PLOS ONE

Maximally Informative "Stimulus Energies" in the Analysis of Neural Responses to Natural Signals

Kanaka Rajan*, William Bialek

Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America

Abstract

The concept of feature selectivity in sensory signal processing can be formalized as dimensionality reduction: in a stimulus space of very high dimensions, neurons respond only to variations within some smaller, relevant subspace. But if neural responses exhibit invariances, then the relevant subspace typically cannot be reached by a Euclidean projection of the original stimulus. We argue that, in several cases, we can make progress by appealing to the simplest nonlinear construction, identifying the relevant variables as quadratic forms, or "stimulus energies." Natural examples include non–phase–locked cells in the auditory system, complex cells in the visual cortex, and motion–sensitive neurons in the visual system. Generalizing the idea of maximally informative dimensions, we show that one can search for kernels of the relevant quadratic forms by maximizing the mutual information between the stimulus energy and the arrival times of action potentials. Simple implementations of this idea successfully recover the underlying properties of model neurons even when the number of parameters in the kernel is comparable to the number of action potentials and stimuli are completely natural. We explore several generalizations that allow us to incorporate plausible structure into the kernel and thereby restrict the number of parameters. We hope that this approach will add significantly to the set of tools available for the analysis of neural responses to complex, naturalistic stimuli.

Introduction

A central concept in neuroscience is feature selectivity: as our senses are bombarded by complex, dynamic inputs, individual neurons respond to specific, identifiable components of these data [1,2]. Neurons early in a processing pathway are thought to be sensitive to simpler features [3,4], and one can think of subsequent stages of processing as computing conjunctions of these features, so that neurons later in the pathway respond to more complex structures in the sensory world [5]. A major challenge for theory is to make this intuition mathematically precise, and to use such a precise formulation to build tools that allow us to analyze real neurons as they respond to naturalistic inputs. There is a long history of such work, but much of it rests on the identification of "features" with filters or templates. Filtering is a linear operation, and matching to a template can be thought of as a cascade of linear and nonlinear steps. As we will see, however, there are many examples of neural feature selectivity, well known from experiments on visual and auditory systems in many organisms, for which such a description in linear terms does not lead to much simplification.

In this paper we use examples to motivate the simplest nonlinear definition of a feature, in which the relevant variable is a quadratic form in stimulus space. Because the resulting variable is connected to the "energy in frequency bands" for auditory signals, we refer to these quadratic forms as "stimulus energies." To be useful, we have to be able to identify these

structures in experiments where neurons are driven by complex, naturalistic inputs. We show that, generalizing the idea of maximally informative dimensions [6], we can find the maximally informative stimulus energies using methods that don't require special assumptions about the structure of the input stimulus ensemble. We illustrate these ideas on model neurons, and explore the amount of data that will be needed to use these methods in the analysis of real neurons.

Motivation

To motivate the problems that we address, let us start by thinking about an example from the auditory system. This starting point is faithful to the history of our subject, since modern approaches for estimating receptive fields and filters have their origins in the classic work of de Boer and coworkers on the "reverse correlation" method [7], which was aimed at separating the filtering of acoustic signals by the inner ear from the nonlinearities of spike generation in primary auditory neurons. We will see that mathematically identical problems arise in thinking about complex cells in visual cortex, motion sensitive neurons throughout the visual pathway, and presumably in other problems as well.

We begin with the simplest model of an auditory neuron. If the sound pressure as a function of time is $s(t)$, it is plausible that the activity of a neuron is controlled by some filtered version of this stimulus, so that the probability per unit time of generating a spike is

$$r(t) = r_0 g \left[\int d\tau f(\tau) s(t-\tau) \right], \qquad (1)$$

where $f(\tau)$ is the relevant temporal filter and $g[\cdot]$ is a nonlinearity; the spikes occur at times $\{t_i\}$. The statement that neurons are tuned is that if we look at the filter in Fourier space,

$$\tilde{f}(\omega) = \int dt f(t) e^{i\omega t}, \qquad (2)$$

then the magnitude of the filter response, $|\tilde{f}(\omega)|$, has a relatively sharp peak near some characteristic frequency ω_c. If we choose the stimulus waveforms from a Gaussian white noise ensemble, then the key result of reverse correlation is that if we compute the average stimulus in the prior to a spike, we will recover the underlying filter, independent of the nonlinearity,

$$\langle s(t - t_i) \rangle \propto f(-t). \qquad (3)$$

We emphasize that this is a theorem, not a heuristic data analysis method [13–15]. If the conditions of the theorem are not met, then the spike–triggered average stimulus may provide a biased estimate of the filter. *If* the conditions of the theorem are met, then this analysis is guaranteed to give the right answer in the limit of large amounts of data.

As an aside, we note that the spike–triggered averaging method does not make assumptions about the statistics of the spike train itself. The theorem states that *if* the probability per unit time of generating single spikes can be written as in Eq. (1), and if stimuli are chosen from a Gaussian white noise ensemble, then spike–triggered averaging recovers the underlying filter. It might be that the probability of spiking is influenced by the time since previous spikes, but even if this is true there is, at each moment in time, a rate $r(t)$ that defines the probability of spiking knowing only the history of the stimulus, not the history of spiking. It is this stimulus–dependent probability that we are writing in Eq. (1).

Equation (1) is an example of dimensionality reduction. In principle, the neuron's response at time t can be determined by the entire history of the stimulus for times $t' \leq t$. Let us suppose that we sample (and generate) the stimulus in discrete time steps spaced by dt. Then the stimulus history is a list of numbers

$$\mathbf{s}_t \equiv \{s(t), s(t-dt), s(t-2dt), \cdots, s(t-(D-1)dt)\}, \qquad (4)$$

where D is the effective stimulus dimensionality, set by $D = T/dt$, and T the longest plausible estimate of the integration time for the neural response. We can think of $\mathbf{s}_t$ as a D–dimensional vector. If we know that the neural response is controlled by a linearly filtered version of the sound pressure stimulus, even followed by an arbitrary nonlinearity, then only one direction in this D–dimensional space matters for the neuron. Further this really is a "direction," since we can write the response as the Euclidean projection of $\mathbf{s}$ onto one axis, or equivalently the dot product between $\mathbf{s}$ and a vector $\mathbf{W}$,

$$r(t) = r_0 g(\mathbf{W} \cdot s_t), \qquad (5)$$

where

$$\mathbf{W} = dt \times \{f(0), f(dt), f(2dt), \cdots, f(T)\}. \qquad (6)$$

This explicit formulation in terms of dimensionality reduction suggests a natural generalization in which several dimensions, rather than just one, are relevant,

$$r(t) = r_0 g(\mathbf{W}_1 \cdot s_t, \mathbf{W}_2 \cdot s_t, \cdots, \mathbf{W}_K \cdot s_t). \qquad (7)$$

As long as we have $K \ll D$, it still holds true that the neuron responds only to some limited set of stimulus dimensions, but this number is not as small as in the simplest model of a single filter.

Notice that if an auditory neuron responds according to Eq. (1), then it will exhibit "phase locking" to periodic stimuli. Specifically, if $s(t) = A\cos(\omega t)$ and $\tilde{f}(\omega) = |\tilde{f}(\omega)| e^{+i\phi}$, then $r(t) = r_0 g[A|\tilde{f}(\omega)|\cos(\omega t - \phi)]$. So long as there is a nonzero response to the stimulus, this response will be modulated at the stimulus frequency ω, and more generally if we plot the spiking probability versus time measured by the phase $\psi = \omega t$ of the stimulus oscillation, then the probability will vary with, or "lock" to this phase.

While almost all auditory neurons are tuned, not all exhibit phase locking. We often summarize the behavior of tuned, non–phase–locked neurons by saying that they respond to the power in a given bandwidth or to the envelope of the signal at the output of a filter. The simplest model for such behavior, which has its roots in our understanding of hair cell responses [8–11], is to imagine that the output of a linear filter passes through a weak nonlinearity, then another filter. The second stage of filtering is low-pass, and will strongly attenuate any signals at or near the characteristic frequency ω_c. Then, to lowest order, the neuron's response depends on

$$p(t) = \int d\tau f_2(\tau) \left[\int dt' f_1(t - \tau - t') s(t') \right]^2, \qquad (8)$$

where f_1 is the bandpass filter that determines the tuning of the neuron and f_2 is a smoothing filter which ensures that the cell responds only to the slow variations in the power in its preferred frequency band. The probability of spiking depends on this power $p(t)$ through a nonlinearity, as before,

$$r(t) = r_0 g[p(t)]. \qquad (9)$$

Intuitively, this simple model for a non–phase–locked neuron also represents a substantial reduction in dimensionality – all that matters is the power passing through a given frequency band, defined by the filter f_1. On the other hand, we cannot collapse this model into the one dimensional form of Eq. (5). To be concrete, suppose that the filter f_1 has a relatively narrow bandwidth around its characteristic frequency ω_c. Then we can write

$$f_1(\tau) = A(\tau)\sin(\omega_c \tau + \phi), \qquad (10)$$

where the amplitude $A(\tau)$ varies slowly compared to the period of the oscillation. Let us denote by τ_1 the temporal width of $A(\tau)$, since this corresponds to the time over which the filter f_1 integrates the stimulus, and similarly τ_2 will denote the temporal width of $f_2(\tau)$. To make sure that the power $p(t)$ does not oscillate at (twice) the frequency ω_c, we must have $\tau_2 \gg 2\pi/\omega_c$. But we still have two possibilities, (a) $\tau_1 \gg \tau_2 \gg 1/\omega_c$ and (b) $\tau_2 \gg \tau_1 \gg 1/\omega_c$. If (a) is true and there is adequate separation in the two time scales, we can show that $p(t)$ is the Pythagorean sum of the outputs of two filters that form a quadrature pair,

$$p(t) \approx \left[\int dt' f_\alpha(t-t')s(t')\right]^2 + \left[\int dt' f_\beta(t-t')s(t')\right]^2, \quad (11)$$

where

$$f_\alpha(\tau) \approx A(\tau)\sin(\omega_c\tau + \phi) \text{ and, } f_\beta(\tau) \approx A(\tau)\cos(\omega_c\tau + \phi). \quad (12)$$

On the other hand, if (b) is true, there is no simple decomposition, and the minimum number of dimensions that we need to describe this model is $K \sim \tau_2/dt$, which can be quite large.

We can measure the number of relevant dimensions using the spike–triggered covariance matrix. Specifically, if the stimulus vector $\mathbf{s}_t$ has components $s_i(t)$, then we can form the matrix

$$\Delta C_{ij} = \langle s_i(t)s_j(t)\rangle_{t=t_{spike}} - \langle s_i(t)s_j(t)\rangle, \quad (13)$$

where in the first term we average over the arrival time of the spikes and in the second term we average over all time. If the spiking probability behaves as in Eq. (7), and we choose the stimulus from a Gaussian ensemble, then ΔC has K nonzero eigenvalues [12,13,16–18,29]. In Fig. 1 we schematize the model auditory neuron we have been describing, and in Fig. 2 we show the spike–triggered covariance analysis for models in the two limits, $\tau_2 \ll \tau_1$ and $\tau_2 \gg \tau_1$. Indeed we find that in the first case there are approximately two relevant dimensions, a quadrature pair of filters, whereas in the second case there are many relevant dimensions; these dimensions appear as temporally shifted and orthogonalized copies of the filter $f_1(t)$.

We can think of a neuron that does not phase lock as having an invariance: it responds to acoustic waveforms that have energy in a relevant bandwidth near ω_c, but it doesn't discriminate among signals that are shifted by small times. This invariance means that the cell is not just sensitive to one dimension of the stimulus, but to many, although these different dimensions correspond, in effect, to the same stimulus feature occurring at different times relative to the spike. Thus, we have a conflict between the notion of a "single feature" and the mathematical description of a "single dimension" via linear projection. The challenge is to provide a mathematical formulation that better captures our intuition. Before presenting a possible solution, let's see how the same problem arises in other cases.

