sum for the simulated network above are shown in Fig. 7B. Note the presence of a “dominant” mode for the sensitivity. This seems to be a rather systematic effect, which arises because the definition of total covariance (Methods, eq. 50) favors the appearance of a mode almost collinear with  $\bar{b}$ . Even so, this dominant mode accounted only for 71% of the population’s total sensitivity, so the residual sensitivity in the other modes is generally not negligible.

#### Sensitivity and choice covariance as a function of the size $K$ of the readout ensemble.

However, we wish to study the more general case where the readout is built from sub-ensembles of size  $K$ . In such a case, not all modes are equally observable, and we rather need to introduce a set of fractions,  $\{\epsilon_m(K)\}_{m=1 \dots M}$ , that express to what extent each mode  $m$  is “observed”, on average, in sub-ensembles  $\mathcal{E}$  of size  $K$  (see Methods for a precise definition). Modes with larger power  $\lambda_m$  tend to be observed more, so  $\epsilon_m(K)$  globally decreases with  $m$ . Conversely,  $\epsilon_m(K)$  naturally increases with  $K$ . For the full population,  $\epsilon_m(N_{\text{tot}}) = 1$  for all modes  $m$ , meaning that all modes are fully observed (see Fig. 7C; here, the mode observation fractions were empirically computed by averaging over random neural sub-ensembles). Using these fractions, we can analytically approximate the values of  $Y$  and  $Q$  which are expected, on average, if the readout is based on ensembles of size  $K$ :

$$\langle Y \rangle_{\mathcal{E}}(K) \simeq \sum_{m=1}^M \epsilon_m(K) y_m, \quad (21)$$

$$\langle Q \rangle_{\mathcal{E}}(K) \simeq \frac{1}{\langle Y \rangle_{\mathcal{E}}(K)} \sum_{m=1}^M \epsilon_m(K) y_m \lambda_m^2. \quad (22)$$

Thus, the sensitivity  $\langle Y \rangle_{\mathcal{E}}$  grows with  $K$  as mode sensitivities  $y_m$  are progressively revealed by the fractions  $\epsilon_m(K)$ . The sensitivity reaches its maximum value,  $Y(N_{\text{tot}}) = Y^{\text{tot}}$ , when  $\epsilon_m(K) = 1$  for all modes  $m$  with a nonzero  $y_m$  (Fig. 7D). Conversely,  $\langle Q \rangle_{\mathcal{E}}$  decreases with  $K$ . Indeed, it can be viewed as an average of the squared powers  $\{\lambda_m^2\}$ , each mode  $m$  contributing with a weight  $\epsilon_m(K) y_m$ . As  $\epsilon_m(K)$  progressively reveals modes with lower power  $\lambda_m$ , this average power is expected to decrease with  $K$ . Again, the minimum value is reached when all nonzero  $y_m$  are revealed (Fig. 7E).

The results for the simulated network in Fig. 7D–E illustrate that the approximations leading to eq. 21–22 are well justified in practice. As for the third indicator used in Case 2,  $V$ , it can also be expressed in the SVD basis (see Methods). However, being a second-order variance term, its approximation based solely on the average fractions  $\{\epsilon_m(K)\}$ , as in eq. 21–22, is generally poor.

**The extrapolation problem revisited.** What do these results imply in terms of extrapolation to larger neural ensembles than those recorded by the experimenter? Arguably, eq. 21–22 constitute an interesting basis for principled extrapolations to larger sizes  $K$ . These equations show that the evolution of  $Y$  and  $Q$  in growing ensembles of size  $K$  is mostly related to the interplay between the modes’ sensitivity spectrum  $\{y_m\}$  and their power spectrum  $\{\lambda_m\}$ . (Empirically, the observation fractions  $\{\epsilon_m(K)\}$  seem primarily governed by the decay rate of  $\{\lambda_m\}$ , although the analytical link between the two remains elusive.) However, note that the spectra  $\{y_m\}$ ,  $\{\lambda_m\}$  and  $\{\epsilon_m(K)\}$  are generally not accessible to the experimenter—this would precisely require to have recorded at least  $N > M$  neurons, and potentially the whole neural population if  $M = N_{\text{tot}}$ .

To extrapolate sensitivity  $\langle Y \rangle_{\mathcal{E}}(K)$  in ensembles of size  $K$  larger than those monitored, one must (implicitly or explicitly) assume a model for  $\{\lambda_m\}$  and  $\{y_m\}$ —which amounts to characterizing the relative embedding of signal and noise in the full population [28]. A number of



reasonable heuristics could be used to produce such a model. For example, one may assume a simple distribution for  $\{\lambda_m\}$ , such as a power law, and estimate its parameters from recorded data. Alternatively, it is often assumed that the noise covariance matrix is “smooth” with respect to the signal covariance matrix, so that the former can be predicted on the basis of the latter [19, 29]. Finally, the extrapolation could rely on more specific assumptions about how neural activities evolve, e.g., through linear dynamics with additive noise [30]. In all cases, the additional assumptions impose (implicit) constraints on the structure of the spectra  $\{\lambda_m\}$  and  $\{y_m\}$ .

However, most likely, any chosen model will be (1) difficult to fit rigorously on the basis of experimental data, (2) subject to pathological situations when extrapolations fail to produce the correct predictions. For example, one can imagine scenarios in which the most sensitive modes (those with highest  $y_m$ ) correspond to very local circuits of neurons, independent from the rest of the population, and thus invisible to the experimenter (see also [19]). Another pathological situation could be a neural network specifically designed to dispatch information non-redundantly across the full population [31, 32], resulting in a few ‘global’ modes of activity with very large SNR—meaning high  $y_m$  and low  $\lambda_m$ . As a result, extrapolation to neural populations larger than those recorded is never trivial, and always subject to some *a priori* assumptions. The most judicious assumptions, and the extent to which they are justified, will depend on each specific context.

## Discussion

We have proposed a framework to interpret sensitivity and choice signals in a standard model of perceptual decision-making. Our study describes percept formation within a full sensory population, and proposes novel methods to estimate its characteristic readout scales on the basis of realistic samples of experimental data. Here, we briefly discuss the underlying assumptions and their restrictions, the possibility of further extensions, and the applicability to real data.

### The linear readout assumption

The readout model (eq. 2) used to analyze sensitivity and choice signals is an installment of the “standard”, feed-forward model of percept formation [17, 19]. As such it makes a number of hypotheses which should be understood when applying our methods to real experimental data. First, it assumes that the percept  $\hat{s}$  is built linearly from the activities of the neurons—a common assumption which greatly simplifies the overall formalism (but see, e.g., [33] for a recent example of nonlinear decoding). Even if the real percept formation departs from linearity, fitting a linear model will most likely retain meaningful estimates for the coarse information (temporal scales, number of neurons involved) that we seek to estimate in our work.

More precisely, the model in eq. 1–2 assumes that spikes are integrated using a kernel that is separable across neurons and time, that is  $A_i(t) = a_i h(t/w)/w$ . Theory does not prevent us from studying a more general integration, where each neuron  $i$  contributes with a different time course  $A_i(t)$ . The readout’s characteristic equations are derived equally well in that case. Rather, assuming a separable form reflects our intuition that the time scale of integration is somewhat uniform across the population. This time scale,  $w$ , is then the one crucial parameter of the integration kernel. Although the shape  $h(t)$  of the kernel could also be fit from data in theory, it seems more fruitful to assume a simple shape from the start. We assumed a classic square kernel in our applications. Other shapes may be more plausible biologically, such as a decreasing exponential mimicking synaptic integration by downstream neurons. However, given that our goal is to estimate the (coarse) time scale of percept formation, our method will likely be robust

to various simple choices for  $h$ . As a simple example, we tested our method, assuming a square kernel, on data produced by an exponential readout kernel, and still recovered the correct parameters  $w$ ,  $t_R$  and  $K$  (data not shown).

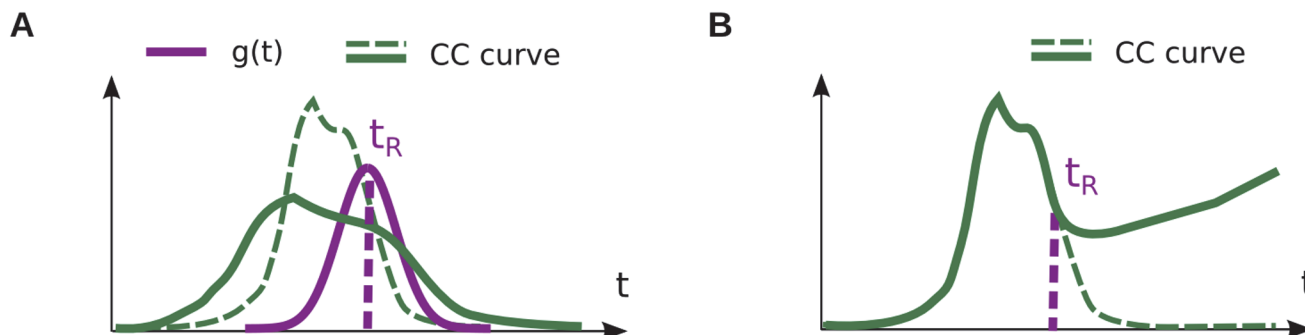
Through the process of integration across time and neurons, each instant in time could be associated to an “ongoing percept”, i.e., the animal’s estimate of stimulus value at current time. In our model, the animal’s estimate at time  $t_R$  serves as the basis for its behavioral report (Fig. 2A), and we designate this single number  $\hat{s}$  as the “percept”. A second strong assumption of our model is that this perceptual readout occurs at the same time  $t_R$  on every stimulus presentation. In reality, there is indirect evidence that  $t_R$  could vary from trial to trial, as suggested by the subjects’ varying reaction times (RT) when they are allowed to react freely [34, 35]. In such tasks, we expect the variations in  $t_R$  to be moderate—because subjects generally react as fast as they can—and we may even try to correct for fluctuations across trials by measuring RTs. On the other hand, when subjects are forced to wait for a long period of time before responding, there is room for ample variations in  $t_R$  from trial to trial, and the model presented above may become insufficient.

