Learning quadratic receptive fields from neural responses to natural signals: information theoretic and likelihood methods

Kanaka Rajan
Lewis-Sigler Institute for Integrative Genomics
Princeton University
Princeton, NJ 08544, USA krajan@princeton.edu

Olivier Marre
Department of Molecular Biology
Princeton University
Princeton, NJ 08544, USA omarre@princeton.edu

Gašper Tkačik
Institute of Science and Technology Austria
A-3400 Klosterneuburg, Austria
gtkacik@ist.ac.at

Abstract

Models of neural responses to rich stimuli often assume that neurons are only selective for a small number of linear projections of a potentially high-dimensional stimulus. Here we address the case where the response depends on the quadratic form of the input rather than on its linear projection, that is, the neuron can be sensitive to the local covariance structure of the stimulus preceding the spike. To infer this quadratic dependence in the presence of arbitrary (e.g. naturalistic) stimulus distribution, we present both an information-theory-based approach and a likelihood-based approach. The first can be viewed as an extension of maximally informative dimensions to quadratic stimulus dependence, while the second analogously extends the generalized linear model framework. We analyze the formal connection between the likelihood- and information-based approaches to show how they lead to consistent inference, and we demonstrate the practical feasibility of the two procedures on a synthetic neuron model responding to natural scenes. These new tools should be directly applicable for probing the feature selectivity in higher sensory areas.

1 Introduction

The concept of a receptive field, i.e., the region of stimulus space where changes in the stimulus modulate the spiking behavior of the neuron, is central to current understanding of how sensory neurons map stimuli onto patterns of spiking and silence. For instance, a ganglion cell in the retina may be sensitive only to specific changes in light intensity that occur within a small visual angle. One productive way of mathematically capturing this notion of locality has been to think of a receptive field as one or more linear filters that act on the stimulus; only those stimulus variations that result in the change in filter output have the ability to affect the neural response. In this view, the neurons are performing dimensionality reduction by linear projection, and the success of data analysis techniques based around this idea must depend on whether a small number of linear filters suffices

to fully account for the neuron's sensitivity. We currently lack systematic and tractable methods for inferring neural sensitivities when the initial dimensionality reduction step is not linear, but higher-order (e.g. quadratic). In this paper we present two complementary approaches that can be used to learn quadratic stimulus dependence even when neurons are responding to rich, naturalistic stimuli.

Suppose a neuron is driven by presenting stimulus clips $\mathbf{s}$ (where the N components of $\mathbf{s}$ represent successive stimulus values in time and optionally across space) drawn from some distribution $P(\mathbf{s})$. If the neuron is well described by the linear-nonlinear (LN) model, where the spiking rate r is an arbitrary positive pointwise nonlinear function f of the stimulus projected onto the filter, $r(\mathbf{s}) = f(\mathbf{k} \cdot \mathbf{s})$, and the stimulus distribution is chosen to be spherically symmetric, $P(\mathbf{s}) = P(-\mathbf{s})$, we can use the spike-triggered averaging (STA) to obtain an unbiased estimate of the single linear filter $\mathbf{k}$ [1, 2]. Spike-triggered covariance (STC) generalizes the filter inference to cases where the firing rate depends nonlinearly on $K \geq 1$ projections of the stimulus, $r(\mathbf{s}) = f(\mathbf{k}_1 \cdot \mathbf{s}, \mathbf{k}_2 \cdot \mathbf{s}, \ldots, \mathbf{k}_K \cdot \mathbf{s})$ [3]. The number of relevant linear filters, K, is equal to the number of nonzero eigenvalues of the spike-triggered covariance matrix. A successful application of STC requires that $P(\mathbf{s})$ be Gaussian, and a small K (usually ≤ 3) to ensure reasonable sampling of the filters and the nonlinearity f in a typical experiment. STC has been used successfully, for example, to understand the computations performed by motion sensitive neurons in the blowfly [4], to map out the sensitivity to full field stimuli and contrast gain control in salamander retinal ganglion cells [5, 6], and to understand adaptation in the rat barrel cortex [7].