Since the classical work of Hubel and Wiesel [19,20], cells in the primary visual cortex are sometimes classified as simple and complex (but see [46]). Although Hubel and Wiesel did not give a mathematical description of their data, in subsequent work, simple cells often have been described in the same way that we described the simplest auditory neuron in Eq. (1) [24,25]. If the light intensity falling on the retina varies in space ($\tilde{\mathbf{x}}$) and time (t) as $I(\tilde{\mathbf{x}},t)$, we can define a spatiotemporal receptive field $F(\tilde{\mathbf{x}},\tau)$ and approximate the probability that a simple cell generates a spike per unit time as

$$r(t) = r_0 \, g\left[\int d^2x \int d\tau \, F(\tilde{\mathbf{x}},\tau) \, I(\tilde{\mathbf{x}},t)\right]. \quad (14)$$

If, as before, we assume that the stimulus is generated in discrete time steps (movie frames) with spacing dt, and that the stimulus influences spikes only within some time window of duration T, then we can think of the stimulus at any moment in time as being the T/dt frames of the movie preceding that moment,

$$\mathbf{s}_t \equiv \{I(\tilde{\mathbf{x}},t), I(\tilde{\mathbf{x}},t-dt), I(\tilde{\mathbf{x}},t-2dt), \cdots, I(\tilde{\mathbf{x}},t-T)\}. \quad (15)$$

If the relevant region of space is within $d \times d$ pixels, then this stimulus vector lives in a space of dimension $D = d^2T/(dt)$, which can be enormous. As in the discussion above, Eq. (14) is a restatement of the hypothesis that only one direction in this space is relevant for determining the probability that the simple cell generates a spike, and "direction" is once again a Euclidean or linear projection.

For a complex cell, on the other hand, this single projection is inadequate (simple cells can also show sensitivity to more than one dimension [26]). Complex cells respond primarily to oriented edges and gratings, as do simple cells, but they have a degree of spatial invariance which means that their receptive fields cannot be mapped onto fixed zones of excitation and inhibition. Instead, they respond to patterns of light in a certain orientation within a large receptive field, regardless of precise location, or to movement in a certain direction. Corresponding to this intuition, analysis of complex cells using the spike–triggered covariance method shows that there is more than one relevant dimension [26]. As with non–phase–locked auditory neurons, what defeats the simplest version of dimensionality reduction in complex cells is the invariance of the response, in this case, invariance to small spatial displacement of the relevant, oriented stimulus feature [21–23].

The simplest model of a complex cell is precisely analogous to the quadrature pair of filters that emerge in the analysis of non–phase–locked auditory neurons. To be concrete, let us imagine that receptive fields are described by Gabor patches. The position $\tilde{\mathbf{x}}$ includes two orthogonal coordinates in visual space, which we call x_1 and x_2. Gabor patches have an oscillatory dependence on one of these coordinates, but simply integrate along the other; both the integration and the envelope of the oscillations are described by Gaussians, so that

$$F_1(\tilde{\mathbf{x}},\tau) = \cos(kx_1)\exp\left[-\left(\frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}\right)\right]f(\tau), \quad (16)$$

and the quadrature filter is then

$$F_2(\tilde{\mathbf{x}},\tau) = \sin(kx_1)\exp\left[-\left(\frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}\right)\right]f(\tau). \quad (17)$$

Each of these filters is maximally sensitive to extended features oriented along the x_2 direction, but the optimal patterns of light and dark are shifted for the two filters; for simplicity we have assumed that spatial and temporal filtering is separable. If we form the energy–like quantity

$$p(t) = \left[\int d^2x \int d\tau \, F_1(\mathbf{x},\tau) \, I(\mathbf{x},t)\right]^2 + \left[\int d^2x \int d\tau \, F_2(\mathbf{x},\tau) \, I(\mathbf{x},t)\right]^2, \quad (18)$$

we have a measure of response to oriented features independent of their precise position, and this provides a starting point for the analysis of a complex cell.

One more example of the problem we face is provided by the computation of motion in the visual system. There is a classical model for this computation, the correlator model, that grew out of the experiments by Hassenstein and Reichardt on behavioral
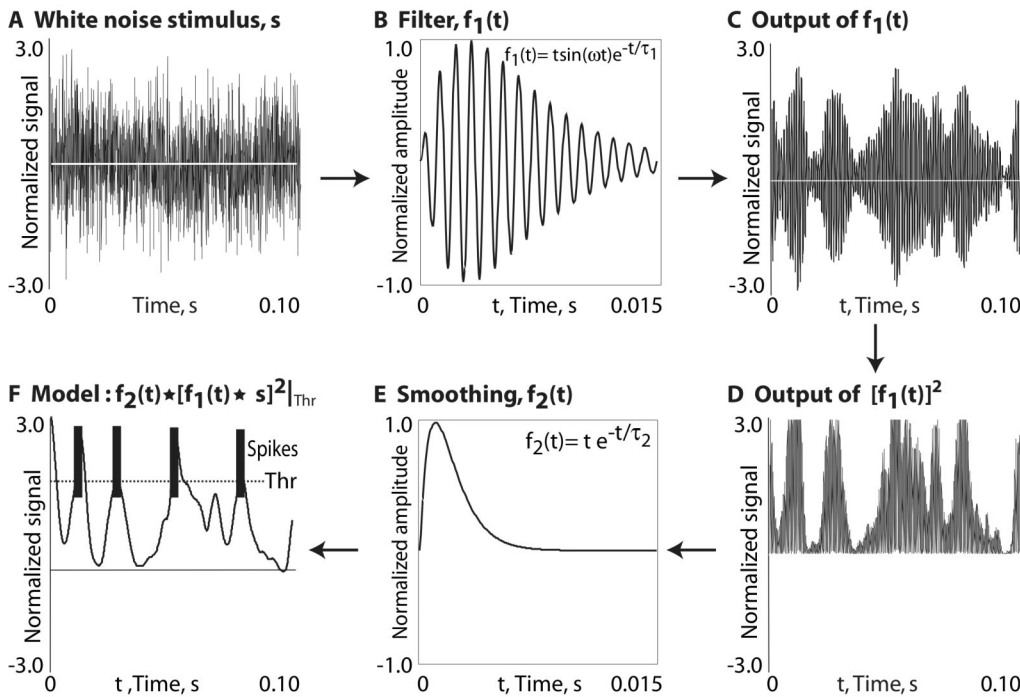
A White noise stimulus, s

B Filter, $f_1(t)$

$$f_1(t) = t\sin(\omega t)e^{-t/\tau_1}$$

C Output of $f_1(t)$

F Model: $f_2(t) \star [f_1(t) \star s]^2 \big|_{Thr}$

Spikes

Thr

E Smoothing, $f_2(t)$

$$f_2(t) = t\, e^{-t/\tau_2}$$

D Output of $[f_1(t)]^2$

Figure 1. Schematic of a non–phase–locked auditory neuron. (a) In this implementation a model auditory neuron responds to a white noise stimulus. (b) The stimulus **s** is filtered through a temporal filter f_1, which has the form $t\sin(\omega t)\exp(-t/\tau_1)$ where $\omega = 2\pi \times 10^3\,\mathrm{Hz}$ and $\tau_1 = 3\,\mathrm{ms}$. (c) The output of f_1 is shown here. The filter f_1 is narrow band, therefore the output oscillates at the characteristic frequency even when the input is white. (d) The output of f_1 is first squared, and in (e), convolved with a second filter f_2 of the form $t\exp(-t/\tau_2)$, with a smoothing time constant $\tau_2 = 1$ ms. (f) The normalized signal is finally thresholded to generate spikes. We assume that time runs discretely in steps of $dt = 1/20000\,\mathrm{s}$.
doi:10.1371/journal.pone.0071959.g001

responses to visual motion in beetles and flies [28]. Briefly, the idea behind this model is that if something is moving at velocity v, then the image intensity $I(x,t)$ must be correlated with the intensity at $I(x+\Delta, t+\tau)$, where $\tau = \Delta/v$. Then we can detect motion by computing this correlation and averaging over some window W in space and f in time,

$$\mathbf{C}_{\Delta,\tau}(t) = \int dx\, W(x) \int dt'\, f(t-t')I(x,t')I(x+\Delta, t'+\tau), \quad (19)$$

where for simplicity, we think just about a single spatial dimension. In principle, with just one value of the delay τ and one value of the displacement Δ, this correlation "detects" only motion at one velocity, $v = \Delta/\tau$, but we can easily generalize this computation to,

$$\mathbf{C}(t) = \int dx\, W(x) \int dt'\, f(t-t') \int d\tau \sum_{\Delta} F(\Delta,\tau)I(x,t')I(x+\Delta, t'+\tau), \quad (20)$$

where the sum over $F(\Delta,\tau)$ operates over different values of these spatiotemporal shifts. Depending on the precise form of this weighting we can arrange for the correlation **C** to have a relatively smooth, graded dependence on the velocity.

In the insect visual system it seems natural to think of the correlations in Eq. (20) as being computed from the outputs of individual photoreceptors, which are typically spaced $\sim 1°$ apart. In mammals, we can imagine computing a similar quantity, but we would use the outputs of larger retinal or cortical receptive fields [27]. We can also think about this computation in Fourier space. If we transform

$$I(x,t) = \int \frac{dk}{2\pi} \int \frac{d\omega}{2\pi}\, \tilde{I}(k,\omega)\, e^{ikx - i\omega t}, \quad (21)$$

then we can think of **C** as integrating power or energy $P(k,\omega) = |\tilde{I}(k,\omega)|^2$ over some region in the $k-\omega$ plane; motion corresponds to having this power concentrated along the line $\omega = vk$. Once again there is an invariance to this computation, since as with the non–phase–locked auditory neuron we are looking for power regardless of phase. More directly, in this case the brain is trying to compute a velocity, more or less independently of the absolute position of the moving objects. Even in an insect visual system, where computing **C** corresponds to correlating the filtered outputs of neighboring photoreceptors in the compound eye, this computation is repeated across an area containing many photoreceptors, and hence there is no way to collapse **C** down to a function of just one or two Euclidean projections in the stimulus space.