As a first step towards addressing this question, we derived a more general version of the characteristic equations 4–6 assuming that  $t_R$  in eq. 1 is itself a random variable, drawn on each trial following some probability distribution  $g(t)$  (supporting S1 Text). The main impact of this modification is on CC curves, which become broader and flatter; essentially, the resulting curve resembles a convolution of the deterministic CC curve by  $g(t)$  (Fig. 8A). This means that if a behavioral task is built such that  $t_R$  can display strong variations from trial to trial, the methods introduced above will produce biased estimates. In theory, this issue could be resolved by adding an additional parameter in the analysis, to describe  $g(t)$  (see supporting S1 Text).

## The decision model

The linear readout provides a percept  $\hat{s}$  on every trial. In principle, behavioral experiments could be set up such that the subject directly reports this percept, so that  $c = \hat{s}$ . Such experiments could be treated completely without a decision model. However, almost all experiments that have been studied in the past involve a more indirect report of the animal’s percept. In these cases, some assumptions about how the percept is transformed into the behavioral report  $c$  need to be made.

In the choice of a decision model, we have followed the logic of the classic random dot motion discrimination task [3], in which a monkey observes a set of randomly moving dots whose



**Fig 8. Discussion.** (A) If the extraction time  $t_R$  varies strongly from trial to trial (with density  $g(t)$ ), it leads to a flattening of CC signals (thick green curve) compared to the case with deterministic  $t_R$  (dashed green curve). (B) If a choice-related signal feedbacks into sensory areas, it leads to an increase of CC signals (thick green curve) after the extraction time  $t_R$ , compared to the case without feedback (dashed green curve).

doi:10.1371/journal.pcbi.1004082.g008

overall motion is slightly biased towards the left ( $s < 0$  in our notations) or towards the right ( $s > 0$ ). The monkey must then press either of two buttons depending on its judgment of the overall movement direction. The simplest decision model assumes a fixed integration time window, additive noise on the percept,  $\hat{s}$ , and an optimal binary decision. A slightly more sophisticated model, the “integration-to-bound” model, assumes that the integration time is not fixed, but rather limited by a desired behavioral accuracy. This model requires variable readout windows, rather than the fixed readout window assumed here, and will require further investigation in the future.

In another classic task [2], the monkey must discriminate the frequencies  $s_1$  and  $s_2$  of two successive vibrating stimuli on their fingertip. They must press either of two buttons depending on whether they consider that  $s_1 > s_2$  or not. In this task, the optimal behavioral model would be  $c = H(\hat{s}_1 - \hat{s}_2)$ . In reality, however, the monkey needs to memorize  $s_1$  for a few seconds before  $s_2$  is presented, so potential effects of memory loss may also come into play (see e.g. [36] for a study of these problems).

More generally, behaving animals can display biases, lapses of attention, various exploratory and reward-maximization policies that lead to deviations from the optimal behavioral model. Choosing a relevant behavioral model is a connected problem that cannot be addressed here, and that will vary depending on the task and individual considered. For most tractable behavioral models, the predicted sensitivities and choice signals will ultimately rely on the quantities introduced in this article.

## The feedforward assumption

Finally, the standard model assumes that percept formation is exclusively feed-forward. The activities  $r_i(t)$  of the sensory neurons are integrated to give rise to the percept  $\hat{s}$  and the animal’s choice  $c$ , yet the formation of this decision does not affect sensory neurons in return. Recent evidence suggests that reality is more complex. By looking at the temporal evolution of CP signals in V2 neurons during a depth discrimination task, Nienborg and Cumming (2009) evidenced dynamics which are best explained by a top-down signal, biasing the activity of the neurons on each trial after the choice is formed [20]. In our notations, the population spikes  $r_i(t)$  would thus display a choice-dependent signal which kicks in on every trial after time  $t_R$ , resulting in CC signals that deviate from their prediction in the absence of feedback (Fig. 8B).

What descriptive power does our model retain, if such top-down effects are strong? The answer depends on the nature of the putative feedback. If the feedback depends linearly on percept  $\hat{s}$  (and thus, on the spike trains), its effects are fully encompassed in our model. Indeed, this feedback signal will then be totally captured by the neurons’ linear covariance structure  $C_{ij}(t, u)$ , so that our predictions will naturally take it into account. On the other hand, if the feedback depends directly on the choice  $c$ —which displays a nonlinear, “all-or-none” dependency on  $\hat{s}$ —then it will not be captured by our model, and lead to possible biases. Even so, our model would still apply if percept and decision were essentially uncoupled before the putative extraction time  $t_R$ , in which case one could simply compare true and predicted CC signals up to (candidate) time  $t_R$  (see Fig. 8B).

## Undersampled neural populations

In most real-life situations, experimenters only have access to samples from a large, unknown population, so they must resort to a statistical description of readout vector  $\mathbf{a}$ . Our solution relies on an assumption of restricted optimality, based on Fisher’s linear discriminant formula (eq. 12). By assuming that readout is made optimally from some unknown neural ensemble  $\mathcal{E}$ ,

we reformulated the problem of characterizing  $\mathbf{a}$  in that of characterizing  $\mathcal{E}$ , and could in turn exploit the characteristic equations 4–6 statistically.

In real experiments, the true readout profile  $\mathbf{a}$  may not match this description: most vectors  $\mathbf{a}$  do not implement optimal readout from a sub-ensemble. This potential discrepancy from the true readout is inescapable, once we start representing  $\mathbf{a}$  through a statistical model. However, note that our model uses *two* distinct sources of non-optimality: (1) the size  $K$  of the readout ensemble, which can be much smaller than the full population, and (2) the decision noise  $\sigma_d$ , which adds a ‘global’ non-optimality to the readout. Arguably, by combining both factors, our chosen model for  $\mathbf{a}$  will be flexible enough to provide meaningful estimates when fit to real data.

At present, the main limitation is likely to be the size of ensembles of neurons that have been recorded simultaneously. Past work has often shown that small ensembles of neurons are completely sufficient to account for an animal’s behavior [3, 37]. However, there is an inherent trade-off between the number of neurons and the time scale of integration. One simple explanation for the small sizes of previous readout ensembles is that the true readout time scales used by subjects are much shorter. Unfortunately, as detailed above (Case 3), extrapolations from a finite-size recording onto the whole population always come at the price of strong additional assumptions.

However, as experimental techniques advance, and as the number of simultaneously recorded neurons reaches the number of neurons implied in the readout, we will eventually be able to directly infer the readout parameters from the data. In this case, our method can readily be tested on real data, and hopefully provide new insights into the nature of percept formation from populations of sensory neurons.

## Methods

The methods are organized as follows. First, we set our basic notations and definitions. Second, we derive the characteristic equations of the model, both for the linear part and decision part. Third, we detail the predictions in case of an optimal readout from some neural sub-ensemble  $\mathcal{E}$ . Fourth, we re-express these predictions in the basis of the population’s SVD modes. Finally, we detail our methodology to empirically estimate the quantities used in this article, from limited amounts of experimental data. Tables 1–3 summarize the main variables and notations used in the article.

## Statistical notation

In the following, we generally deal with variables  $x$  that assume different values on different trials. An example is the spike count of a single neuron. Trials in turn can be grouped by stimulus  $s$  or choice  $c$ . We can make this explicit by writing  $x^{scq}$  to denote the  $q$ -th trial in which the stimulus was  $s$  and the subject’s choice was  $c$ . Given such a variable, we will write  $E[x]$  for its expectation value, i.e., for the hypothetical value this quantity would take if it could be averaged over infinitely many trials. We will write  $E[x|s]$  for the expectation value conditioned on stimulus  $s$ , i.e., for the expectation value computed over all trials in which the stimulus was  $s$ . A similar notation holds when conditioning on choices  $c$ . We note that for quantities that are already conditional expectations, for instance,  $y(s) = E[x|s]$ , their expectation value  $E[y(s)]$  will average out the stimuli according to their relative probabilities, i.e.,  $E[y(s)] = \sum_s p(s)y(s)$ . Thereby, each stimulus  $s$  contributes to the expectation in proportion to the number of trials associated to it. Then the notations are coherent, since we have  $E[E[x|s]] = E[x]$ . Covariances are generically defined as  $\text{Cov}[x, y] = E[xy] - E[x]E[y]$ , and variances as  $\text{Var}[x] = \text{Cov}[x, x]$ . For vectorial

quantities, we assume  $\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x} \mathbf{y}^\top] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}^\top]$ , and introduce the shorthand  $\text{Cov}[\mathbf{x}] := \text{Cov}[\mathbf{x}, \mathbf{x}]$ .

## Experimental statistics of neural activity and choice

Classic measures in decision-making experiments can be interpreted as estimates of the first- and second-order statistics of choice  $c$  and recorded spike trains  $r_i(t)$ , across all trials with a fixed stimulus value  $s$ :

$$\psi(s) := \mathbb{E}[c|s], \quad (23)$$

$$m_i(t; s) := \mathbb{E}[r_i(t)|s], \quad (24)$$

$$C_{ij}(t, u; s) := \text{Cov}[r_i(t), r_j(u)|s], \quad (25)$$

$$d_i(t; s) := \text{Cov}[r_i(t), c|s]. \quad (26)$$

Here,  $\psi(s)$  is the psychometric curve,  $m_i(t; s)$  is known as the PSTH, and  $C_{ij}(t, u; s)$  as the JPSTH. The choice covariance (CC) curve  $d_i(t; s)$  is our proposal for measuring each neuron's "choice signal". Theoretically, the temporal signals in [eq. 24–26](#) are well-defined quantities in the framework of continuous-time point processes [\[38\]](#). In practice, they are estimated by binning spike trains  $r_i(t)$  with a finite temporal precision, depending on the amount of data available.