Despite their utility and simplicity, spike-triggered methods require the use of statistically simple stimuli and in particular, exclude the use of stimuli with naturalistic statistics, e.g. with $1/f$ spectra, non-gaussian histograms and/or higher-order correlations. This is a big challenge when studying neurons beyond the sensory periphery that are responsible for extracting higher-order structure, or neurons unresponsive to white noise presentations, for example in the auditory pathway. To address this issue and recover the filter(s) in an unbiased way with an arbitrary stimulus distribution, maximally informative dimensions (MID) [8, 9] have been developed and utilized, for example to recover simple cell receptive fields. MID looks for a linear filter $\mathbf{k}$ that maximizes the information between the presence / absence of a spike and the projection x of the stimulus onto $\mathbf{k}$, $x = \mathbf{k} \cdot \mathbf{s}$. Information per spike is then given by the Kullback-Leibler divergence of $P(x|\text{spike})$, the *spike-triggered distribution* (the distribution of stimulus projections preceding the spike) and $P(x)$, the *prior distribution* (the overall distribution of projections):

$$I_{\text{spike}} = D_{KL}\left[P(x|\text{spike})||P(x)\right] = \int dx\; P(x|\text{spike}) \log_2 \frac{P(x|\text{spike})}{P(x)}. \tag{1}$$

Given the spike train and the stimulus, finding $\mathbf{k}$ becomes an information optimization problem in I_{spike} that can be solved using various annealing methods, although care must be taken due to the existence of local extrema.

Spike-triggered methods and MID do not explicitly assume a form for the nonlinearity $f(\cdot)$ in the LN model; instead, they provide unbiased estimates of the filter(s), and once the filters are known, the nonlinearity can be reconstructed using the Bayes' rule from sampled spike-triggered and prior distributions:

$$f(x) \propto P(\text{spike}|x) = \frac{P(x|\text{spike})P(\text{spike})}{P(x)}, \tag{2}$$

where $P(\text{spike})$ is directly related to the average firing rate during the experiment.

Let us now consider a situation where the neuron has a vanishing linear filter, as is the case with a complex cell or high-frequency auditory nerves [10]. In that case we should look for more than one linear filter, which for maximally informative approaches means finding a set of orthogonal vectors $\mathbf{k}_1, \ldots, \mathbf{k}_K$ such that the projections of the stimuli $x_i = \mathbf{k}_i \cdot \mathbf{s}$ jointly maximize the information with the spike in Eq (1). For a model complex cell, we would have two phase-shifted vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ that together form a quadrature pair, such that the most informative variable about the firing is the "power",

$$r(\mathbf{s}) = f\left[(\mathbf{k}_1 \cdot \mathbf{s})^2 + (\mathbf{k}_2 \cdot \mathbf{s})^2\right]. \tag{3}$$

Similarly, models of contrast gain control in the retina also include sensitivity to second-order features in the stimulus, with the spiking probability being [5]

$$r(\mathbf{s}) = \frac{f(\mathbf{k_0} \cdot \mathbf{s})}{\sum_{i=1}^{M} w_i (\mathbf{k}_i \cdot \mathbf{s})^2 + \sigma^2}, \tag{4}$$

where the quadratic terms in the denominator scale down the gain at high contrast (in this case however, the neuron does not have a vanishing linear filter $\mathbf{k}_0$).

We can describe these and other cases by a generic "quadratic" model neuron which is sensitive to a second-order form of the input (parametrized by a real symmetric matrix $\mathbf{Q}$) in addition to the linear projection (parametrized by the filter $\mathbf{k}_0$):

$$r(\mathbf{s}) = f(\mathbf{k}_0 \cdot \mathbf{s},\ \ \mathbf{s}^{\mathrm{T}}\mathbf{Q}\mathbf{s}). \tag{5}$$

For the contrast gain control model described in Eq (4) the matrix $\mathbf{Q}$ is of rank M, with eigenvalues w_i and eigenvectors $\mathbf{k}_i$. The complex cell example described in Eq (3) has $\mathbf{k}_0 = 0$ and $\mathbf{Q} = \sum_{i=1}^{2} \mathbf{k}_i\mathbf{k}_i^{\mathrm{T}}$; in other words, $\mathbf{Q}$ is a rank-2 matrix.

Graphically, while a threshold LN model with a linear filter corresponds to a classifier whose separating hyperplane is perpendicular to the filter, the proposed LN model with a threshold nonlinearity and a quadratic filter $\mathbf{Q}$ is selective for all stimuli that lie in an N-dimensional ellipsoid whose axes correspond to the eigenvectors of $\mathbf{Q}$.