What do these examples – non–phase–locked auditory neurons, complex cells in the visual cortex, and visual motion detectors – have in common? In all three cases, the natural, simplest starting point is a model in which the brain computes not a linear projection of the stimulus onto a receptive field, but rather a quadratic form. More precisely, if stimuli are the vectors $\mathbf{s} \equiv \{s_1, s_2, \cdots s_D\}$, then Eq's. (8), (18) and (20) all correspond to computing a quantity

$$x = \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s} \equiv \sum_{i,j=1}^{D} s_i\, Q_{ij}\, s_j. \quad (22)$$
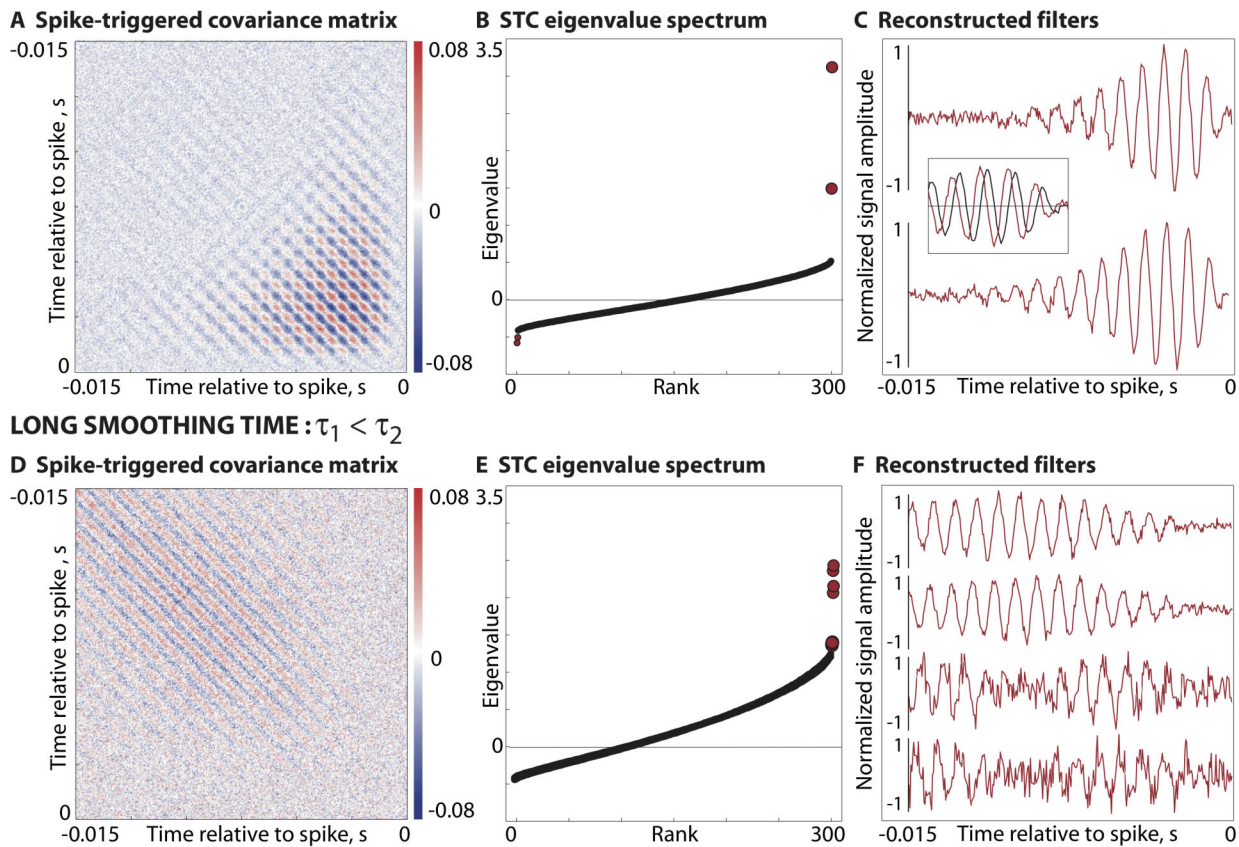
A Spike-triggered covariance matrix

B STC eigenvalue spectrum

C Reconstructed filters

LONG SMOOTHING TIME $: \tau_1 < \tau_2$

D Spike-triggered covariance matrix

E STC eigenvalue spectrum

F Reconstructed filters

Figure 2. Covariance analysis of the non–phase–locked auditory neuron. (a) If the time constant of the smoothing filter f_2 is much shorter than that of the filter f_1, the spike–triggered covariance matrix has a relatively simple structure. Here, $\tau_1 = 3$ms and $\tau_2 = 1$ms. (b) Diagonalizing this covariance matrix yields 2 leading eigenvalues, enlarged for clarity (the rest remain close to 0). (c) The eigenvectors corresponding to the 2 non-zero eigenvalues are the reconstructed filters plotted here. These form a quadrature pair, as shown in the inset. (d) If the smoothing time of the filter f_2 is larger than that of the first, the covariance matrix has a much richer structure. Here, $\tau_1 = 3$ms and $\tau_2 = 30$ms. (e) The spike–triggered covariance matrix decomposes into multiple non-unique eigenvalues, enlarged for clarity. (f) The eigenvectors corresponding to the non-zero eigenvalues give multiple time-shifted copies of the same filter.
doi:10.1371/journal.pone.0071959.g002

Generalizing the use of terms such as "energy in a frequency band" and "motion energy," we refer to x as a "stimulus energy."

If the matrix **Q** is of low rank, this means that we can build x out of projections of the stimulus onto a correspondingly low dimensional Euclidean subspace, and we can try to recover the relevant structure using methods such as the spike–triggered covariance or maximally informative dimensions. Still, if **Q** is of rank 5 (for example), once we identify the five relevant dimensions we might run out of data before we can calculate the nonlinear input/output relation of the neuron (this problem has also been addressed in [39]) and confirm our intuition that there is indeed a single stimulus feature x to which the cell responds.

In many cases, it is plausible that neurons could be responding just to one energy x, but the underlying matrix **Q** could be of full rank; a clear example is provided by correlator models for wide–field motion sensitive neurons, as in [28,29]. But in this case there is no real "dimensionality reduction" associated with the mapping from $\mathbf{s} \rightarrow x$, if all we know how to do is to search for linear projections or Euclidean subspaces. On the other hand, the mapping really is a tremendous reduction in complexity, because the full stimulus **s** is described by D parameters, while x is just one number.

Suppose that the response of a neuron to complex stimuli can be described by saying that the probability of spiking depends solely on a single stimulus energy x as in Eq. (22), so that

$$r(t) = r_0 g(\mathbf{s}_t{}^T \cdot \mathbf{Q} \cdot \mathbf{s}_t). \tag{23}$$

Our task becomes one of showing that we can recover the underlying matrix **Q** by analyzing the spike train in relation to the stimuli, without making any assumptions about the statistics of the stimuli.

Methods

Core of the method

If the probability of generating an action potential depends on the stimulus **s**, then observing the arrival of even a single spike provides information about the stimulus. Importantly, the data processing inequality [30] tells us that if we look not at the full stimulus but only at some limited or compressed description of the stimulus – a single feature, for example – we can only lose information. If the neuron really is sensitive to only one stimulus feature, however, we lose none of the available mutual information between spikes and stimuli by focusing on this one feature. This

suggests a natural strategy for identifying relevant stimulus features, by searching for those which preserve the maximum amount of information [6]. Further, we can put the success of such a search on an absolute scale, by estimating more directly the information that spikes provide about the stimulus [31,33], and asking what fraction of this information is captured by the best feature we could find.

Setting up the problem

To make this precise, we recall that the information about the stimulus **s** that is conveyed by the observation of a single spike can be written, following [32], as

$$I_{\text{spike}} = \int d^D\mathbf{s}\, P(\mathbf{s}|\text{spike}) \log_2\left[\frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})}\right] \text{bits}, \qquad (24)$$

where $P(\mathbf{s})$ is the stimulus distribution and $P(\mathbf{s}|\text{spike})$ is the distribution of stimuli given that a spike occurred at a particular time (the response conditional ensemble [18]). If we consider a mapping $\mathcal{M} : \mathbf{s} \rightarrow x$, then we can also compute

$$I_{\text{spike}}(\mathcal{M}) = \int dx\, P(x|\text{spike}) \log_2\left[\frac{P(x|\text{spike})}{P(x)}\right], \qquad (25)$$

knowing that for any mapping $\mathcal{M}$, $I_{\text{spike}}(\mathcal{M}) \le I_{\text{spike}}$. If we restrict ourselves to some class of mappings, then we can search for an $\mathcal{M}^*$ which maximizes $I_{\text{spike}}(\mathcal{M})$, and see how close this is to the real value of I_{spike}, provided this can be computed from data [31]. While I_{spike} is an integral over all stimuli, and involves distributions over the full, high–dimensional, stimulus space, $I_{\text{spike}}(\mathcal{M})$ only involves distributions of a single variable x. If $\mathcal{M}$ is fixed, it is relatively easy for experiments to have sufficient data that $P(x)$ and $P(x|\text{spike})$ will be sampled well. Thus, as first emphasized in [6], searching for maximally informative features provides a practical and principled way to analyze neural responses to high dimensional, naturalistic stimuli.

The work in [6] considered the case where the stimulus features are one or more linear projections of the stimulus, $x_\alpha = \mathbf{W}_\alpha \cdot \mathbf{s}$, so that the mapping $\mathcal{M}$ is parameterized by the vectors $\{\mathbf{W}_\alpha\}$; in the simplest case there being just one vector $\mathbf{W}$. Here we are interested in quadratic mappings, corresponding to the stimulus energies in Eq. (22). Now the mapping $\mathcal{M}$ is parameterized by a symmetric matrix $\mathbf{Q}$. In principle, all the arguments of [6] for the linear case should hold here because the quadratic form for a full rank symmetric matrix can be rewritten as a dot product between the vectorized matrix and the vectorized product $\mathbf{s}\mathbf{s}^T$. This is equivalent to the "kernel trick" in machine learning, but we note that viewing the matrix $\mathbf{Q}$ in this way, we lose track of all its structure. Before starting, we note an obvious problem related to the number of parameters we are looking for. If we are searching for a vector $\mathbf{W}$ in a D–dimensional stimulus space, we are looking an object described by D numbers. In fact, the length of the vector is irrelevant, so that maximizing $I_{\text{spike}}(\mathcal{M})$ corresponds to optimization in a $D-1$ dimensional space. But if we are searching for a symmetric matrix that acts on stimulus vectors, there are $D(D+1)/2-1$ free parameters. This is a problem both because we have to optimize in a space of much higher dimensionality, and because determining more parameters reliably must require larger data sets. We will address these problems shortly, but let's start by following the path laid out in [6], which involves searching for the maximum of $I_{\text{spike}}(\mathcal{M})$ by gradient ascent.

If our stimulus feature is the energy in Eq. (22), then the distribution of x is

$$P_{\mathbf{Q}}(x) = \int d^D s\, \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s})\, P(\mathbf{s}), \qquad (26)$$

where the subscript explicitly denotes that $P(x)$ depends on $\mathbf{Q}$. We take the derivative of this with respect to an element Q_{ij} in the matrix $\mathbf{Q}$,

$$\frac{\partial P_{\mathbf{Q}}(x)}{\partial Q_{ij}} = \int d^D s\, \frac{\partial}{\partial Q_{ij}} \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) \qquad (27)$$

$$= -\int d^D s\, s_i s_j\, \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s})\, P(\mathbf{s}). \qquad (28)$$

Similarly, we can differentiate the distribution of x conditional on a spike,

$$\frac{\partial P_{\mathbf{Q}}(x|\text{spike})}{\partial Q_{ij}} = -\int d^D s\, s_i s_j\, \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s})\, P(\mathbf{s}|\text{spike}). \qquad (29)$$

Putting these terms together, we can differentiate the information:

$$\frac{\partial I_{\text{spike}}(\mathcal{M})}{\partial Q_{ij}} =$$

$$\int dx \left[\frac{\delta I_{\text{spike}}(\mathcal{M})}{\delta P_{\mathbf{Q}}(x)} \frac{\delta P_{\mathbf{Q}}(x)}{\partial Q_{ij}} + \frac{\delta I_{\text{spike}}(\mathcal{M})}{\delta P_{\mathbf{Q}}(x|\text{spike})} \frac{\partial P_{\mathbf{Q}}(x|\text{spike})}{\partial Q_{ij}} \right]$$

$$= \frac{1}{\ln 2} \int dx \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \int d^D s\, s_i s_j \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s})$$

$$- \frac{1}{\ln 2} \int dx \left(1 + \ln\left[\frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)}\right] \right) \int d^D s\, s_i s_j \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}) \qquad (30)$$

$$= -\frac{1}{\ln 2} \int dx \int d^D s\, s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) \frac{d}{dx}\left[\frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)}\right]$$

$$+ \frac{1}{\ln 2} \int dx \frac{P_{\mathbf{Q}}(x)}{P_{\mathbf{Q}}(x|\text{spike})} \int d^D s\, s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}) \qquad (31)$$

$$\frac{d}{dx}\left[\frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)}\right].$$