From the psychometric curve, we also derive two simpler quantities: the animal's *just-noticeable difference* (JND),  $Z$ , and *decision bias*  $\mu_d$ . We obtain them as the best (MSE) fit to the following formula:

$$\psi(s) = \Phi\left(\frac{s + \mu_d - s_0}{Z}\right), \quad (27)$$

where  $\Phi$  is the standard cumulative normal distribution.  $Z$  measures the inverse slope of the psychometric curve (up to a scaling factor  $\sqrt{2\pi}$ ). The decision bias  $\mu_d$ , when non-zero, represents a bias towards one button when  $s = s_0$ . This formula for the psychometric curve arises naturally when we model the decision task (see below).

## Choice covariance and choice probability

Throughout the article, we consider the special case of a binary choice  $c = \{0, 1\}$ . In this case, the variance of the choice conditioned on  $s$  is given by

$$\sigma_c^2(s) := \text{Var}[c|s] = \psi(s)(1 - \psi(s)), \quad (28)$$

and a straightforward computation shows that

$$d_i(t; s) = \sigma_c^2(s) (\mathbb{E}[r_i(t)|s, c = 1] - \mathbb{E}[r_i(t)|s, c = 0]). \quad (29)$$

(These formulas, and all those below, assume that the choice takes values 0 and 1. Any other binary parametrization should first be reparametrized to  $\{0, 1\}$ .)

The term in brackets is the difference between the two conditional PSTHs, computed only from trials where the animal took one decision vs. the other (stimulus  $s$  keeping a fixed value). This measure is sometimes used as a simpler alternative to choice probabilities [\[3\]](#). In fact, CC curves and CP curves can be analytically related if one assumes Gaussian statistics: see [\[19\]](#) or supporting [S1 Text](#).

## Simplified dependencies on the stimulus

The neural statistics in [eq. 24–26](#) are defined conditionally for each stimulus  $s$  used in the task. To ease the subsequent analysis, we assume that the activity of each neuron is well approximated by a time-varying, linear dependency on the stimulus  $s$ , and that  $C_{ij}(t, u; s)$  is independent of  $s$ . Consequently,

$$\begin{aligned} m_i(t; s) &= m_i^0(t) + b_i(t)s, \\ C_{ij}(t, u; s) &= C_{ij}(t, u). \end{aligned}$$

Since we are modeling a discrimination task, in which stimuli  $s$  display only small variations around the central value  $s_0$ , the linearity assumption seems reasonable. In turn, we can write

$$b_i(t) := \partial_s E[r_i(t)|s]. \quad (30)$$

We will refer to  $b_i(t)$  as the neural tuning. More precisely, it is the slope of the neuron's tuning curve at each time point.

Naturally, actual data (even from a synthetic simulation) always somewhat deviate from this idealized situation. In practice, we obtain the best fits for  $b_i(t)$  and  $C_{ij}(t, u)$  using linear regression, so that

$$b_i(t) = \frac{E[sm_i(t; s)] - E[s]E[m_i(t; s)]}{E[s^2] - E[s]^2}, \quad (31)$$

$$C_{ij}(t, u) = E[C_{ij}(t, u; s)]. \quad (32)$$

Similarly, it is convenient to integrate the various CC curves  $d_i(t; s)$  ([eq. 26](#)) into a single CC curve for each neuron, say  $d_i(t)$ . There is no obvious choice for this simplification, because  $d_i(t; s)$  has to change with  $s$ . For example, the CC signal is non-zero only if stimulus  $s$  and threshold  $s_0$  are close enough for the animal to make occasional mistakes (this is reflected in [eq. 29](#), since  $\sigma_c^2(s)$  tends to zero when the animal makes no mistakes). In the experimental literature, a common choice is to focus only on the CC curve at threshold, that is  $d_i(t) = d_i(t; s_0)$ . In experiments with a limited number of trials, this has the inconvenience of losing the statistical power from nearby stimulus values  $s$  that were also tested. We thus propose an alternative definition:

$$d_i(t) := E[d_i(t; s)], \quad (33)$$

which exploits each stimulus  $s$  in proportion to the number of associated trials. In our model, this averaging also limits the influence of the JND  $Z$  on the magnitude of CC signals: see [eq. 45–46](#).

## Derivation of the linear characteristic equations

The readout defined in [eq. 1–2](#) is linear with respect to the underlying spike trains  $\{r_i(t)\}$ . To clarify the equations, let us introduce the temporal averaging kernel

$$k(t | w, t_R) := \frac{1}{w} h\left(\frac{t_R - t}{w}\right), \quad (34)$$

where parameters  $w$  and  $t_R$  are generally implicit. Then, the integrated spike counts from [eq. 1](#) are simply  $\bar{r}_i = \int_t r_i(t) k(t) dt$ .



Using this notation, [eq. 2](#) becomes  $\hat{s} = a_0 + \sum_i \int_t a_i k(t) r_i(t) dt$ . Thanks to the linear structure, the two first moments of  $\hat{s}$  can easily be developed:

$$E[\hat{s}|s] = a_0 + \sum_i a_i \int_t E[r_i(t)|s] k(t) dt,$$

$$\text{Var}[\hat{s}|s] = \sum_{ij} a_i a_j \int_u \int_t \text{Cov}[r_i(t), r_j(u)|s] k(t) k(u) dt du,$$

$$\text{Cov}[r_i(t), \hat{s}|s] = \sum_j a_j \int_u \text{Cov}[r_i(t), r_j(u)|s] k(u) du.$$

Given our various definitions ([eq. 24–25](#)), and after differentiating the first line with respect to  $s$ , see [eq. 30](#), we obtain:

$$\partial_s E[\hat{s}|s] = \sum_i a_i \int_t b_i(t) k(t) dt = \mathbf{a}^\top \bar{\mathbf{b}}, \quad (35)$$

$$\text{Var}[\hat{s}|s] = \sum_{ij} a_i a_j \int_u \int_t C_{ij}(t, u) k(t) k(u) dt du = \mathbf{a}^\top \bar{\bar{\mathbf{C}}} \mathbf{a}, \quad (36)$$

$$\text{Cov}[r_i(t), \hat{s}|s] = \sum_j a_j \int_u C_{ij}(t, u) k(u) du = [\bar{\mathbf{C}}(t) \mathbf{a}]_i. \quad (37)$$

These are exactly the characteristic equations [4–6](#) from the main text, after introducing the following vectors and matrices:

$$\bar{b}_i := \int_t b_i(t) k(t) dt, \quad (38)$$

$$\bar{C}_{ij}(t) := \int_u C_{ij}(t, u) k(u) du, \quad (39)$$

$$\bar{\bar{C}}_{ij} := \int_t \bar{C}_{ij}(t) k(t) dt, \quad (40)$$

$$\bar{d}_i := \int_t d_i(t) k(t) dt, \quad (41)$$

which simply correspond to the statistics of activity for the integrated spike counts  $\bar{r}_i$  ([eq. 1](#)). Indeed,  $\bar{b}_i = \partial_s E[\bar{r}_i|s]$  (tuning vector),  $\bar{\bar{C}}_{ij} = \text{Cov}[\bar{r}_i, \bar{r}_j|s]$  (noise covariance matrix), and  $\bar{d}_i = E[\text{Cov}[\bar{r}_i, c|s]]$  (choice covariance vector). Given our assumptions above, the resulting quantities are all independent of the stimulus  $s$ . Note though, that all quantities depend on the readout parameters  $w$  and  $t_R$ . Importantly, one can show that the noise covariance matrix  $\bar{\bar{\mathbf{C}}}$  scales as  $w^{-1}$ , under mild assumptions (supporting [S1 Text](#), section 2).

## The decision model of a fine-discrimination task

To produce a binary choice, the (continuous) percept  $\hat{s}$  is fed into the decision model  $c = H(\hat{s} - s_0 + \xi_d)$ , where  $H$  is the Heaviside function,  $s_0$  is the (task-imposed) decision

threshold, and  $\xi_d \sim \mathcal{N}(\mu_d, \sigma_d)$  is a Gaussian variable representing additional noise and biases. The mean  $\mu_d$  implements a possible bias towards one button when  $s = s_0$ . The standard deviation  $\sigma_d$  implements additional sources of noise in the animal's decision process.

Using this decision model, and mild additional assumptions, we can relate the left-hand sides of [eq. 35–37](#) to experimental data. First, we assume that  $E[\hat{s}|s] = s$ , meaning that  $\hat{s}$  follows  $s$  on average. (In statistical terminology,  $\hat{s}$  is an *unbiased estimator* of  $s$ .) Then, the left-hand side of [eq. 35](#) is simply equal to

$$\partial_s E[\hat{s}|s] = 1. \quad (42)$$

Second, we assume that the distribution of  $\mathbf{r}(t)$  (given  $s$ ) is Gaussian. (In theory, this assumption is violated at small time scales due to the binary nature of  $r_i(t)$ . But in practice this is not an issue, as the spike trains always undergo some form of temporal integration afterwards.) Then,  $\hat{s}$  (given  $s$ ) is normally distributed, and [eq. 36](#) ensures that its variance  $\text{Var}[\hat{s}|s]$  is independent of  $s$  (see [Fig. 2B](#)). In these conditions, the predicted formula for the psychometric curve is exactly that of [eq. 27](#), namely,

$$\psi(s) = \Phi\left(\frac{s + \mu_d - s_0}{Z}\right),$$

and the JND,  $Z$ , is given by the following expression:

$$Z^2 = \text{Var}[\hat{s}|s] + \sigma_d^2. \quad (43)$$

Furthermore, under the same assumptions, we can predict the CC curve for each neuron. We use the following general result: for any bivariate normal variables  $(X, Y)$  and threshold  $t$ ,  $\text{Cov}[X, H(Y - t)] = \text{Cov}[X, Y] \mathcal{G}(t; \mu_Y, \sigma_Y)$ , where  $\mathcal{G}(\cdot; \mu, \sigma)$  is the normal density function. Here, we take  $X = r_i(t)$ ,  $Y = \hat{s} + \xi_d$  and  $t = s_0$ , to obtain:

$$d_i(t; s) = \mathcal{G}(s; s_0 - \mu_d, Z) \text{Cov}[r_i(t), \hat{s}|s]. \quad (44)$$

With  $d_i(t)$  defined as an average CC curve over tested stimuli ([eq. 33](#)), we finally obtain

$$d_i(t) = \kappa(Z) \text{Cov}[r_i(t), \hat{s}|s], \quad (45)$$

$$\text{with } \kappa(Z) = E[\mathcal{G}(s; s_0 - \mu_d, Z)]. \quad (46)$$

The final equation for CC signals ([eq. 9](#)) is obtained by combining [eq. 37](#) and [45](#).