Since every real symmetric matrix can be spectrally decomposed into $\mathbf{Q} = \sum_{i=1}^{N} \lambda_i \mathbf{k}_i\mathbf{k}_i^{\mathrm{T}}$, we could try recovering the quadratic dependence of $\mathbf{Q}$ in Eq (5) by multidimensional MID, hoping to infer all $\{\mathbf{k}_i\}$ as orthogonal informative dimensions. While formally true, this is infeasible in practice because the distributions in Eq (1) would be N-dimensional and impossible to sample. The same sampling problem would reappear when trying to estimate the nonlinearity in Eq (2). In contrast, the relevant distributions for a (purely) quadratic model are only one-dimensional functions of $x = \mathbf{s}^{\mathrm{T}}\mathbf{Q}\mathbf{s}$, making such inference possible even when $\mathbf{Q}$ is not of low rank. Clearly, this advantage has been gained by assuming that projections onto eigenvectors of $\mathbf{Q}$ combine as a sum of squares. This assumption is not only a mathematical convenience: well-known phenomena of phase invariance, adaptation to local contrast or sensitivity to the signal envelope are all examples of second-order stimulus sensitivity. Furthermore, various other response phenomena in the visual cortex grouped together as relating to the *nonclassical receptive field* could be manifestations of quadratic or higher-order sensitivity [13]. Examples of recent work establishing connections between higher-order structure of natural scenes and neural mechanisms beyond the sensory periphery (e.g. [14, 15]) make the development of corresponding methods for neural characterization, such as the one presented here, very timely.

2 Finding quadratic filters using information maximization

One way of reconstructing the quadratic filter $\mathbf{Q}$ from a recorded spike train is to maximize the information in Eq (1), where x is now given by $x = \mathbf{s}^{\mathrm{T}}\mathbf{Q}\mathbf{s}$. Taking a derivative of Eq (1) with respect to $\mathbf{Q}$ gives

$$\nabla_{\mathbf{Q}} I = \int dx\ P_{\mathbf{Q}}(x) \left[\left\langle \mathbf{s}\mathbf{s}^{\mathrm{T}} \middle|\, x, \mathrm{spike} \right\rangle - \left\langle \mathbf{s}\mathbf{s}^{\mathrm{T}} \middle|\, x \right\rangle \right] \frac{d}{dx} \left(\frac{P_{\mathbf{Q}}(x|\mathrm{spike})}{P_{\mathbf{Q}}(x)} \right), \tag{6}$$

where brackets indicate averaging over the spike-triggered or prior distributions respectively, and the subscript $\mathbf{Q}$ makes the dependence of the distributions explicit. Only the symmetric part of $\mathbf{Q}$ contributes to x, and the overall scale of the matrix is irrelevant to the information, making the number of parameters $N(N+1)/2 - 1$.

To learn the "maximally informative matrix" or the quadratic filter $\mathbf{Q}$, we can ascend the gradient in successive learning steps,

$$\mathbf{Q} \to \mathbf{Q} + \epsilon\, \nabla_Q I. \tag{7}$$

The probability distributions that enter the gradient term are obtained by computing x for all stimuli, choosing an appropriate binning for the variable x, and sampling binned versions of the spike-triggered and prior distributions. The $\langle \mathbf{s}\mathbf{s}^{\mathrm{T}} \rangle$ averages are computed separately for each bin; and the integral in Eqs (1,6) and the derivative in Eq (6) are approximated as a sum over bins and as a finite difference, respectively. To deal with local maxima in the objective function, we use a large starting value of ϵ and gradually decrease ϵ during learning. This basic algorithm can be extended by using kernel density estimation and stochastic gradient ascent / annealing methods, but we do not report these technical improvements here.

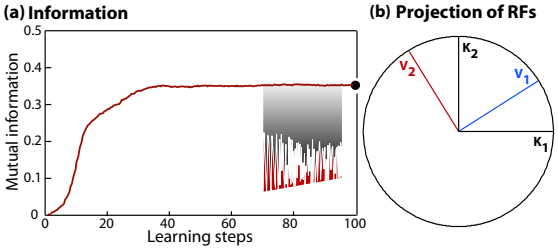

Figure 1: **(a)** I_Q as a function of the number of learning steps peaks and then plateaus. ϵ is decreased from a starting value of 0.1 to 0.01 near the end (illustrated between the 75th and 95th step in the inset). The black dot is the point where the reconstructed RFs are shown in Fig. 2(b) and in panel (b) here. **(b)** Reconstructed $\mathbf{v}_1$, $\mathbf{v}_2$ are rotated versions of $\mathbf{k}_1$, $\mathbf{k}_2$, but span the same linear subspace (all vectors are normalized to unit length).