But we notice that

$$\int d^D s\, s_i s_j\, \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) = \langle s_i s_j | x \rangle\, P_{\mathbf{Q}}(x), \qquad (32)$$

where $\langle s_i s_j | x \rangle$ is the expectation value of $s_i s_j$ conditional on the value of the stimulus energy x, and similarly

$$\int d^D s\, s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}) =$$
$$\langle s_i s_j | x, \text{spike} \rangle P_{\mathbf{Q}}(x|\text{spike}), \qquad (33)$$

where $\langle s_i s_j | x, \text{spike} \rangle$ is the expectation value conditional on the energy x *and* the occurrence of a spike. We can combine these terms to give

$$\frac{\partial I_{\text{spike}}(\mathcal{M})}{\partial Q_{ij}} =$$

$$\int dx\, P_{\mathbf{Q}}(x)\left[\langle s_i s_j | x, \text{spike}\rangle - \langle s_i s_j | x\rangle\right]\frac{d}{dx}\left[\frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)}\right], \quad (34)$$

or, more compactly,

$$\nabla_{\mathbf{Q}} I = \int dx\, P_{\mathbf{Q}}(x)\left[\langle \mathbf{ss}^{\mathrm{T}}|x,\text{spike}\rangle - \langle \mathbf{ss}^{\mathrm{T}}|x\rangle\right]\frac{d}{dx}\left[\frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)}\right]. \quad (35)$$

To learn the maximally informative energy, or the best choice of the matrix $\mathbf{Q}$, we can ascend the gradient in successive learning steps,

$$\mathbf{Q} \rightarrow \mathbf{Q} + \gamma\, \nabla_{\mathbf{Q}} I \quad \text{where } \gamma \text{ is small.} \quad (36)$$

Multiple matrices

In the same way that the idea of linear projection can be generalized to have the probability of spiking depend on multiple linear projections, we can generalize to the case where the are multiple relevant stimulus energies. Perhaps the simplest example Eq. (23) can be generalized to 2 matrices, is the computation of a (regularized) ratio between two stimulus energies, such that,

$$r = r_0 g\left(\frac{\mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_1 \cdot \mathbf{s}}{1 + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_2 \cdot \mathbf{s}}\right). \quad (37)$$

Some biological examples of this formulation include gain control or normalization in V1 [26], optimal estimation theory of motion detection in visual neurons of insects [29] and complex spectro-temporal receptive fields of neurons responsible for song processing in songbirds [34,35]. The inference task becomes one of estimating both matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ by information maximization.

As before, we can compute the gradient, noting that this time, there are two different gradients of $I_{\text{spike}}(\mathcal{M})$,

$$\nabla_{\mathbf{Q}_1} I = \int dx\, P_{\mathbf{Q}_1}(x)\left[\langle \frac{\mathbf{ss}^{\mathrm{T}}}{1 + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_2 \cdot \mathbf{s}}\Big| x, \text{spike}\rangle\right.$$

$$\left. - \langle \frac{\mathbf{ss}^{\mathrm{T}}}{1 + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_2 \cdot \mathbf{s}}\Big| x\rangle\right]\frac{d}{dx}\left[\frac{P_{\mathbf{Q}_1}(x|\text{spike})}{P_{\mathbf{Q}_1}(x)}\right],$$

$$\nabla_{\mathbf{Q}_2} I = - \int dx\, P_{\mathbf{Q}_2}(x)\left[\langle \frac{\mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_1 \cdot \mathbf{s}}{(1 + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_2 \cdot \mathbf{s})^2}\, \mathbf{ss}^{\mathrm{T}}\Big| x, \text{spike}\rangle\right.$$

$$\left. - \langle \frac{\mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_1 \cdot \mathbf{s}}{(1 + \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q}_2 \cdot \mathbf{s})^2}\, \mathbf{ss}^{\mathrm{T}}\Big| x\rangle\right]\frac{d}{dx}\left[\frac{P_{\mathbf{Q}_2}(x|\text{spike})}{P_{\mathbf{Q}_2}(x)}\right] \quad (38)$$

Analogous to Eq. (36), at every learning step, we update each matrix $\mathbf{Q}_i$ by the appropriate i^{th} gradient,

$$\mathbf{Q}_i \rightarrow \mathbf{Q}_i + \gamma \nabla_{\mathbf{Q}_i} I. \quad (39)$$

where $i = 1, 2$ for the x in Eq. (37). In principle the formalism in Eq. (37) could yield a more complete description of a neuron's nonlinear response properties compared to a single kernel, but

there are some data-requirement challenges which we will address later.

Technical aspects of optimization

In order to implement Eq. (36) as an algorithm, we have to evaluate all the relevant probability distributions and integrals. In practice, this means computing x for all stimuli, choosing an appropriate binning along the x–axis, and sampling the binned versions of the spike–triggered and prior distributions. We compute the expectation values $\langle \mathbf{ss}^{\mathrm{T}}\rangle$ separately for each bin, approximate the integrals as sums over the bins, and derivatives as differences between neighboring bins. To deal with local extrema in the objective function, we use a large starting value of γ and gradually decrease γ during learning. This basic prescription can be made more sophisticated, but we do not report these technical improvements here. An example of these ideas is shown in Fig. 3. We used very small patches from natural images as inputs, reshaping the intensities in nearby pixels into a D–component stimulus vector $\mathbf{s}$ where $D = 10$. To describe the neuron we chose a random symmetric matrix $\mathbf{K}$ to use as the kernel, and generated spikes when the stimulus energy, $\mathbf{s}^{\mathrm{T}} \cdot \mathbf{K} \cdot \mathbf{s}$, crossed a threshold, as illustrated in Fig. 3a. We fixed the spiking threshold such that the fraction of bins containing spikes is 0.1, and we generated ~ 1000 spikes. We then tried to extract the neuron's receptive field by starting with a random initial matrix $\mathbf{Q}$, and following the gradient of mutual information, as in Eq. (36).

We let the one parameter of the algorithm, γ, gradually decrease from a starting value of 0.5 to 0.05, in order to minimize the fluctuations around the true maximum of the information. The relatively large initial value of γ allows us to avoid simulated annealing by letting the simulation hop over false maxima. The large value of γ also explains why the information settles at a value slightly lower than its peak. The solution bounces back and forth from one side of this peak to the other, plateauing at a lower value than the true maximum. After this approximate solution is found, it is improved by continuing the procedure with smaller γ values. We normalize the resulting matrix to unit length after each gradient step.

Mutual information, the red trace in Fig. 3b, peaks at the 40th learning step and remains unchanged after that. The 3 black dots in Fig. 3b correspond to the steps during the optimization when we extract and plot the initial guess, the intermediate and the optimal/maximally informative matrix $\mathbf{Q}$. It is interesting to note that the intermediate matrix appears completely different from the optimal $\mathbf{Q}$ even though the corresponding mutual information is relatively close to its maximum (a similar observation was made in the context of maximally informative dimensions [6]). In general the number of steps required for the mutual information to peak and plateau depends inversely on the smoothness of the true matrix of the neuron and the amount of data available, and directly with the stimulus dimensionality. While these are only trends, a more concrete treatment of this exists in the linear case, shown in [6].

In Fig. 3c the root–mean–square (RMS) reconstruction error $\langle (\mathbf{K} - \mathbf{Q})^2\rangle^{1/2}$ is plotted as a function of the number of learning steps for a randomly initialized $\mathbf{Q}$ (solid line) and when $\mathbf{Q}$ is initialized to the spike–triggered covariance (STC) matrix (dashed line). RMS error at the start of the algorithm ≈ 1 when the "true" matrix $\mathbf{K}$ and the initial guess for $\mathbf{Q}$ are symmetric, random matrices, uncorrelated with each other, but is slightly lower when $\mathbf{Q}$ is initialized to the STC. This difference becomes smaller as the stimulus dimensionality D increases or as the stimulus departs more strongly from Gaussianity. Both traces decrease, and stop
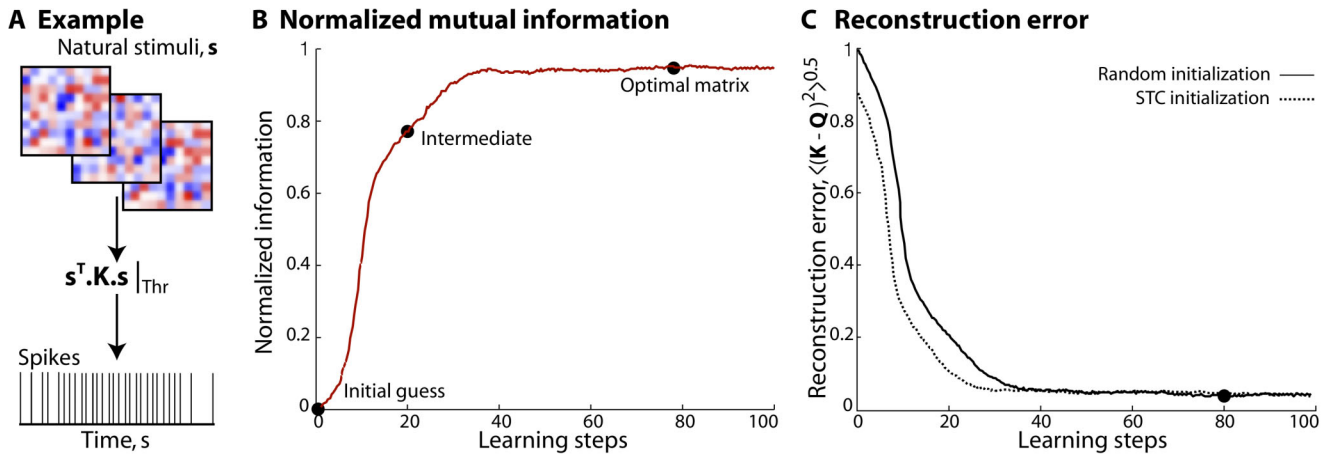
A Example
Natural stimuli, **s**

$s^T \cdot K \cdot s \big|_{Thr}$

Spikes

Time, s

B Normalized mutual information

Optimal matrix

Intermediate

Initial guess

Learning steps

C Reconstruction error

Random initialization ———
STC initialization ········

Learning steps

Figure 3. Core of the method. (a) A general implementation is shown here. The stimuli **s** are natural image clips which are $D \times D$ pixel patches resized from a natural image database, as described in [36]. 1000 spikes are generated with a probability per time bin of 0.1 from the model neuron by a thresholding the term, $s^T \cdot K \cdot s$ where the 10×10 matrix **K** is the receptive field of the neuron. **(b)** Mutual information between the spiking response of the model neuron and the quadratic stimulus projection x is plotted as a function of the number of learning steps. Information, normalized by its value when $K = Q$, peaks at the 40th learning step and then plateaus. The 3 black dots on the trace denote the points at which we extract the initial, the intermediate and the optimal matrices for comparison. The maximally informative matrix **Q** reconstructed at the 80th step, agrees well with **K**, indicating convergence. For this implementation the step size $\gamma = 0.5$ at the start and 0.05 at the end of the algorithm. **(c)** Root–mean–square (RMS) reconstruction error calculated as $\langle (K - Q)^2 \rangle^{1/2}$, is plotted as a function of the number of learning steps. This error decreases steadily until either the randomly initialized matrix (solid line) or the matrix initialized to the spike–triggered covariance matrix (dashed line) matches **K**. If **Q** is initialized to the covariance matrix, the initial RMS error is smaller and the convergence is faster (30th learning step) compared to that for a randomly initialized **Q**. For this example, both **K** and **Q** are 10×10 matrices and the black dot on the solid trace is at the same learning step as in panel (b).
doi:10.1371/journal.pone.0071959.g003

changing once our estimate of the optimal **Q** matches **K**. This occurs at the 40th step for the randomly initialized **Q** and slightly sooner (30th step) when initialized to the STC. If we had fewer spikes, our estimate for the optimal **Q** could still match **K** adequately, but the actual RMS error in reconstruction will be higher. We explore such performance measures and data requirement issues next.