In many experimental setups, the averaging over stimuli  $s$  will ensure that  $\kappa(Z)$  has only a mild dependency on its argument  $Z$ . Indeed, note the rough approximation  $\kappa(Z) \propto \int ds \mathcal{G}(s; s_0 - \mu_d, Z) = 1$ , valid whenever the tested stimuli  $s$  are uniformly distributed over a range of values comparable to  $Z$ . This is another practical argument for considering the stimulus-averaged CC signal  $d_i(t)$ , from [eq. 33](#).

## Signal, noise, and sensitivity

The just-noticeable difference (JND) and the sensitivity can be related to the variances of signal and noise in the population. Here, we briefly review these relations. The variance of any scalar variable  $x$  that changes from trial to trial can be decomposed in a signal term  $\sigma_x^2 := \text{Var}[E[x|s]]$  and a noise term  $Z_x^2 := E[\text{Var}[x|s]]$ . Then, note that  $\text{Var}[x] = \sigma_x^2 + Z_x^2$ .

The noise term  $Z_x$  defines the minimal level past which fluctuations in  $x$  can be attributed to  $s$  rather than intrinsic noise—hence the term JND. When a decision is taken on the basis of variable  $x$ , the JND governs the inverse slope of the corresponding psychometric curve (see [eq.](#)

27). We also define the *sensitivity* of variable  $x$  as

$$Y_x := \frac{\sigma_x^2}{\sigma_x^2 + Z_x^2}, \quad (47)$$

which is simply the ratio of the signal to the total variance. The sensitivity  $Y_x$  takes values between 0 and 1. It thus avoids singularities which may occur when  $Z_x$  tends to 0 or  $+\infty$ .

We can also distinguish between signal-related and noise-related variance for the (time-averaged) neural activities  $\bar{\mathbf{r}}$ . The signal covariance matrix,  $\Sigma$ , noise covariance matrix,  $\bar{\bar{\mathbf{C}}}$ , and total covariance matrix,  $\mathbf{A}$ , are given by the following relations:

$$\Sigma := \text{Cov}[\mathbf{E}[\bar{\mathbf{r}}|s]] = \text{Cov}[\bar{\mathbf{m}}(s)] = \sigma_s^2 \bar{\mathbf{b}} \bar{\mathbf{b}}^\top \quad (48)$$

$$\bar{\bar{\mathbf{C}}} := \mathbf{E}[\text{Cov}[\bar{\mathbf{r}}|s]] \quad (49)$$

$$\mathbf{A} := \text{Cov}[\bar{\mathbf{r}}] = \bar{\bar{\mathbf{C}}} + \sigma_s^2 \bar{\mathbf{b}} \bar{\mathbf{b}}^\top. \quad (50)$$

The last equality is the classic decomposition of total covariance into noise and signal terms. Note that  $\Sigma$  is a rank-1 matrix, owing to the system's assumed linearity wrt. stimulus  $s$ .

In turn, these matrices allow to compute the signal- and noise- variances for any weighted sum of the neural activities. For our linear readout (with added decision noise  $\xi_d$ ), we have  $x = \mathbf{a}^\top \bar{\mathbf{r}} + \xi_d$ , and thus:

$$\sigma_x^2 = \mathbf{a}^\top \Sigma \mathbf{a} = \sigma_s^2 (\mathbf{a}^\top \bar{\mathbf{b}})^2, \quad (51)$$

$$Z_x^2 = \mathbf{a}^\top \bar{\bar{\mathbf{C}}} \mathbf{a} + \sigma_d^2, \quad (52)$$

$$\sigma_x^2 + Z_x^2 = \mathbf{a}^\top \mathbf{A} \mathbf{a} + \sigma_d^2. \quad (53)$$

## Optimal readout from a neural ensemble $\mathcal{E}$

We now assume that the readout vector  $\mathbf{a}$  has support only on some neural ensemble  $\mathcal{E}$ . Formally, we introduce the  $K \times N_{\text{tot}}$  projection matrix  $\mathbf{H}(\mathcal{E})$ , such that for  $i \in \mathcal{E}$  and every neuron  $j$ ,  $H_{ij}(\mathcal{E}) = \delta_{ij}$ . Then, the restrictions of vectors and matrices in neuron space, such as  $\bar{\mathbf{b}}$  and  $\bar{\bar{\mathbf{C}}}$ , to ensemble  $\mathcal{E}$  will be denoted by a subscript  $r$  (for restriction), so that

$$\bar{\mathbf{b}}_r := \mathbf{H} \bar{\mathbf{b}}, \quad (54)$$

$$\bar{\bar{\mathbf{C}}}_r := \mathbf{H} \bar{\bar{\mathbf{C}}} \mathbf{H}^\top. \quad (55)$$

Our principle of (restricted) optimality selects the readout vector  $\mathbf{a}$  which maximizes the signal-to-noise ratio of the resulting percept  $\hat{s}$ . Since  $\mathbf{a}^\top \bar{\mathbf{b}} = 1$  (unbiased percept, eq. 35 and 42), the signal variance is imposed to be  $\sigma_x^2 = \sigma_s^2$  (eq. 51). Under this constraint, optimality is achieved by minimizing the noise variance  $\mathbf{a}^\top \bar{\bar{\mathbf{C}}} \mathbf{a}$  (eq. 52)—or equivalently, the total variance  $\mathbf{a}^\top \mathbf{A} \mathbf{a}$  (eq. 53). The solution, known as Fisher's Linear Discriminant, is easily found with Lagrange multipliers (either based on  $\bar{\bar{\mathbf{C}}}$  or  $\mathbf{A}$ ):

$$\mathbf{a}_r = \frac{(\bar{\bar{\mathbf{C}}}_r)^{-1} \bar{\mathbf{b}}_r}{\bar{\mathbf{b}}_r^\top (\bar{\bar{\mathbf{C}}}_r)^{-1} \bar{\mathbf{b}}_r} = \frac{(\mathbf{A}_r)^{-1} \bar{\mathbf{b}}_r}{\bar{\mathbf{b}}_r^\top (\mathbf{A}_r)^{-1} \bar{\mathbf{b}}_r}. \quad (56)$$

The second formulation of  $\mathbf{a}_r$ , based on the total covariance matrix  $\mathbf{A}_r$ , will prove more useful when we turn to the SVD analysis. It also has the advantage of avoiding the singularity which may occur when vector  $\bar{\mathbf{b}}_r$  lies outside the span of matrix  $\bar{\bar{\mathbf{C}}}_r$ . In that case one simply replaces  $(\mathbf{A}_r)^{-1}$  by the (Moore-Penrose) pseudoinverse  $(\mathbf{A}_r)^+$ .

When combining the optimal readout in [eq. 56](#) with the equation for the JND ([eq. 52](#)), we obtain the JND predicted by the model:

$$Z^2 = \left( \bar{\mathbf{b}}_r^\top (\bar{\bar{\mathbf{C}}}_r)^{-1} \bar{\mathbf{b}}_r \right)^{-1} + \sigma_d^2. \quad (57)$$

Equivalently, using the formulations based on total variance ([eq. 47](#), [53](#), [56](#)) we obtain the model's prediction for sensitivity:

$$Y = \frac{\sigma_s^2}{\left( \bar{\mathbf{b}}_r^\top (\mathbf{A}_r)^{-1} \bar{\mathbf{b}}_r \right)^{-1} + \sigma_d^2}. \quad (58)$$

## CC signals for the optimal readout

When combining the optimal readout in [eq. 56](#) with the characteristic equation for the CC curves ([eq. 9](#)), we obtain the CC curves predicted by the model,

$$d_i(t) = \kappa(Z) (Z^2 - \sigma_d^2) \bar{\mathbf{C}}_{ir}(t) (\bar{\bar{\mathbf{C}}}_r)^{-1} \bar{\mathbf{b}}_r. \quad (59)$$

Here,  $d_i(t)$  is the resulting, predicted CC curve for every neuron  $i$  in the population (not only in ensemble  $\mathcal{E}$ ). Note that  $\bar{\mathbf{C}}_{ir}(t)$  is the restriction of vector  $\bar{\mathbf{C}}_i(t)$  ([eq. 39](#)) to neurons  $j \in \mathcal{E}$ , but that  $i = 1 \dots N_{\text{tot}}$  still runs over all neurons. [Equation 59](#) can also be expressed in its temporally-integrated form, using the definition  $\int_t \bar{\mathbf{C}}(t) k(t) dt = \bar{\bar{\mathbf{C}}}$ :

$$\bar{d}_i = \kappa(Z) (Z^2 - \sigma_d^2) \bar{\bar{\mathbf{C}}}_{ir} (\bar{\bar{\mathbf{C}}}_r)^{-1} \bar{\mathbf{b}}_r. \quad (60)$$

If neuron  $i$  belongs to the readout ensemble  $\mathcal{E}$ , matrix  $\bar{\bar{\mathbf{C}}}_r$  simplifies away from [eq. 60](#), yielding:

$$\bar{d}_i^{(\mathcal{E})} = \kappa(Z) (Z^2 - \sigma_d^2) \bar{b}_i^{(\mathcal{E})}. \quad (61)$$

This equation, first shown in [\[19\]](#), means that choice signals within the readout ensemble are simply proportional to tuning. This is not true, however, for neurons outside the readout ensemble.

This has two important implications. First, it proves that choice signals are markedly different for neurons inside or outside the readout ensemble (an observation made empirically by [\[12\]](#)). Second, as we consider readout ensembles  $\mathcal{E}$  larger and larger, [eq. 61](#) will become true for more and more neurons. As a result the statistical indicator  $V$  ([eq. 15](#)), which measures the population-wide deviation from linearity between  $\bar{d}_i$  and  $\bar{b}_i$ , is expected to decrease with the readout ensemble's size  $K$ .