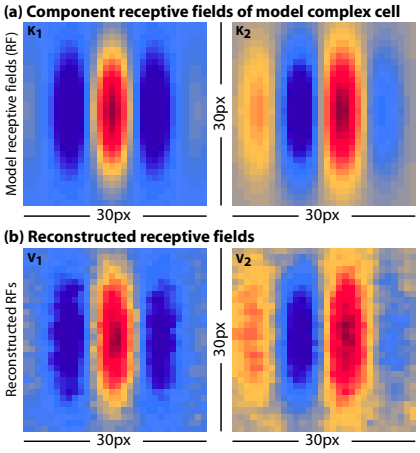

Figure 2: **(a)** The RF of the model complex cell is given by two linear filters in Eq (10): $\mathbf{k}_1$ (left) and $\mathbf{k}_2$ (right). **(b)** The reconstructed RF at the 100th learning step with filters $\mathbf{v}_1$ (left) and $\mathbf{v}_2$ (right) rotated to best align with $\mathbf{k}_1$, $\mathbf{k}_2$. A quadratic extension of GLM recovers this quadrature pair of Gabor filters equally well (not shown here).

It is possible to select an approximate linear basis in which to expand the matrix $\mathbf{Q}$, by writing

$$\mathbf{Q} = \sum_{\mu=1}^{M} \alpha_\mu \mathbf{B}^{(\mu)}. \tag{8}$$

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

$$\alpha_\mu \to \alpha_\mu + \epsilon \sum_{i,j=1}^{N} \frac{\partial I}{\partial \mathbf{Q}_{ij}} \mathbf{B}_{ij}^{(\mu)}, \tag{9}$$

where the chain rule on $\nabla_Q I$ leads to the Trace$(\nabla_Q(\alpha) \cdot \mathbf{B})$ update term.

To illustrate this approach, we consider two examples. In the first example, we consider an energy model for a complex cell, whose spatial receptive field is defined by a quadrature-pair of Gabor functions $\mathbf{k}_1$ and $\mathbf{k}_2$:

$$\exp\left(-\frac{i^2/\sigma_1^2 + j^2/\sigma_2^2}{2}\right)\cos(\kappa i) \quad \text{and} \quad \exp\left(-\frac{i^2/\sigma_1^2 + j^2/\sigma_2^2}{2}\right)\sin(\kappa i) \tag{10}$$

with $\kappa = 2\pi/3$, $\sigma_1 = 1.6$ and $\sigma_2 = 5$, and where i and j represent pixel coordinates in a 30×30 pixel frame. The stimuli were 20,000 grayscale 30×30 pixel image patches extracted from a calibrated natural image database [16]; both stimulus clips $\mathbf{s}$ as well as filters $\mathbf{k}_1$, $\mathbf{k}_2$ were represented as 900-dimensional linear vectors. Spikes were generated with some probability whenever the sum $(\mathbf{k}_1 \cdot \mathbf{s})^2 + (\mathbf{k}_2 \cdot \mathbf{s})^2$, or equivalently, the term $\mathbf{s}^{\mathrm{T}}\mathbf{K}\mathbf{s}$ (with $\mathbf{K} = \sum_{i=1}^{2} \mathbf{k}_i \mathbf{k}_i^{\mathrm{T}}$), exceeded a threshold value. For this optimization, we assumed that the sought-after information-maximizing matrix $\mathbf{Q}$ is of rank 2, and looked for its two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