Performance measures and data requirements

We expect the the accuracy of our reconstruction will depend of the number of spikes N_{spikes}. Further, we expect that when the stimulus dimensionality D increases, the fact that there are more parameters needed to describe the kernel **K** means that performance will deteriorate unless we have access to more data. To address these issues we explore model neurons as in Fig. 3, but systematically vary D and N_{spikes}. Again we consider the most difficult case, with naturalistic stimuli and kernels that are random symmetric matrices with elements drawn from a Gaussian distribution. The results are shown in Fig. 4. Our intuition is that the number of spikes should scale with the number of free parameters in the model, $D(D+1)/2$, so we always start with $N_{spikes} = D(D+1)/2$. If we scale N_{spikes} by the number of free parameters, the data collapse, as shown in Fig. 4b. Evidently accurate reconstructions of very large matrices or multiple energies requires very many spikes. This suggests that it will be important to have some more constrained models, which we now explore.

Constrained frameworks

The most general stimulus energy is described by $D(D+1)/2$ parameters, and this quickly becomes large for high dimensional stimuli. In many cases it is plausible that the matrix kernel of the stimulus energy has some simpler structure, which can be used to reduce the number of parameters.

One way to simplify the description is to use a matrix that has low rank. If, for example, the rank of the matrix **Q** in Eq. (22) is $p \leq D$, then we can find a set of orthogonal vectors $\{v_i\}$ such that

$$Q = \sum_{i=1}^{p} v_i v_i^{\mathrm{T}}. \quad (40)$$

In terms of these vectors, the stimulus energy is just $x = \sum_{i=1}^{p} (v_i^{\mathrm{T}} \cdot s)^2$.

The low rank approximation reminds us of the simpler, Euclidean notion of dimensionality reduction discussed above. Thus, we could introduce variables $x_i = v_i^{\mathrm{T}} \cdot s$ for $i = 1, 2, \ldots, p$. The response would then be approximated as depending on all of these variables, $r(x_1, x_2, \ldots, x_p)$, as in Eq. (7). In the stimulus energy approach, all of these multiple Euclidean projections are combined into $x = \sum_i x_i^2$, so that have a more constrained but potentially more tractable description. When **Q** is written as $Q = \sum_{i=1}^{p} v_i v_i^{\mathrm{T}}$, the relevant gradient of information, analogous to Eq. (35) is
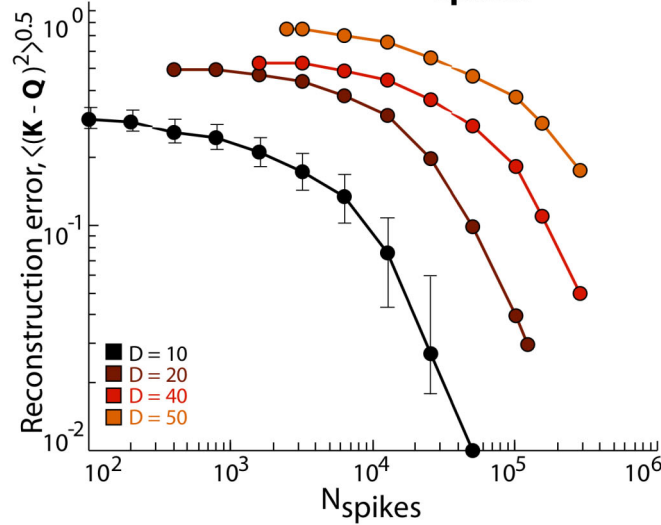
$$\nabla_{v_i} I =$$
$$2 \int dx \, P_Q(x) \left[\langle s(v_i^{\mathrm{T}} s) | x, \mathrm{spike} \rangle - \langle s(v_i^{\mathrm{T}} s) | x \rangle \right] \frac{d}{dx} \left[\frac{P_Q(x|\mathrm{spike})}{P_Q(x)} \right], \quad (41)$$

and we can turn this into an algorithm for updating our estimates of the v_i,

$$v_i \rightarrow v_i + \gamma \nabla_{v_i} I. \quad (42)$$

There is a free direction for the overall normalization of the matrix **Q** [6,26] which makes the mutual information invariant to reparameterization of the quantities. To be sure that Eq. (40) is

A Reconstruction error vs. N$_{spikes}$
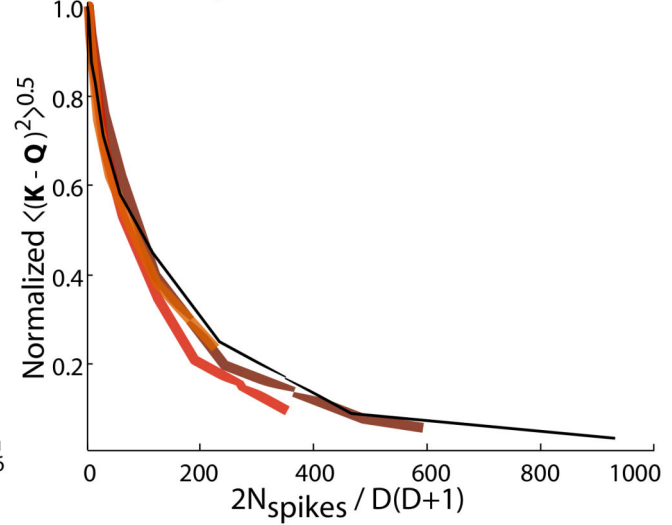

B Data collapse

Figure 4. Data requirement and performance issues. (a) Reconstruction error $\langle(\mathbf{K}-\mathbf{Q})^2\rangle^{1/2}$ is plotted as a function of number of spikes (N_{spikes}) for matrices corresponding to stimuli of increasing dimensionality ($D = 10, 20, 40$ & 50). Both matrices are normalized consistently. **(b)** The traces for different values of D collapse when the error is normalized by the (maximum) value at $N_{spikes} = D(D+1)/2$ for each D, and plotted as a function of $2N_{spikes}/(D(D+1))$.
doi:10.1371/journal.pone.0071959.g004

sufficiently general, we cannot constrain these vectors to be of unit length.

Another way of constraining the kernel of the stimulus energy is to assume that it is smooth as we move from one stimulus dimension to the next. Smooth matrices can be expanded into weighted sums of basis functions,

$$\mathbf{Q} = \sum_{\mu=1}^{M} \alpha_\mu \, \mathbf{B}^{(\mu)}, \qquad (43)$$

and finding the optimal matrix then is equivalent to calculating the most informative M–dimensional vector of weights.

The basis can be chosen so that systematically increasing the number of basis components M allows the reconstruction of progressively finer features in $\mathbf{Q}$. For example, we can consider $\{\mathbf{B}^{(\mu)}\}$ to be a set of Gaussian bumps tiling the $D \times D$ matrix $\mathbf{Q}$, and whose scale (standard deviation) is inversely proportional to $\sqrt{M}$. For $M \rightarrow D^2/2$ the basis matrix set becomes a complete basis, allowing every $\mathbf{Q}$ to be exactly represented by the vector of coefficients α. In any matrix basis representation, the learning rule becomes,

$$\alpha_\mu \rightarrow \alpha_\mu + \gamma \sum_{i,j=1}^{M} \frac{\partial I}{\partial \mathbf{Q}_{ij}} \mathbf{B}_{ij}^{(\mu)}. \qquad (44)$$

This is equivalent to taking projections of our general learning rule, Eq. (36), onto the basis elements.

Results

A model auditory neuron

As a first example, we return to the model auditory neuron whose response properties from Eq's (8) and (9) were schematized in Fig. 1. Rather than studying its responses to white noise stimuli, however, we consider the responses to bird song, as shown in

Fig. 5a. We start with Eq. (8) and see that it is equivalent to a stimulus energy with kernel $\mathbf{K}$ defined through

$$p(t) = \int dt_1 \int dt_2 \, s(t_1) \, \mathbf{K}(t-t_1, t-t_2) \, s(t_2), \qquad (45)$$

$$\mathbf{K} = \int d\tau f_2(\tau) f_1(t-\tau-t_1) f_1(t-\tau-t_2). \qquad (46)$$

We used the same filters as we showed in Fig. 1b and Fig. 1e to construct $\mathbf{K}$, which is plotted in Fig. 5b. We can also look in a mixed time–frequency representation to generate a spectro-temporal "sensitivity," $\tilde{\mathbf{K}}$, as follows:

$$\tilde{\mathbf{K}}(\omega,t) = \int d\tau \, \mathbf{K}\left(t+\frac{\tau}{2}, t-\frac{\tau}{2}\right) e^{+i\omega\tau}. \qquad (47)$$

$\tilde{\mathbf{K}}$ is a characteristic of the model neuron and describes its selectivity to the stimulus. We see that this neuron responds to sound with frequencies around 1kHz with a temporal dependence dictated by the time constants of the 2 filters that make up the neuron's receptive field matrix $\mathbf{K}$, as shown in Fig. 5c. This description has the flavor of a spectrotemporal receptive field (STRF), but in the usual implementations of the STRF idea a spectrogram representation is imposed onto the stimulus, fixing the shapes of the elementary bins in the time–frequency plane and assuming that the cell responds only to stimulus power in each frequency band. Here, in contrast, Fourier transforms are in principle continuous, and the general quadratic form allows for more than just stimulus power to be relevant.

The natural stimuli we used to probe this model auditory neuron's receptive field came from recordings of zebra finch songs, modified into stimulus clips **s**. The songs were interpolated down from their original sampling rate to retain the same discrete time

steps ($dt = 1/20000$ s) that we use in Fig. 1. An example sound pressure wave of a song stimulus is plotted in Fig. 5a.

This model neuron, as illustrated in Fig. 1, emitted spikes when the power in Eq. (46) exceeded a threshold. We set the threshold of firing so that the mean spike rate was $20\,\mathrm{s}^{-1}$. We presented ~ 42 minutes of bird song stimuli to the model neuron, collecting roughly 50,000 spikes. We follow the gradient ascent procedure exactly as described in Eq. (36). Note that $\mathbf{K}$ in this model is a 300×300 matrix, and we make no further assumptions about its structure; we start with a random initial condition (Fig. 5d). The maximally informative matrix that we find is shown in Fig. 5e, and is in agreement with the matrix $\mathbf{K}$; we can see this in the frequency domain as well, shown in Fig. 5f. Quantitatively, the RMS reconstruction error for this inference is less than 5% of the maximum for any two uncorrelated random, symmetric matrices of the same size.

Model complex cells

We consider the model complex cell described earlier in Eq. (18), but allow integration over just one frame of a movie, so we don't need to describe the temporal filter. We chose parameters $k = 2\pi/3$, $\sigma_1 = 1.6$ and $\sigma_2 = 5$, with positions measured in pixels. The stimuli were 20,000 grayscale 30×30 pixel image patches extracted from a calibrated natural image database [36]. Spikes

were generated as before, with a threshold set to ensure that the fraction of bins containing spikes is 0.1. We note that, in this model, the kernel is explicitly of rank 2, and so we followed the algorithm in Eq. (41). The results are shown in Fig. 6. As expected, the best possible reconstruction is a vector pair $\mathbf{v}_1$, $\mathbf{v}_2$ that is equal to the pair F_1, F_2 up to a rotation.