Finally, under the assumption of (restricted) optimality, the time-averaged statistical indicator  $\bar{q}$  is always positive. Indeed, averaging over all neurons  $i$  in the population is akin to a scalar product:  $\bar{q} = \langle \bar{b}_i \bar{d}_i \rangle_i = N_{\text{tot}}^{-1} \bar{\mathbf{b}}^\top \bar{\mathbf{d}}$ . Using this relation and [eq. 60](#), we get

$$\bar{q} = N_{\text{tot}}^{-1} \kappa(Z) (Z^2 - \sigma_d^2) \bar{\mathbf{b}}^\top \left( \bar{\bar{\mathbf{C}}} \mathbf{H}^\top (\bar{\bar{\mathbf{C}}}_r)^{-1} \mathbf{H} \right) \bar{\mathbf{b}}, \quad (62)$$

which is always positive because both matrices  $\bar{\bar{\mathbf{C}}}$  and  $\mathbf{H}^\top (\bar{\bar{\mathbf{C}}}_r)^{-1} \mathbf{H}$  are symmetric semi-definite positive.

## Singular value decomposition

We denote the time-averaged activities of neuron  $i$  in the  $q$ -th presentation of stimulus  $s$  as  $\bar{r}_i^{sq}$ . We interpret these activities as a very large  $N_{\text{tot}} \times \Omega$  matrix, where  $N_{\text{tot}}$  refers to the number of neurons and  $\Omega$  to an idealized, and essentially infinitely large number of trials.

Next, we consider the singular value decomposition (SVD) of the neural activities. The (compact) SVD is a standard decomposition which can be applied to any rectangular matrix  $\mathbf{R}$ . It is given by  $\mathbf{R} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top$ , where  $\mathbf{\Lambda}$  is an  $M \times M$  diagonal matrix with strictly positive entries  $\lambda_m$  (the singular values),  $\mathbf{U}$  is an  $N_{\text{tot}} \times M$  matrix of orthogonal columns (meaning  $\mathbf{U}^\top \mathbf{U} = \mathbf{Id}_M$ ), and  $\mathbf{V}$  is an  $\Omega \times M$  matrix of orthogonal columns (meaning  $\mathbf{V}^\top \mathbf{V} = \mathbf{Id}_M$ ).

Using the indices defined above, the SVD decomposition for the neural activities becomes

$$\bar{r}_i^{sq} = \bar{r}_i^0 + \sum_{m=1}^M \lambda_m u_i^m v_m^{sq}, \quad (63)$$

where  $\bar{r}_i^0$  is the average activity of each cell over all trials and stimuli. The orthogonality of  $\mathbf{U}$  implies that for all indices  $m$  and  $n$ , we have  $\sum_i u_i^m u_i^n = \delta^{mn}$ , while the orthogonality of  $\mathbf{V}$  similarly implies  $\sum_{sq} (v_m^{sq} v_n^{sq}) = \delta_{mn}$ .

## Statistics of activity, in the space of modes

The SVD decomposition (eq. 63) is best interpreted as a change of variables re-expressing neural activities  $\{\bar{r}_i^{sq}\}_{i=1 \dots N_{\text{tot}}}$  in terms of mode appearance variables  $\{v_m^{sq}\}_{m=1 \dots M}$ . As a result, we can define the respective equivalents of all statistical quantities in the space of activity modes. Specifically, we can reinterpret sums over trials in the SVD as expectations, thus emphasizing the statistical interpretation of the SVD. First we note that  $\bar{r}_i^0 = E[\bar{r}_i^{sq}]$  for all neurons  $i$ , so that the data for the actual SVD has been “centered”. This centering implies for all modes  $m$  that

$$E[v_m^{sq}] = 0, \quad (64)$$

$$E[v_m^{sq}|s] = \eta_m(s - E[s]), \quad (65)$$

where  $\eta_m$  is the tuning parameter of the  $m$ -th mode, just as  $\bar{b}_i$  was the tuning parameter for the  $i$ -th neuron. Grouping all mode appearance variables in a vector  $\mathbf{v}$ , we obtain the signal covariance and total covariance matrices in mode space as

$$\Sigma_v := \text{Cov}[E[\mathbf{v}|s]] = \text{Cov}[\eta s] = \sigma_s^2 \eta \eta^\top, \quad (66)$$

$$\mathbf{A}_v := \text{Cov}[\mathbf{v}] = E[\mathbf{v} \mathbf{v}^\top] = \mathbf{Id}_M. \quad (67)$$

where the last relation follows from the orthogonality of  $\mathbf{V}$  explained in the previous section. The singular values  $\lambda_m$  and distribution vectors  $\mathbf{u}^m$  then allow us to relate the statistics at the levels of neurons and modes. Using the SVD formula (eq. 63) yields (in matrix form):

$$\bar{\mathbf{b}} = \mathbf{U} \mathbf{\Lambda} \eta, \quad (68)$$

$$\mathbf{A} = \mathbf{U} \mathbf{\Lambda}^2 \mathbf{U}^\top. \quad (69)$$

## Sensitivity of sub-ensembles, in the space of modes

We now wish to understand which factors govern the sensitivity embedded in a neural sub-ensemble  $\mathcal{E}$  of cardinality  $K$ . For simplicity, we will consider the case for which the decision noise

is negligible, i.e.,  $\sigma_d \rightarrow 0$ . Then, from [eq. 58](#), we have

$$Y = \sigma_s^2 \bar{\mathbf{b}}_r^\top (\mathbf{A}_r)^+ \bar{\mathbf{b}}_r. \quad (70)$$

Here we use explicitly the most general formula, based on the pseudo-inverse of matrix  $\mathbf{A}_r$ . To re-express this sensitivity of finite sub-ensembles  $\mathcal{E}$  into mode space, we need to find the equivalent, restricted expressions of [eq. 68–69](#). For that purpose, we introduce the *design matrix* associated to ensemble  $\mathcal{E}$  in mode space:

$$\mathbf{X} := \Lambda \mathbf{U}^\top \mathbf{H}^\top, \quad (71)$$

where  $\mathbf{H}$  is the restriction operator from [eq. 54](#).  $\mathbf{X}$  is an  $M \times K$  matrix with elements  $x_i^m := \lambda_m u_i^m$ . Using this matrix, we obtain from [eq. 68–69](#) that  $\bar{\mathbf{b}}_r = \mathbf{X}^\top \mathbf{Z}$  and  $\mathbf{A}_r = \mathbf{X}^\top \mathbf{X}$ , so that [eq. 70](#) becomes

$$\begin{aligned} Y &= \sigma_s^2 \eta^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \eta \\ &= \sigma_s^2 \eta^\top \mathbf{P} \eta, \end{aligned} \quad (72)$$

where we have defined the  $M \times M$  matrix

$$\mathbf{P} := \mathbf{X} (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top. \quad (73)$$

Note that  $\mathbf{P}$  is simply the orthogonal projector on  $\text{Im}(\mathbf{X})$ , since  $\mathbf{P} = \mathbf{P}^2 = \mathbf{P}^\top$ , and  $\text{Im}(\mathbf{P}) = \text{Im}(\mathbf{X})$ .

The projector  $\mathbf{P} = \mathbf{P}(\mathcal{E})$  spans more and more space as the size  $K$  of ensemble  $\mathcal{E}$  increases. In the limiting case, when  $K$  is larger than the number of modes  $M$ , then necessarily  $\mathbf{P} = \mathbf{Id}_M$ , and we obtain

$$Y_{\text{tot}} = \sigma_s^2 \eta^\top \eta = \sum_{m=1}^M \sigma_s^2 \eta_m^2. \quad (74)$$

In other words, all modes are available experimentally, and sensitivity estimates saturate to their maximum value, independently of ensemble  $\mathcal{E}$ . We can explicitly denote the sensitivity of each mode's activation variable  $v_m$  by defining

$$y_m := \sigma_s^2 \eta_m^2. \quad (75)$$

By solving [eq. 68](#) for  $\eta$ , we obtain  $\eta_m = \sum_i \lambda_m^{-1} u_i^m \bar{b}_i$ , which in turn yields [eq. 20](#) from the main text.

## CC signals, in the space of modes

Similarly, we can express CC signals in mode space. First, we re-express the CC equation ([eq. 10](#)) as a function of the total covariance  $\mathbf{A}$  ([eq. 50](#)) to obtain

$$\begin{aligned} \bar{\mathbf{d}} &= \kappa(Z) \bar{\mathbf{C}} \mathbf{a} \\ &= \kappa(Z) (\mathbf{A} - \sigma_s^2 \bar{\mathbf{b}} \bar{\mathbf{b}}^\top) \mathbf{a}. \end{aligned}$$

We further recall that  $\mathbf{a}^\top \bar{\mathbf{b}} = 1$  (unbiased percept, see [eq. 35](#) and [42](#)). Hence, up to a scaling and shift, the CC vector  $\bar{\mathbf{d}}$  can be replaced by the *total percept covariance* vector

$$\mathbf{e} := \mathbf{A} \mathbf{a} = \kappa(Z)^{-1} \bar{\mathbf{d}} + \sigma_s^2 \bar{\mathbf{b}}. \quad (76)$$

In the case of an optimal readout, vector  $\mathbf{a}$  is given by [eq. 56](#), so that we obtain

$$\mathbf{e} = \frac{\mathbf{A} \mathbf{H}^\top \mathbf{A}_r^+ \bar{\mathbf{b}}_r}{\bar{\mathbf{b}}_r^\top (\mathbf{A}_r)^+ \bar{\mathbf{b}}_r}. \quad (77)$$



Second, using the corresponding sensitivity  $Y$  (eq. 70), and the SVD expressions for  $\mathbf{A}$  and  $\bar{\mathbf{b}}$  (eq. 68–69), and for  $\mathbf{A}_r$  and  $\bar{\mathbf{b}}_r$  as a function of matrix  $\mathbf{X}$  (eq. 71), we write:

$$\begin{aligned}\mathbf{e} &= \sigma_s^2 Y^{-1} \mathbf{A} \mathbf{H}^\top \mathbf{A}_r^\top \bar{\mathbf{b}}_r \\ &= \sigma_s^2 Y^{-1} \mathbf{U} \mathbf{\Lambda} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \boldsymbol{\eta} \\ &= \sigma_s^2 Y^{-1} \mathbf{U} \mathbf{\Lambda} \mathbf{P} \boldsymbol{\eta}.\end{aligned}\quad (78)$$

Here also, the final result can be expressed as a function of  $\mathbf{P}$ , the projection matrix associated to ensemble  $\mathcal{E}$  in the space of modes (eq. 73). Note again that  $\mathbf{e}$  provides the CC signal for every neuron  $i$  in the population (not only in ensemble  $\mathcal{E}$ ). As  $\mathcal{E}$  tends to the full population,  $\mathbf{P} = \mathbf{P}(\mathcal{E})$  tends to  $\mathbf{I}_{\mathbf{d}_M}$  and we recover  $\mathbf{e}(\infty) = \sigma_s^2 Y_{\text{tot}}^{-1} \bar{\mathbf{b}}$ , the prediction for choice signals in the case of a (globally) optimal readout [19].

Using eq. 78, we can finally compute the analytical predictions for the two CC statistical indicators,  $\bar{q}$  and  $V$ . Precisely, we compute the following population-wide regression coefficient between  $\mathbf{e}$  and  $\bar{\mathbf{b}}$ :

$$\begin{aligned}Q &:= \langle e_i \bar{b}_i \rangle_i \\ &= N_{\text{tot}}^{-1} \bar{\mathbf{b}}^\top \mathbf{e} \\ &= \sigma_s^2 N_{\text{tot}}^{-1} Y^{-1} \boldsymbol{\eta}^\top \mathbf{\Lambda} \mathbf{U}^\top \mathbf{U} \mathbf{\Lambda} \mathbf{P} \boldsymbol{\eta} \\ &= \sigma_s^2 N_{\text{tot}}^{-1} Y^{-1} \boldsymbol{\eta}^\top \mathbf{\Lambda}^2 \mathbf{P} \boldsymbol{\eta}.\end{aligned}\quad (79)$$

Again, we made use of the SVD expressions for  $\bar{\mathbf{b}}$  (eq. 68) and  $\mathbf{e}$  (eq. 78). Note that, since  $\mathbf{e}$  is a linear rescaling of  $\bar{\mathbf{d}}$ ,  $Q$  is a similar rescaling of indicator  $\bar{q}$ , as pointed in the main text (eq. 18). Finally, a very similar computation leads to the expression of indicator  $V$  (eq. 15) in the space of modes:

$$V = \kappa(Z)^2 N_{\text{tot}}^{-2} \sigma_s^4 Y^{-2} \boldsymbol{\eta}^\top \mathbf{\Lambda}^2 (\boldsymbol{\eta} \boldsymbol{\eta}^\top \mathbf{P} - \mathbf{P} \boldsymbol{\eta} \boldsymbol{\eta}^\top) \mathbf{\Lambda}^2 \mathbf{P} \boldsymbol{\eta}.\quad (80)$$

## Sensitivity and CC signals as a function of $K$

We are now better armed to understand how sensitivity and CC indicators vary as a function of the readout ensemble  $\mathcal{E}$ . We are mostly interested in averages of these quantities over very large numbers of randomly chosen ensembles  $\mathcal{E}$  of size  $K$ ; we thus use the generic notation  $E[x|K] := E[x(\mathcal{E}) | \text{Card}(\mathcal{E}) = K]$  to denote the expected value of a variable  $x$  when averaging over ensembles of size  $K$ . Note that this notation is equivalent to the more explicit notation used in the main text, so that  $E[x|K] = \langle x \rangle_{\mathcal{E}(K)}$ . From eq. 72 we find:  $E[Y|K] = \sigma_s^2 \boldsymbol{\eta}^\top E[\mathbf{P}|K] \boldsymbol{\eta}$ .

To understand the properties of the  $(M \times M)$  matrix  $E[\mathbf{P}|K]$ , we view the  $(M \times K)$  design matrix  $\mathbf{X}(\mathcal{E})$  (eq. 71) as a collection of  $K$  random vectors  $\mathbf{x}_i$  in mode space, viewing neuron identities  $i$  as the random variable. Thus,  $\mathbf{P}(\mathcal{E})$  is the orthogonal projector on the linear span of the  $K$  sample vectors  $\{\mathbf{x}_i\}_{i \in \mathcal{E}}$ . As a projector, its trace is equal to its rank, so we have  $\text{Tr}(E[\mathbf{P}|K]) = K$ . Furthermore, since  $K+1$  samples span on average more space than  $K$  samples, we are ensured that  $E[\mathbf{P} | K+1] \succeq E[\mathbf{P}|K]$ , in the sense of positive semidefinite matrices.

Finally, intuition and numerical simulations suggest that  $E[\mathbf{P}|K]$  is almost diagonal. Indeed, as the various modes are linearly independent, there is no linear interplay between the different dimensions of  $\mathbf{x}_i$  across samples  $i$ . More precisely, the expectation value over neurons is

$$\langle x_i^m x_i^n \rangle_i = N_{\text{tot}}^{-1} \lambda_m^2 \delta^{mn}.$$
 This leads to the matrix expression:

$$E[\mathbf{X} \mathbf{X}^\top | K] = K N_{\text{tot}}^{-1} \mathbf{\Lambda}^2.$$

Let us consider the (compact) SVD decomposition  $\mathbf{X} \mathbf{X}^\top := \mathbf{W} \mathbf{D} \mathbf{W}^\top$ , with  $\mathbf{W}^\top \mathbf{W} = \mathbf{Id}$ , and  $\mathbf{D}$  an invertible diagonal matrix. Then, the projection matrix  $\mathbf{P}$  is simply equal to  $\mathbf{W} \mathbf{W}^\top$ . As for the previous equation, it rewrites

$$\mathbb{E}[\mathbf{W} \mathbf{D} \mathbf{W}^\top | K] = K N_{\text{tot}}^{-1} \mathbf{\Lambda}^2.$$

Here, both matrices  $\mathbf{D}$  and  $\mathbf{\Lambda}$  are diagonal. So, if we assume a form of independence between  $\mathbf{W}$  and  $\mathbf{D}$ , it is reasonable to suppose that  $\mathbb{E}[\mathbf{W} \mathbf{W}^\top | K] = \mathbb{E}[\mathbf{P} | K]$  is close to diagonal as well. (Actually, we postulate that  $\mathbb{E}[\mathbf{P} | K]$  is exactly diagonal when the random vectors  $\mathbf{x}_i$  follow a normal distribution. In the general case, small or moderate deviations from diagonality can be observed.) We denote these diagonal terms as

$$\epsilon(K) := \text{diag}(\mathbb{E}[\mathbf{P} | K]). \quad (81)$$

The properties of  $\mathbb{E}[\mathbf{P} | K]$  stated above imply that  $\sum_m \epsilon_m(K) = K$  (trace property), and  $\epsilon_m(K+1) \geq \epsilon_m(K)$  (growth property). Finally, we can consider the resulting approximations of sensitivity (eq. 72) and CC indicator (eq. 79):

$$\mathbb{E}[Y | K] \simeq \sigma_s^2 \sum_{m=1}^M \epsilon_m(K) \eta_m^2, \quad (82)$$

$$\mathbb{E}[YQ | K] \simeq N_{\text{tot}}^{-1} \sigma_s^2 \sum_{m=1}^M \epsilon_m(K) \lambda_m^2 \eta_m^2. \quad (83)$$

In this expression, we recognize the individual mode sensitivities  $y_m = \sigma_s^2 \eta_m^2$ . For CC signals, we also make the approximation  $\mathbb{E}[YQ | K] \simeq \mathbb{E}[Y | K] \mathbb{E}[Q | K]$ , and recover eq. 21–22 from the main text. Unfortunately, there is no such simple approximation for indicator  $V$ , that would lead from eq. 80 to  $\mathbb{E}[V | K]$ .

## Validation on a simulated neural network

In this final part of the Methods, we provide additional information for applying our inference method (Case 2) to experimental data. The neural network used to test our methods is described in detail in supporting S1 Text (section 3). Briefly, on each trial, 2000 input Poisson neurons fire with rate  $s$ , taking one of three possible values 25, 30 and 35 Hz (so in our simulation, stimulus units are Hz). The encoding population *per se* consists of 5000 leaky integrate-and-fire (LIF) neurons. 1000 of these neurons receive sparse excitatory projections from the input Poisson neurons, which naturally endows them with a positive tuning to stimulus  $s$ . Another 1000 neurons receive sparse inhibitory projections from the Poisson neurons, which naturally endows them with negative tuning. The remaining 3000 neurons receive no direct projections from the input. Instead, all neurons in the encoding population are coupled through a sparse connectivity with random delays up to 5 msec. Synaptic weights are random and balanced, leading to a mean firing rate of 21.8 Hz in the population. We implemented and simulated the network using Brian, a spiking neural network simulator in Python [39].

The “true” perceptual readout from this network was built from a fixed random set of  $K^* = 80$  neurons, with temporal parameters  $w^* = 50$  msec and  $t_R^* = 100$  msec, and decision noise  $\sigma_d^* = 1$  stimulus units (Hz). The readout vector  $\mathbf{a}^*$  was built optimally given these constraints (eq. 12). The trials used to learn  $\mathbf{a}^*$  were not used in the subsequent analysis. The resulting JND for the “animal” was  $Z^* \approx 3$  stimulus units (Hz).