Suppose that the real neuron, as in our model, is described by a kernel of rank 2, but we don't know this and hence search for a kernel of higher rank. As shown in Fig. 7c, higher rank fits do not increase the information that we capture, either for random stimuli or for natural stimuli. Interestingly, however, the "extra" components of our model are not driven to zero, but appear as (redundant) linear combinations of the two true underlying vectors, so that the algorithm still finds a genuinely two dimensional, albeit over-complete, solution. We could also initialize the algorithm with a full rank matrix, $\mathbf{Q} = \sum_{i=1}^{2} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} + \eta$, where η is Gaussian random noise. The convergence of $\mathbf{Q}$ to the "true" matrix $\mathbf{K}$ can be determined by looking at the projections of the leading eigenvectors of the matrix $\mathbf{Q}$ at the end of the maximization algorithm. In this complex cell example where spikes are generated from a rank 2 matrix, the eigenvectors corresponding to the two leading eigenvalues of the fit matrix $\mathbf{Q}$ should be identical to F_1, F_2. The remaining eigenvalues should be driven to be 0, and this is indeed what we see in Fig. 7d.
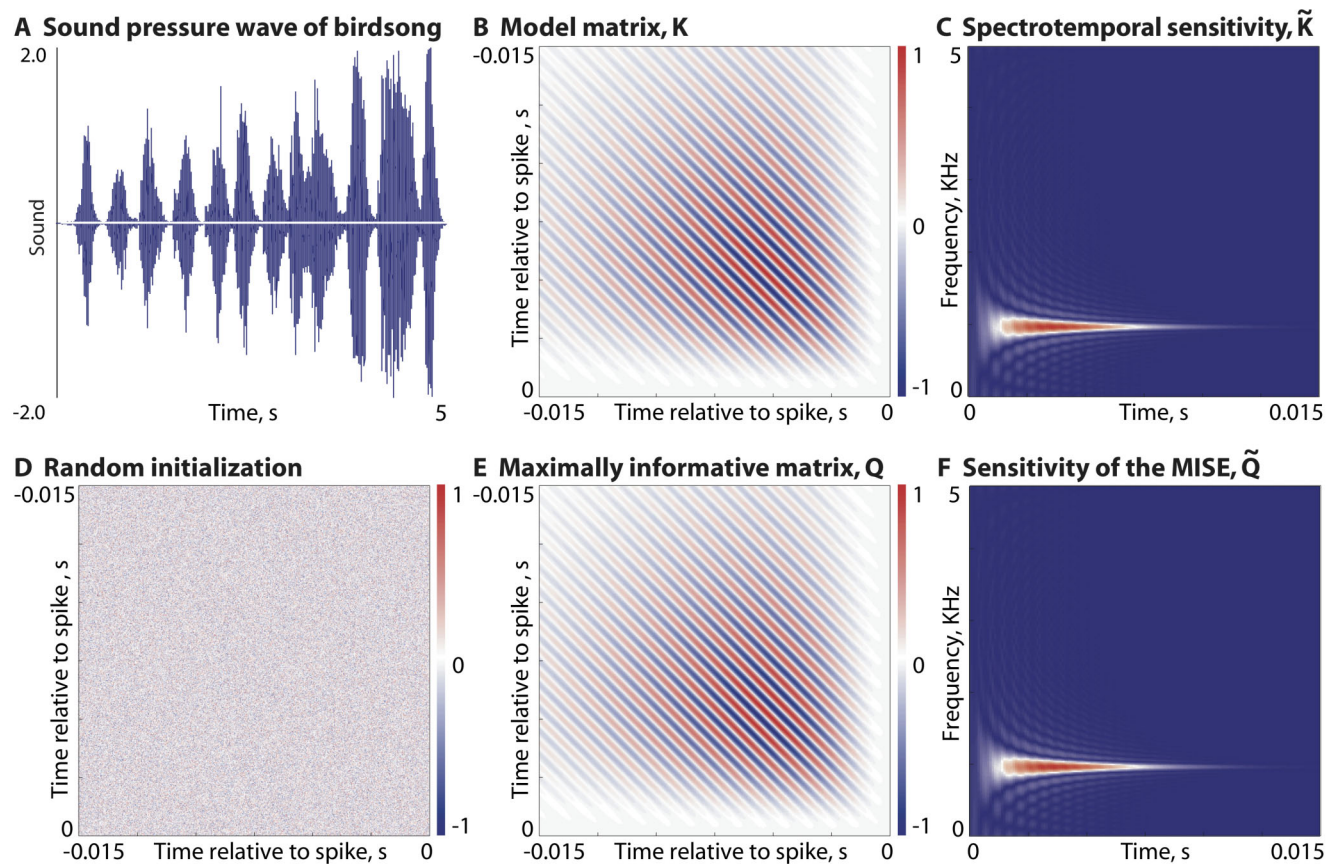

Figure 5. Analyzing the responses of the model auditory neuron to a bird song. (**a**) The sound pressure wave of a zebra finch song used as stimulus to the model neuron is shown here. (**b**) The equivalent matrix $\mathbf{K}$, constructed from the two filters as described in Eq. (8) is 300×300 in size but has a relatively simple structure. (**c**) Taking a Fourier transform over t_2 of $\mathbf{K}$ yields a spectrotemporal sensitivity matrix, $\tilde{\mathbf{K}}$ with a peak at 1kHz. (**d**) The initial guess for $\mathbf{Q}$ is the random symmetric matrix plotted here. (**e**) The optimal matrix $\mathbf{Q}$ that maximizes the mutual information between the spiking response of the model neuron and the projection $x = \mathbf{s}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{s}$ matches $\mathbf{K}$ well at the end of 100 learning steps. (**f**) The spectrotemporal sensitivity $\tilde{\mathbf{Q}}$, corresponding to the maximally informative stimulus energy has the same response preferences as $\tilde{\mathbf{K}}$ at 1kHz.
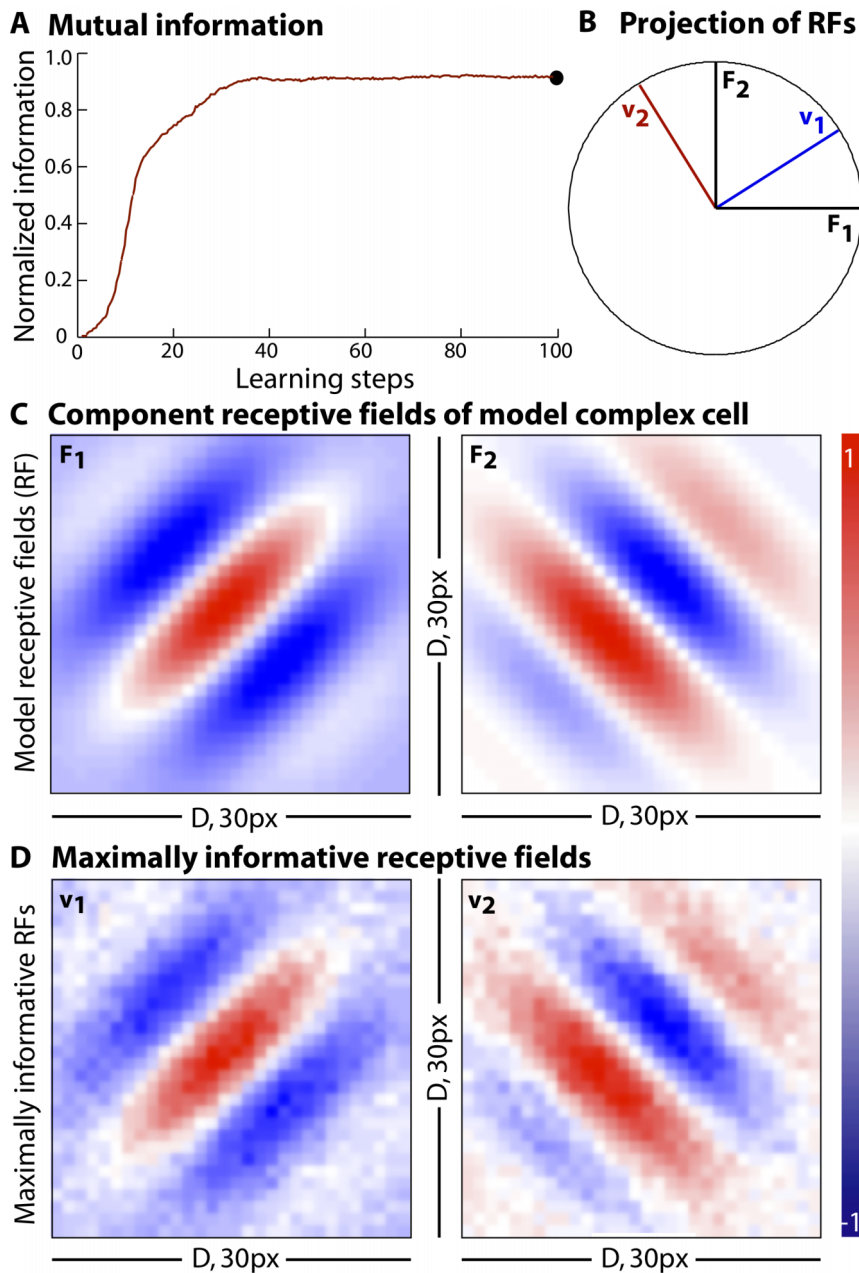doi:10.1371/journal.pone.0071959.g005

A Mutual information

B Projection of RFs

C Component receptive fields of model complex cell

D Maximally informative receptive fields

Figure 6. Receptive fields of a model complex cell and the reconstructed maximally informative pair. (a) Information as a function of the number of learning steps peaks and then plateaus. The black dot is the point where the reconstructed receptive fields are shown in panel (d) below. (b) Reconstructed vectors v_1, v_2 are rotated versions of the receptive fields F_1, F_2, but span the same linear subspace (all vectors are normalized to unit length). (c) The receptive field of the model complex cell is given by the two linear filters in Eq. (18): F_1 (left) and F_2 (right). (d) The reconstructed receptive fields at the 100th learning step (black dot in panel (a) above) with filters v_1 (left) and v_2 (right) rotated to best align with the $F_1 - F_2$ pair.
doi:10.1371/journal.pone.0071959.g006

Matrix basis formalism

Here we explore the matrix basis formalism, as described in Eq. (43), to infer a matrix $\mathbf{K}$ that could be of arbitrarily high rank. For $\mathbf{K}$ we used a symmetrized 250×250 pixel image of a fluid jet as shown in Fig. 8a. While this is not an example of a receptive field from biology, it illustrates the validity of our approach even when the response has an atypical and complex dependence on the stimulus. Spikes were generated by thresholding the energy $\mathbf{s^T \cdot K \cdot s}$, and the same naturalistic visual stimulus ensemble was used as before. Gaussian basis matrices shown in the inset of Fig. 8b were