Then, “experimentally”, neural activity was observed through 15 pools of 170 simultaneously recorded neurons, each pool being recorded on  $3 \times 180$  trials. For the statistical inference

method, we assumed a square integration kernel  $h$ . We tested all combinations of the following readout parameters (in matrix notation):  $K = 10:10:150$  neurons,  $w = 10:10:100$  msec,  $t_R = 10:10:200$  msec,  $\sigma_d = 0:0.25:3$  stimulus units (Hz). For each tested size  $K$ , we picked 2000 random candidate ensembles  $\mathcal{E}$  (always within one of the 15 simultaneous pools) to build the predictions. For each ensemble  $\mathcal{E}$ , another ensemble  $\mathcal{I}$  of 20 neurons, segregated from  $\mathcal{E}$ , were used to predict CC signals outside the readout ensemble (this was always possible since recording pools had size 170, and  $K \leq 150$ ). The details of these predictions are explained in the following paragraph. Finally, the three terms in the “statistical” loss function (eq. 16) were weighted according to the power of the respective, true measures. That is:

$$\lambda = \frac{(Z^*)^4}{\iint dt du (q^*(u, t))^2} \quad \text{and} \quad \mu = \frac{(Z^*)^4}{(V^*)^2}.$$

## Experimental predictions for CC indicators

Here, we detail how to compute the CC indicators  $q(u, t)$  and  $V$  (eq. 14–15) from actual data. For the *measured* versions  $q^*(u, t)$  and  $V^*(w, t_R)$ , this is straightforward. One considers the true, *measured* CC signals  $d_i^*(t)$ , and computes the population averages in eq. 14–15 over as many neurons  $i$  as were recorded. Note however that the final indicators can be corrupted by noise, whenever each measure  $d_i^*(t)$  comes from too few recording trials (this problem is addressed in the next section). Also note that, since the definition of  $V$  requires a temporal integration, we actually have to produce a different “true”  $V^*$  for each tested set of temporal parameters  $w$  and  $t_R$ .

Conversely, special care must be taken when it comes to *predicted* CC indicators. Whenever a candidate ensemble  $\mathcal{E}$  is proposed as the source of the readout, eq. 59 predicts the resulting CC signal  $d_i(t|\mathcal{E})$  for every neuron  $i$  in the population. However, in practice, the noise covariance term  $\tilde{C}_{ir}(t)$  is required in the computation, so neuron  $i$  and ensemble  $\mathcal{E}$  must have been recorded simultaneously during the same run. This limits the number of neurons  $i$  which can participate in the population averages.

Furthermore, choice covariances will generally differ between neurons that are part of the readout ensemble and neurons that are not (see eq. 61 and the associated discussion). As a result, the two following averages must be predicted separately:

$$q_{\mathcal{E}}(u, t | \mathcal{E}) := \langle b_i(u) d_i(t | \mathcal{E}) \rangle_{i \in \mathcal{E}}, \quad (84)$$

$$q_{\text{out}}(u, t | \mathcal{E}) := \langle b_i(u) d_i(t | \mathcal{E}) \rangle_{i \notin \mathcal{E}}, \quad (85)$$

before one can recombine them in the correct proportions:

$$p(\mathcal{E}) := \frac{K}{N_{\text{tot}}}, \quad (86)$$

$$q(u, t | \mathcal{E}) = p(\mathcal{E}) q_{\mathcal{E}}(u, t | \mathcal{E}) + (1 - p(\mathcal{E})) q_{\text{out}}(u, t | \mathcal{E}), \quad (87)$$

and similarly for  $V(\mathcal{E})$ . To compute  $q_{\text{out}}$  experimentally, each tested candidate ensemble  $\mathcal{E}$  (of size  $K$ ) is associated to a complementary set of neurons  $\mathcal{I}$  (of size  $I$ ), which we use to approximate the average in eq. 85:

$$q_{\mathcal{I}}(u, t | \mathcal{E}) := \langle b_i(u) d_i(t | \mathcal{E}) \rangle_{i \in \mathcal{I}}. \quad (88)$$

All neurons in ensembles  $\mathcal{E}$  and  $\mathcal{I}$  must have been recorded during the same run, which imposes that  $I+K \leq N$ . Hence in our simulations, we chose a size  $I = 170 - 150 = 20$  neurons.

Clearly, 20 neurons is not sufficient for  $q_{\mathcal{I}}$  to be a reliable population average. So in practice, we cannot estimate reliably each prediction  $q(u, t | \mathcal{E})$  from [eq. 87](#). Luckily, we are not interested in their value for each individual readout ensemble  $\mathcal{E}$ . We simply need to estimate their means across all tested ensembles  $\mathcal{E}$  of similar size:

$$\langle q(u, t) \rangle_{\mathcal{E}} := \langle q(u, t | \mathcal{E}) \rangle_{\mathcal{E} \text{ with Card}(\mathcal{E})=K} \quad (89)$$

$$\langle V \rangle_{\mathcal{E}} := \langle V(\mathcal{E}) \rangle_{\mathcal{E} \text{ with Card}(\mathcal{E})=K} \quad (90)$$

which will be reliable as soon as we test a sufficient amount of candidate ensembles  $\mathcal{E}$ .

Note that in the final inference ([eq. 16](#)), a match is sought between the true indicators  $q^*$  and  $V^*$ —which arise from a single readout ensemble  $\mathcal{E}^*$ , and the predictions  $\langle q \rangle_{\mathcal{E}}$  and  $\langle V \rangle_{\mathcal{E}}$ —which are average values across all readout ensembles  $\mathcal{E}$  of size  $K$ . Thus, a prediction error can occur whenever the true readout ensemble  $\mathcal{E}^*$  is not a “typical” representative of its size  $K^*$ . To quantify these potential errors, one should also estimate the indicators’ *variance* across ensembles  $\mathcal{E}$  of same size.

## Correcting for the finite amounts of data

The computations of  $Z$ ,  $q$  and  $V$ , as described above, can produce imprecise results when the data are overly limited. Generically, for any quantity  $X$  estimated from the data, we can write

$$X_{\text{noisy}} = X_{\text{ideal}} + \xi,$$

where  $\xi$  represents the measurement error on  $X$  due to the finite amounts of data. If we could recompute  $X$  from a different set of neurons and/or a different set of trials, variable  $\xi$  would take a different value—meaning that  $\text{Var}(\xi) > 0$ . This is an inescapable phenomenon for experimental measures.

More problematically, variable  $\xi$  can display a systematic *bias*, meaning that  $E(\xi) \neq 0$ . Since the bias is generally different for the ‘true’ and ‘predicted’ versions, the comparison between the two ([eq. 16](#)) will be systematically flawed. To counteract this effect, we applied a number of correction procedures when computing indicators  $Z$ ,  $q$  and  $V$ , to ensure that they are globally unbiased. We only provide an overview here, and refer to supporting [S1 Text](#) for a detailed description.

First, when the optimal vector  $\mathbf{a}$  is computed with Fisher’s linear discriminant, it systematically underestimates the JND  $Z$  (overestimates the sensitivity  $Y$ ). Essentially, vector  $\mathbf{a}_r$  computed through [eq. 12](#) finds artificial “holes” in matrix  $\bar{\bar{\mathbf{C}}}_r$  which are only due to its imprecise measurement—a phenomenon known as statistical *overfitting*. The less recording trials, the more overfitting there will be [[40](#), [41](#)]. We addressed this problem with a regularization technique, inspired by Bayesian linear regression [[42](#)]. We replaced [eq. 12](#) by the following:

$$\mathbf{a}_r = \frac{(\bar{\bar{\mathbf{C}}}_r + \lambda \text{Id})^{-1} \bar{\mathbf{b}}_r}{\bar{\mathbf{b}}_r^T (\bar{\bar{\mathbf{C}}}_r + \lambda \text{Id})^{-1} \bar{\mathbf{b}}_r},$$

where the strength of parameter  $\lambda$  imposes the degree of regularization. We chose  $\lambda$  according to an ‘empirical Bayes’ principle, to maximize the likelihood of the data under a given statistical model (supporting [S1 Text](#), section 4). It largely mitigated the effects of overfitting, without totally suppressing them—as can be seen in [Fig. 5D-E](#).

Second, indicator  $V$  (eq. 15) can also display substantial biases ( $E(\xi) \neq 0$  in the above discussion). Indeed, its computation relies on squared quantities—such as  $\bar{d}_i^2$  or  $\bar{q}^2$ —that systematically transform measurement errors into positive biases. The required corrections are very similar to the classic “ $N/(N-1)$ ” correction for the naive variance estimator, with the additional difficulty that  $V$  is affected by *two* sources of noise: the finite number of recording trials, and the finite number of recorded neurons. The exact corrections to ensure an unbiased estimation of  $V$  are detailed in supporting [S1 Text](#), section 5.

Third, indicator  $q(u, t)$  displays little or no measurement bias—because its computation is essentially linear. Yet, it can display an important level of measurement noise ( $\text{Var}(\xi) \gg 0$  in the above discussion) that may deteriorate the subsequent inference procedure. We mitigated this measurement noise by applying a bi-temporal Gaussian smoothing to  $q^*(u, t)$  and predictions  $q(u, t)$ , with time constant 10 msec.

To estimate the measurement errors due to the finite number of trials, we produced 14 sets of surrogate data by sampling our original trials with replacement (bootstrap procedure). These resamplings were used to derive some of the correction terms for  $V$ , and *also* to derive confidence intervals on our final estimators, as shown in [Fig. 6](#). This departure from the statistical canon was imposed by the length of the whole inference procedure (see supporting [S1 Text](#), section 5, for details).

## Reproduction of our results and implementation

In the Supporting Information, we provide a generic implementation of the inference method (“Case 2” above) in MATLAB, which can be applied to any data from a 2AFC discrimination task. We also provide the Python code for the network simulation, and MATLAB scripts for the reproduction of the experimental Figures in this article ([Fig. 4–7](#)).

## Supporting Information

**S1 Text. Supporting text.** Contains additional information about Choice Probabilities (section 1), the influence of parameter  $w$  on stimulus sensitivity (section 2), the encoding neural network used for testing the method (section 3), the Bayesian regularization procedure on Fisher’s linear discriminant (section 4), unbiased computation of CC indicators in the presence of measurement noise (section 5), and an extended readout model with variable extraction time  $t_R$  (section 6).