used to represent the quadratic kernel, reducing the number of free parameters from $\sim 6 \times 10^4$ to $M = 225$. We start the gradient ascent with a large γ value of 1 and progressively scale it down to 0.1 near the end of the algorithm; Fig. 8b shows the information plateauing in about 40 learning steps. The maximally informative quadratic kernel $\mathbf{Q}$ reconstructed from these 225 basis coefficients is shown in Fig. 8c. Optimizing the coefficients of the 225 basis functions captures the overall structure of the kernel, and this can be improved to an almost perfect reconstruction (at a pixel–by–pixel resolution) by increasing M, as shown in Fig. 8d. The
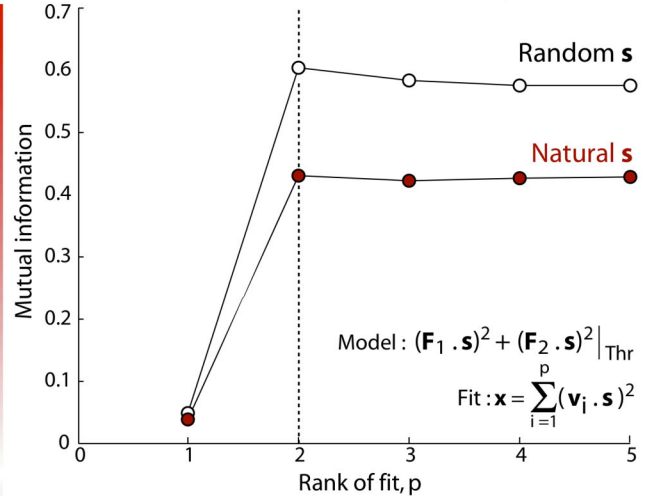
A Overfitting the model complex cell
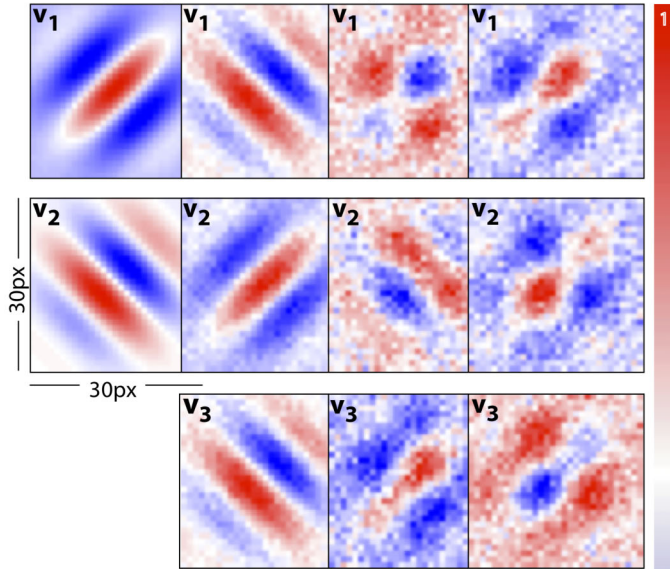
B Projection plot
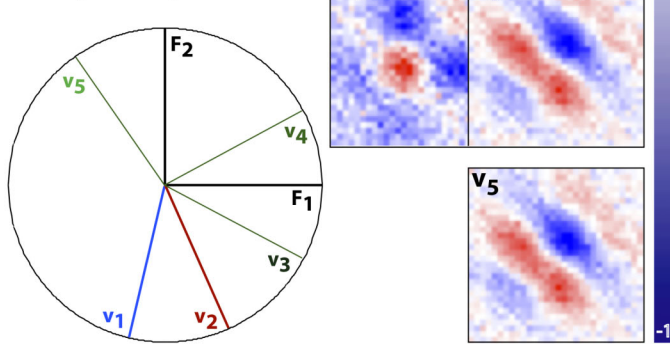
C Information saturates at rank of complex cell

Model : $(\mathbf{F}_1 . \mathbf{s})^2 + (\mathbf{F}_2 . \mathbf{s})^2 \big|_{Thr}$

Fit : $\mathbf{x} = \sum_{i=1}^{p} (\mathbf{v}_i . \mathbf{s})^2$

D Eigenvalue profile of maximally informative Q

Model : $(\mathbf{F}_1 . \mathbf{s})^2 + (\mathbf{F}_2 . \mathbf{s})^2 \big|_{Thr}$

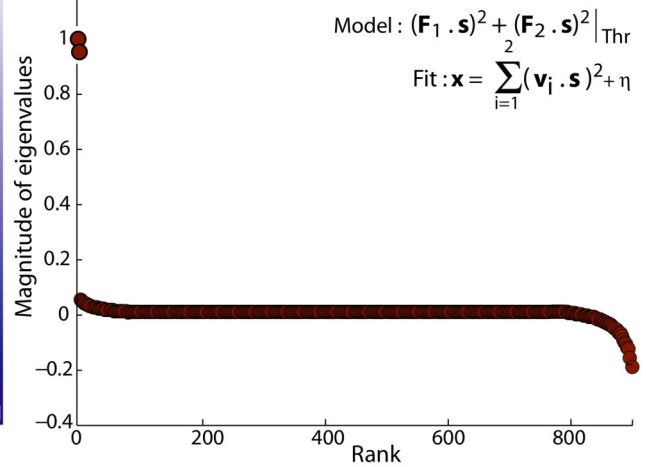Fit : $\mathbf{x} = \sum_{i=1}^{2} (\mathbf{v}_i . \mathbf{s})^2 + \eta$

Figure 7. Over-fitting the model complex cell with matrices of successively increasing rank. (a) Receptive fields reconstructed after mutual information is maximized with matrices of rank $p = 2,3,4$ and 5 (from left to right). (b) The resulting vectors, $\mathbf{v}_1$ through $\mathbf{v}_5$, at the end of the information maximization are no longer orthogonal but project fully into a unit circle in the F_1–F_2 plane. (c) Maximum mutual information as a function of the rank of fit, p, for random stimuli (open circles) or for the stimulus matrix generated using natural scenes (filled circles), peaks at the rank equal to that of the "data" (rank $p = 2$ for the model complex cell), and remains unchanged as the rank of $\mathbf{Q}$ increases. (d) Over-fitting the model matrix $\mathbf{K}$ with Gaussian noise does not add to the mutual information $I_\mathbf{Q}$ and the algorithm successfully finds a two dimensional solution. Eigenvalue profile of matrix $\mathbf{Q} = \sum_{i=1}^{2} \mathbf{v}_i \mathbf{v}_i^T + \eta$ where η is a 30×30 sized-random matrix after maximizing information with respect to the complex cell. Aside from two leading eigenvalues with magnitude 1, the rest $\rightarrow 0$.
doi:10.1371/journal.pone.0071959.g007

reduction in the number of parameters to be estimated through this simplification means that this method can be easily extended to the inference of much larger matrices, with a precision dictated by the amount of data available.

Multiple matrices

We consider a model neuron similar to the one we described in Eq. (37), with stimulus dimensionality $D = 5$; the two matrices, $\mathbf{K}_1$ and $\mathbf{K}_2$, are plotted in Fig. 9a, along with a schematic of the spikes produced when the nonlinearity $g(\cdot)$ is a threshold. Again we constructed stimuli from nearby pixels of natural images, so that the distribution is strongly correlated and non–Gaussian; the spiking threshold was set so that the fraction of bins with spikes to 0.3, and we generated $\sim 10,000$ spikes. We followed the algorithm in Eq's (38) and (39) to find the maximally informative matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$. Mutual information normalized by the maximum when

$\mathbf{K}_1 = \mathbf{Q}_1$ and $\mathbf{K}_2 = \mathbf{Q}_2$, and RMS reconstruction error in green (calculated as $\langle [(\mathbf{K}_1 + \mathbf{K}_2) - (\mathbf{Q}_1 + \mathbf{Q}_2)]^2 \rangle^{1/2}$) are plotted as a function of the learning steps in Fig. 9c. While convergence is definitely possible, our estimates of the maximally informative matrices are noisier than in the single matrix instances, even with a relatively large amount of data. We expect that realistic searches for multiple stimulus energies will require us to impose some simplifying structure on the underlying matrices.

Discussion

There is a long history in neuroscience of trying to extract receptive fields from neural responses to a variety of synthetic and natural stimuli [40]. While the field began with a focus on linear models for feature selectivity, more complex, nonlinear notions of "features" have been explored as well; some of these studies go back several decades. In the Y-type retinal ganglion cells of the cat
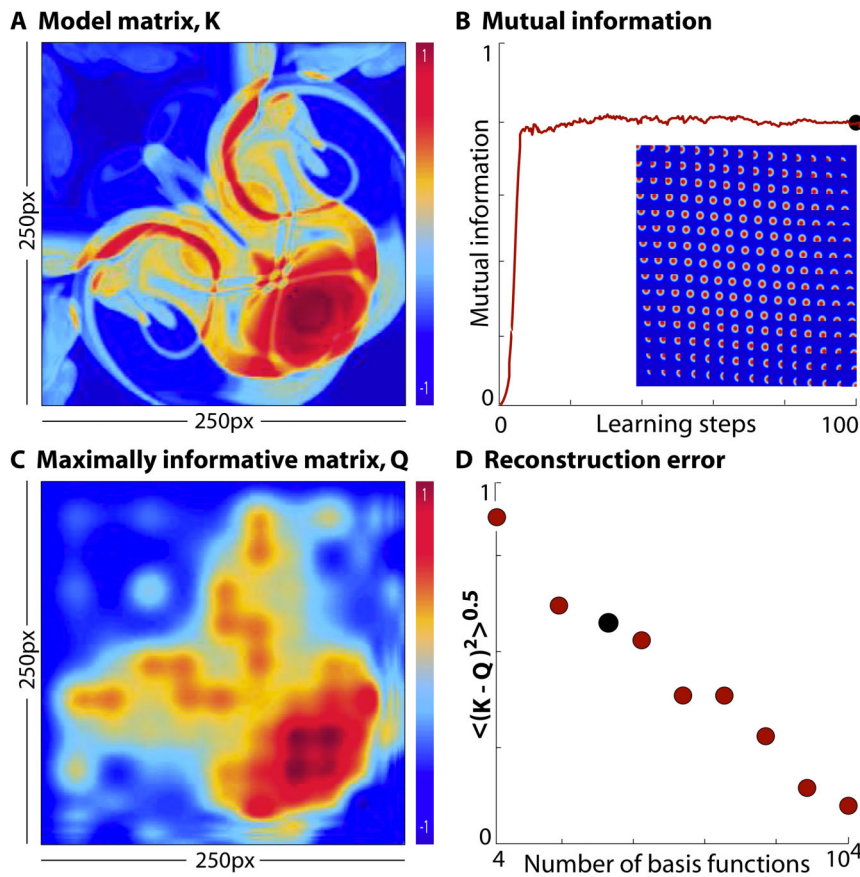
A Model matrix, K

B Mutual information

C Maximally informative matrix, Q

D Reconstruction error

Figure 8. Matrix basis formalism for constraining the number of free parameters. (**a**) A complex, structured full rank matrix **K** is generated by symmetrizing a 250×250 pixel image of a fluid jet, and used as the "true" kernel for our model neuron. (**b**) Mutual information increases with the number of learning steps, peaks at the 40th step and remains unchanged thereafter. Inset shows the collection of 225 Gaussian matrix basis functions whose peaks densely tile the space of **K**. A trial matrix is constructed as a linear sum (with coefficients $\{\alpha_\mu\}$) of the basis matrices, and information optimization is performed over $\{\alpha_\mu\}$. The black dot at the 100th learning step is the point where **Q** is extracted. (**c**) The reconstructed matrix kernel **Q** after maximizing mutual information using the 225 basis coefficients making up the kernel is shown here. (**d**) The RMS reconstruction error $\langle (\mathbf{K} - \mathbf{Q})^2 \rangle^{1/2}$ decreases as the number of basis functions M increases from 4 to 10^4. With enough data perfect reconstruction is possible as M approaches the number of independent pixels in **K**.
doi:10.1371/journal.pone.0071959.g008

responding to multi-sinusoidal signals, for instance, the responses were fit by incorporating second order kernels [41,42]. If we start with a model in which, as discussed here, the probability of spiking is determined by a quadratic "stimulus energy," then the second order kernel is precisely the first term in a series expansion of the probability as a function of the energy; our approach thus connects to a large body of work using series expansion methods (e.g., Wiener and Volterra series) to characterize the input/output relations of neurons [43]. The class of models we consider allows for a more or less arbitrary nonlinear relationship between the spike probability and the stimulus energy, thus summing up a selection of all higher order terms in a series expansion.