(PDF)

**S1 Compressed file archive. Supporting code for the article.**

(GZ)

## Author Contributions

Conceived and designed the experiments: AW CM. Performed the experiments: AW. Analyzed the data: AW. Wrote the paper: AW CM.

## References

1. Renart A, Machens CK (2014) Variability in neural activity and behavior. *Current Opinion in Neurobiology* 25: 211–220. doi: [10.1016/j.conb.2014.02.013](https://doi.org/10.1016/j.conb.2014.02.013) PMID: [24632334](https://pubmed.ncbi.nlm.nih.gov/24632334/)
2. Mountcastle V, Steinmetz MA, Romo R (1990) Frequency discrimination in the sense of flutter: psychophysical measurements correlated with postcentral events in behaving monkeys. *Journal of Neuroscience* 10: 3032–3044. PMID: [2118947](https://pubmed.ncbi.nlm.nih.gov/2118947/)

3. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience* 12: 4745–4765. PMID: [1464765](#)
4. Werner G, Mountcastle V (1965) Neural activity in mechanoreceptive cutaneous afferents: Stimulus-response relations, weber functions, and information transmission. *Journal of Neurophysiology* 28. PMID: [14283062](#)
5. Talbot W, Darian-Smith I, Kornhuber H, Mountcastle V (1968) The sense of flutter-vibration: comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. *Journal of Neurophysiology* 31. PMID: [4972033](#)
6. Romo R, Salinas E (2003) Flutter discrimination: neural codes, perception, memory and decision making. *Nature Reviews Neuroscience* 4: 203–18. doi: [10.1038/nrn1058](#) PMID: [12612633](#)
7. Gold JI, Shadlen MN (2007) The neural basis of decision making. *Annual Review of Neuroscience* 30: 535–74. doi: [10.1146/annurev.neuro.29.051605.113038](#) PMID: [17600525](#)
8. Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience* 18: 3870–3896. PMID: [9570816](#)
9. Abbott LF, Dayan P (1999) The effect of correlated variability on the accuracy of a population code. *Neural computation* 11: 91–101. doi: [10.1162/089976699300016827](#) PMID: [9950724](#)
10. Averbach BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. *Nature Reviews Neuroscience* 7: 358–66. doi: [10.1038/nrn1888](#) PMID: [16760916](#)
11. Uka T, DeAngelis GC (2003) Contribution of middle temporal area to coarse depth discrimination: comparison of neuronal and psychophysical sensitivity. *Journal of Neuroscience* 23: 3515–30. PMID: [12716961](#)
12. Cohen MR, Newsome WT (2009) Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *Journal of Neuroscience* 29: 6635–48. doi: [10.1523/JNEUROSCI.5179-08.2009](#) PMID: [19458234](#)
13. Price NSC, Born RT (2010) Timescales of sensory- and decision-related activity in the middle temporal and medial superior temporal areas. *Journal of Neuroscience* 30: 14036–45. doi: [10.1523/JNEUROSCI.2336-10.2010](#) PMID: [20962225](#)
14. Green D, Swets J (1966) *Signal detection theory and psychophysics*, volume 1974. Wiley, New York, USA.
15. Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon AJ (1996) A relationship between behavioral choice and the visual response of neurons in macaque MT. *Visual Neuroscience* 13: 87–100. doi: [10.1017/S095252380000715X](#) PMID: [8730992](#)
16. de Lafuente V, Romo R (2006) Neural correlate of subjective sensory experience gradually builds up across cortical areas. *Proceedings of the National Academy of Sciences of the United States of America* 103: 14266–71. doi: [10.1073/pnas.0605826103](#) PMID: [16924098](#)
17. Shadlen MN, Britten KH, Newsome WT, Movshon AJ (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience* 16: 1486–1510.
18. Nienborg H, Cumming BG (2010) Correlations between the activity of sensory neurons and behavior: how much do they tell us about a neuron's causality? *Current opinion in neurobiology* 20: 376–381. doi: [10.1016/j.conb.2010.05.002](#) PMID: [20545019](#)
19. Haefner RM, Gerwinn S, Macke JH, Bethge M (2013) Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nature Neuroscience* 16: 235–242. doi: [10.1038/nn.3309](#) PMID: [23313912](#)
20. Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature* 459: 89–92. doi: [10.1038/nature07821](#) PMID: [19270683](#)
21. Hernández A, Zainos A, Romo R (2000) Neuronal correlates of sensory discrimination in the somatosensory cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97: 6191–6. doi: [10.1073/pnas.120018597](#) PMID: [10811922](#)
22. Aertsen AM, Gerstein GL, Habib MK, Palm G (1989) Dynamics of neuronal firing correlation: modulation of “effective connectivity”. *Journal of neurophysiology* 61: 900–17. PMID: [2723733](#)
23. Luna R, Hernández A, Brody CD, Romo R (2005) Neural codes for perceptual discrimination in primary somatosensory cortex. *Nature Neuroscience* 8: 1210–9. doi: [10.1038/nn1513](#) PMID: [16056223](#)
24. Ahrens MB, Orger MB, Robson DN, Li JM, Keller PJ (2013) Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods* 10: 413–420. doi: [10.1038/nmeth.2434](#) PMID: [23524393](#)



25. Panier T, Romano SA, Olive R, Pietri T, Sumbre G, et al. (2013) Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy. *Frontiers in neural circuits* 7. doi: [10.3389/fncir.2013.00065](https://doi.org/10.3389/fncir.2013.00065) PMID: [23576959](https://pubmed.ncbi.nlm.nih.gov/23576959/)
26. Portugues R, Feierstein CE, Engert F, Orger MB (2014) Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron* 81: 1328–1343. doi: [10.1016/j.neuron.2014.01.019](https://doi.org/10.1016/j.neuron.2014.01.019) PMID: [24656252](https://pubmed.ncbi.nlm.nih.gov/24656252/)
27. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning*. Springer Verlag, New York, USA.
28. Wohrer A, Humphries MD, Machens CK (2013) Population-wide distributions of neural activity during perceptual decision-making. *Progress in Neurobiology* 103: 156–193. doi: [10.1016/j.pneurobio.2012.09.004](https://doi.org/10.1016/j.pneurobio.2012.09.004) PMID: [23123501](https://pubmed.ncbi.nlm.nih.gov/23123501/)
29. Wohrer A, Romo R, Machens CK (2010) Linear readout from a neural population with partial correlation data. In: *Advances in Neural Information Processing*, volume 23, pp. 2469–2477.
30. Turaga S, Buesing L, Packer AM, Dagleish H, Pettit N, et al. (2013) Inferring neural population dynamics from multiple partial recordings of the same neural circuit. In: *Advances in Neural Information Processing Systems*. pp. 539–547.
31. Boerlin M, Machens CK, Denève S (2013) Predictive coding of dynamical variables in balanced spiking networks. *PLoS computational biology* 9: e1003258. doi: [10.1371/journal.pcbi.1003258](https://doi.org/10.1371/journal.pcbi.1003258) PMID: [24244113](https://pubmed.ncbi.nlm.nih.gov/24244113/)
32. Boerlin M, Denève S (2011) Spike-based population coding and working memory. *PLoS computational biology* 7. doi: [10.1371/journal.pcbi.1001080](https://doi.org/10.1371/journal.pcbi.1001080) PMID: [21379319](https://pubmed.ncbi.nlm.nih.gov/21379319/)
33. Schaub M, Schultz S (2012) The ising decoder: reading out the activity of large neural ensembles. *Journal of Computational Neuroscience* 32: 101–118. doi: [10.1007/s10827-011-0342-z](https://doi.org/10.1007/s10827-011-0342-z) PMID: [21667155](https://pubmed.ncbi.nlm.nih.gov/21667155/)
34. Cook EP, Maunsell JHR (2002) Dynamics of neuronal responses in macaque mt and vip during motion detection. *Nature neuroscience* 5: 985–994. doi: [10.1038/nn924](https://doi.org/10.1038/nn924) PMID: [12244324](https://pubmed.ncbi.nlm.nih.gov/12244324/)
35. Stanford TR, Shankar S, Massoglia DP, Costello MG, Salinas E (2010) Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience* 13: 379–385. doi: [10.1038/nn.2485](https://doi.org/10.1038/nn.2485) PMID: [20098418](https://pubmed.ncbi.nlm.nih.gov/20098418/)
36. Ashourian P, Loewenstein Y (2011) Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one* 6: e19551. doi: [10.1371/journal.pone.0019551](https://doi.org/10.1371/journal.pone.0019551) PMID: [21589867](https://pubmed.ncbi.nlm.nih.gov/21589867/)
37. Miura K, Mainen ZF, Uchida N (2012) Odor representations in olfactory cortex: distributed rate coding and decorrelated population activity. *Neuron* 74: 1087–1098. doi: [10.1016/j.neuron.2012.04.021](https://doi.org/10.1016/j.neuron.2012.04.021) PMID: [22726838](https://pubmed.ncbi.nlm.nih.gov/22726838/)
38. Daley D, Vere-Jones D (2007) *An introduction to the theory of point processes*, volume 1. Springer Verlag, New York, USA.
39. Goodman D, Brette R (2008) Brian: a simulator for spiking neural networks in python. *Frontiers in neuroinformatics* 2. doi: [10.3389/neuro.11.005.2008](https://doi.org/10.3389/neuro.11.005.2008) PMID: [19115011](https://pubmed.ncbi.nlm.nih.gov/19115011/)
40. Raudys S, Duin R (1998) Expected classification error of the fisher linear classifier with pseudoinverse covariance matrix. *Pattern Recognition Letters* 19: 385–392. doi: [10.1016/S0167-8655\(98\)00016-6](https://doi.org/10.1016/S0167-8655(98)00016-6)
41. Hoyle DC (2011) Accuracy of pseudo-inverse covariance learning—a random matrix theory analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33: 1470–1481. doi: [10.1109/TPAMI.2010.186](https://doi.org/10.1109/TPAMI.2010.186)
42. Bishop CM (2006) *Pattern recognition and machine learning*. Springer Verlag, New York, USA.