There have also been other more recent efforts that point to the importance of quadratic forms in stimulus space. It is shown in [37] that a logistic dependence of the spike probability on a quadratic form is the maximum entropy, and hence least structured, model consistent with a measurement of the spike–triggered covariance matrix, and that this identification is correct without any assumptions about the distributions of inputs. In addition [38], compares information theoretic extensions of spike-triggered covariance to non-Gaussian stimuli and concludes that

when the number of free parameters is large, the non-convex nature of multiple MID optimizations could raise issues. In a different direction, the work in [39] considers models in which the spike probability depends exponentially on a quadratic form, and spiking is explicitly a Poisson process. They show that this model, and some generalizations, lends itself to a Bayesian formulation, in which various simplifying structures (see Section IV) can be imposed through prior distributions on **Q**. In some limits, the different recent approaches are equivalent to one another, and to the search for maximally informative stimulus energies that we propose here (see [45]).

The maximally informative stimulus energies we have developed can be generalized to include the inference of an arbitrary two-dimensional, nonlinear quantity of the form $I(\{\mathbf{b}{\cdot}\mathbf{s}, \mathbf{Q}{\cdot}\mathbf{s}\}; \text{spike})$ without an explicit assumption about how the linear and the quadratic part combine. This idea is explored further in [45]. In a similar spirit, searching for multiple stimulus energies (shown in Fig. 9) would allow us to discover sensitivity to different sorts of nonlinear combinations, ranging from divisive nonlinearities and normalization, to logical AND operations among multiple features, without having to assume a specific form for these interactions in advance.
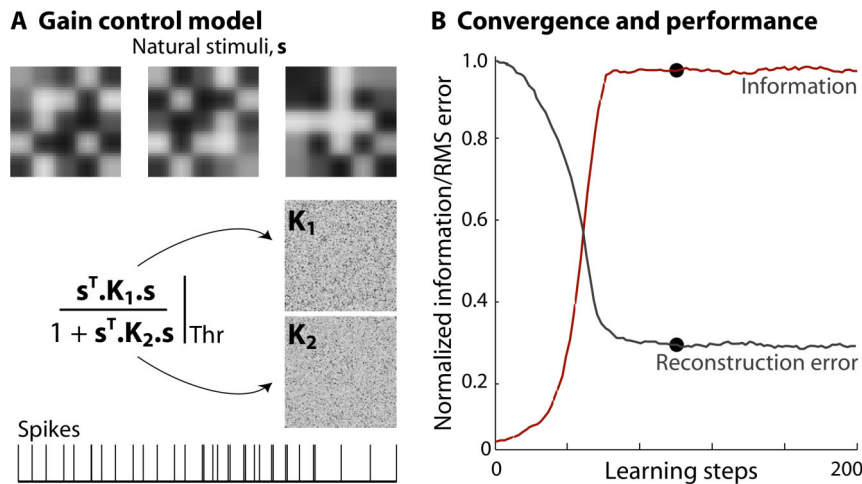
A Gain control model

Natural stimuli, **s**

B Convergence and performance

Figure 9. Inferring 2 maximally informative energies simultaneously. (**a**) Schematic of a model neuron with divisive gain control as described in the text. Natural images were reshaped and used as stimulus clips, **s**. The "true" matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ were generated as random, symmetric 25×25 matrices. (**b**) Mutual information normalized by the maximum when $\mathbf{K}_1 = \mathbf{Q}_1$ and $\mathbf{K}_2 = \mathbf{Q}_2$, peaks at the 80th learning step and remains unchanged after. RMS reconstruction error (green), computed as $\langle [(\mathbf{K}_1 + \mathbf{K}_2) - (\mathbf{Q}_1 + \mathbf{Q}_2)]^2 \rangle^{1/2}$ decays to a steady low at the same step. Using more spikes will decrease this further.
doi:10.1371/journal.pone.0071959.g009

An important aspect of methods based on information theoretic measures is that the quantities we compute have a meaning beyond more detailed models, and our results can be calibrated on an absolute scale. As an example, almost no neuron in the brain behaves exactly as a modulated Poisson process, in which the arrival times of successive spikes are independent given the stimulus. Nonetheless, there is a well defined question as to which stimulus features are most informative about the arrival times of individual spikes. Single spikes may acquire sensitivity to multiple stimulus features through interactions with other spikes, but finding these features, or the most informative single feature, does not require us to make assumptions about the spike statistics. Similarly, in experiments where stimuli are repeated we can estimate the information carried by the arrival times of single spikes or more complex events, without assumptions about the interactions among these events [32,44], and this provides an absolute calibration for the analysis of feature selectivity: we can ask not just for the features that capture the maximum information, but we can compare this feature-based information to the total, and ask if a description in terms of a small number of features really works.

In conclusion, while the notion that neurons respond to multiple projections of the stimulus onto orthogonal filters is powerful, it has been difficult to develop a systematic framework to infer a neuron's response properties when there are more than two filters. To get around this limitation, we propose an alternative model in which the neural response is characterized by features that are quadratic functions of the stimulus. In other words, instead of

being described by multiple linear filters, the selectivity of the neuron is described by a single quadratic kernel. The choice of a quadratic form is motived by the fact that many neural phenomena previously studied in isolation can be viewed as instances of quadratic dependences on the stimulus. We presented a method for inferring maximally informative stimulus energies based on information maximization. We make no assumptions about how the quadratic projections onto the resulting matrices map onto patterns of spiking and silence in the neuron. This approach yields asymptotically unbiased estimates for receptive fields for arbitrary ensembles of stimuli, but requires optimization in a possibly rugged information landscape. While clearly only one step toward a more complete collection of such tools, the methods we have presented should help elucidate how sensitivity to high–order statistical features of natural inputs arises in a wide range of systems.

Acknowledgments

Author Contributions

Conceived and designed the experiments: KR WB. Performed the experiments: KR WB. Analyzed the data: KR WB. Contributed reagents/materials/analysis tools: KR WB. Wrote the paper: KR WB.

References

1. Barlow HB (1972) Single units and sensation: a neuron doctrine for perception. Perception 1: 371–394.
2. Barlow HB (1995) The neuron in perception. The Cognitive Neurosciences, Gazzaniga MS editor MIT Press Cambridge pp. 415–4343.
3. Barlow HB (1953) Summation and inhibition in the frog's retina. J Physiol Lond 119: 69–88.
4. Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frog's brain. Proc IRE 47: 1940–1951.
5. Gross CG (2002) Genealogy of the "grandmother cell." Neuroscientist 8: 512–518.
6. Sharpee TO, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput 16: 223–50.
7. de Boer E, Kuyper P (1968) Triggered correlation. IEEE Trans Biomed Eng 15: 169–179.
8. Javel E, McGee JA, Horst W, Farley GR (1988) Temporal mechanisms in auditory stimulus coding. Auditory Function: Neurological Bases of Hearing, Wiley, New York, Edelman GM, Gall WE, Cowan WM, editors pp. 515–558.
9. Hudspeth AJ, Corey DP (1977) Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli. Proc Nat Acad Sci U S A 74: 2407–2411.

10. Palmer AR, Russell IJ (1986) Phase–locking in the cochlear nerve of the guinea pig and its relation to the receptor potential of inner hair cells. Hear Res 24: 1–15.
11. Johnson DH (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. J Acoust Soc Am 68: 1115–1122.
12. Aguera y Arcas B, Fairhall AL (2003) What causes a neuron to spike? Neural Comput 15: 1789–807.
13. Paninski L (2003). Convergence properties of three spike-triggered analysis techniques. Network 14: 437–464.
14. Bussgang JJ (1952) Cross-correlation function of amplitude-distorted Gaussian signals. Res Lab Elec MIT Cambridge MA Tech Rep 216.
15. Wick GC (1950) The Evaluation of the Collision Matrix. Phys Rev 80: 268–272.
16. Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2004) Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. Neurocomputing Elsevier 2004.
17. Schwartz O, Pillow JW, Rust NC, Simoncelli EP (2006) Spike-triggered Neural Characterization. J Vision 6: 484–507.
18. de Ruyter van Steveninck RR, Bialek W (1988) Real-time performance of a movement sensitive in the blowy visual system: Information transfer in short spike sequences. Proc Roy Soc Lond B 234: 379–414.
19. Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol Lond 160: 106–154.
20. Hubel DH, Wiesel TN (1977) Functional architecture of macaque monkey visual cortex. Proc Roy Soc Lond Ser B 198: 1–59.
21. Dimitrov AG, Sheiko M, Baker J, Yen SC (2009) Spatial and temporal jitter distort estimated functional properties of visual sensory neurons. J Comput Neurosci 27(3): 309–319.
22. Dimitrov AG, Gedeon T (2006) Effects of stimulus transformations on estimates of sensory neuron selectivity. J Comput. Neurosci 20(3): 265–283.
23. Gollisch T (2006) Estimating receptive fields in the presence of spike-time jitter. Network 17(2): 103–129.
24. Victor JD (2005) Analyzing receptive fields, classification images and functional images: challenges with opportunities for synergy. Nat Neurosci 8: 1651–1656.
25. Recio–Spinoso A, Temchin AN, van Dijk P, Fan YH, Ruggero MA (2005) Wiener-kernel analysis of responses to noise of Chinchilla auditory-nerve fibers. J Neurophys 93: 3615–34.
26. Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque V1 receptive fields. Neuron 46: 945–956.
27. Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. J Opt Soc Am A 2: 284–299.
28. Hassenstein S, Reichardt W (1956) Systemtheoretische analyse der zeitreihenfolgen und vorzeichenauswertung bei der bewegungsperzeption des ru sselka fers chlorophanus. Zeitschrift fur Naturforschung 11: 513–524.
29. Bialek W, de Ruyter van Steveninck RR (2005) Features and dimensions: Motion estimation in y vision. arxiv.org:q-bio/0505003.
30. Cover TM, Thomas JA (1991) Elements of Information Theory, John Wiley, New York.
31. Brenner N, de Ruyter van Steveninck RR, Bialek W (2000) Adaptive rescaling maximizes information transmission. Neuron 26: 695–702.
32. Brenner N, Strong SP, Koberle R, Bialek W, de Ruyter van Steveninck RR (2000) Synergy in a neural code. Neural Comp 12: 1531–1552.
33. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W (1998) Entropy and information in neural spike trains. Phys Rev Lett 80: 197–200.
34. Christianson GB, Sahani M, Linden JF (2007) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. J Neurosci 28: 446–455.
35. Gill P, Zhang J, Woolley SMN, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21: 5–20.
36. Tkačik G, Garrigan P, Ratliff C, Milčinski G, Klein JM, et al. (2011) Natural images from the birthplace of the human eye. PLoS ONE 6(6): e20409.
37. Fitzgerald JD, Sincich LC, Sharpee TO (2011) Minimal models of multidimensional computations. PLoS Comput Biol 7(3): e1001111.
38. Fitzgerald JD, Rowekamp RJ, Sincich LC, Sharpee TO (2011) Second order dimensionality reduction using minimum and maximum mutual information models. PLoS Comput Biol 7(10): e1002249
39. Park I, Pillow JW (2011) Bayesian spike-triggered covariance. Advances in Neural Information Processing Systems 25 eds Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, 2357–2365.
40. Pillow JW, Simoncelli EP (2006) Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis. J Vision 6(4): 414–428.
41. Victor JD (1988) The dynamics of the cat retinal Y cell subunit. J Physiol 405: 289–320.
42. Victor JD, Shapley RM (1979) The nonlinear pathway of Y ganglion cells in the cat retina. J Gen Physiol 74: 671–689.
43. Marmarelis PZ, Marmarelis VZ (1978) Analysis of physiological systems: The white-noise approach. New York: Plenum ISBN 030631066X.
44. DeWeese MR, Meister M (1999) How to measure the information gained from one symbol. Network 10: 325–340.
45. Rajan K, Marre O, Tkacik G (2013) Learning quadratic receptive fields from neural responses to natural signals. Neural Comp 25(7): 1661–92.
46. Abbott LF, Chance FS (2002) Rethinking the taxonomy of visual neurons. Nat Neurosci 5: 390–391.