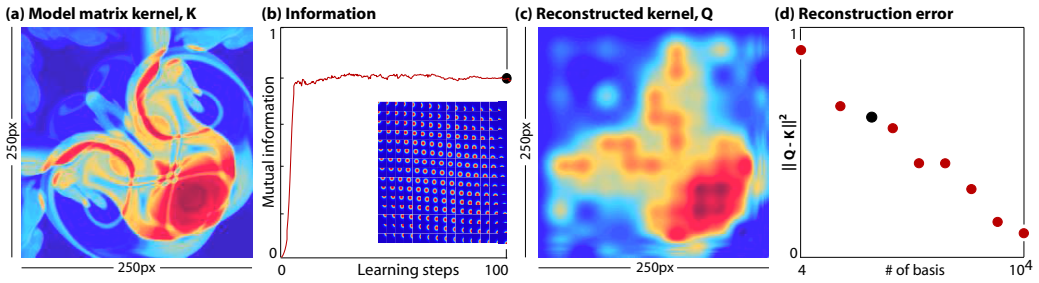

Figure 3: **(a)** A complex full-rank matrix $\mathbf{K}$ is generated by symmetrizing a 250×250 pixel image of a fluid jet. This is the true quadratic filter for our threshold LN model neuron. **(b)** Mutual information increases with learning steps, peaks at step 40 and remains unchanged thereafter. Inset shows a collection of 225 Gaussian matrix basis functions whose peaks densely tile the matrix space; 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 learning step 100 is the point where $\mathbf{Q}$ is extracted. **(c)** The reconstructed matrix kernel $\mathbf{Q}$ after maximizing mutual information. **(d)** The normalized reconstruction error $||\mathbf{Q} - \mathbf{K}||^2$, shown in red dots 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 $\mathbf{K}$. The black dot corresponds to $M = 225$ used to extract the maximally informative $\mathbf{Q}$ shown in panel (c). The spatial resolution of $\mathbf{Q}$ improves as M is increased.

The convergence of the reconstruction procedure is shown in Fig. 1(a). The energy model has an inherent phase ambiguity since the factors $\cos(\kappa i)$ and $\sin(\kappa i)$ in the Gabor functions of Eq (10) can be replaced by $\cos(\kappa i + \phi)$ and $\sin(\kappa i + \phi)$ for any ϕ without changing the responses of the model; this means that the best possible reconstruction is a vector pair $\mathbf{v}_1$, $\mathbf{v}_2$ that is equal to the pair $\mathbf{k}_1$, $\mathbf{k}_2$ up to a rotation. Figure 1(b) shows that this is indeed the case for reconstructed filters $\mathbf{v}_1$ and $\mathbf{v}_2$, which otherwise match the true filters very well, as shown in Fig. 2. The inference method is robust to different choices for the threshold and probability of generating a spike, producing consistent results as long as enough spikes are available.

In the second example we make use of the matrix basis expansion from Eq (8) to infer a quadratic kernel $\mathbf{K}$ that is of arbitrarily high rank. For $\mathbf{K}$ we used a symmetrized 250×250 pixel image of a fluid jet as shown in Fig. 3(a). 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 $\mathbf{s}^{\mathrm{T}}\mathbf{K}\mathbf{s}$, and the same naturalistic stimulus ensemble was used as before. Gaussian basis matrices shown in the inset of Fig. 3(b) were used to represent the quadratic kernel, reducing the number of optimization parameters from $\sim 6 \times 10^4$ to $M = 225$. We start the gradient ascent with a large ϵ value of 1 and progressively scale it down to 0.1 near the end of the algorithm; Fig. 3(b) shows the information plateauing in about 40 learning steps. The maximally informative quadratic filter reconstructed from these basis coefficients is shown in Fig. 3(c). Figure 3(d) demonstrates how reconstruction error systematically decreases as the number of basis functions M is increased from 4 to 10^4, improving precision.

These examples show how quadratic filters can be extracted using information maximization for both low-rank and full-rank matrices with simple (threshold) LN models. To analyze real data, we would have to simultaneously look for the most informative linear filter $\mathbf{k}_0$ as well as the matrix $\mathbf{Q}$ of Eq (5), which should remain a feasible inference task using a matrix basis.

3 Generalized quadratic models for extracting filters in the likelihood framework

A powerful technique for modeling neural spiking behavior is the Generalized Linear Model (GLM) framework [11, 12]. This has been used recently with much success to account for the stimulus dependence, spiking history dependence and interneuronal coupling in a population of 27 retinal ganglion cells in macaque retina [17]. For a single neuron, the model assumes that the instantaneous

spiking rate $\lambda(t)$ is a nonlinear function f of a sum of contributions,

$$\lambda(t) = f\left(\mathbf{k}\cdot\mathbf{s}(t) + \mathbf{g}\cdot\mathbf{y}(t_-) + \mu\right), \tag{11}$$

where $\mathbf{k}$ is a linear filter acting on the stimulus $\mathbf{s}$, $\mathbf{g}$ is the linear filter acting on the past spike train $\mathbf{y}$ of the neuron, and μ is an "offset" or intrinsic bias term towards firing or silence. When the stimulus and the spike train are discretized into timebins of duration Δ, the probability of observing (an integer number of) y_t spikes is Poisson, with the mean given by $\lambda_t\Delta$ (where the subscript indexes the time bin). Here we neglect the history dependence of the spikes (with no loss of generality) and focus instead on the stimulus dependence; since each time bin is conditionally independent given the stimulus (and past spiking), the log likelihood for any spike train $\{y_t\}$ is [18]

$$\log P(\{y_t\}|\mathbf{s}) = \sum_t y_t \log\lambda_t - \Delta\sum_t \lambda_t + c, \tag{12}$$

where c does not depend on the parameters μ and $\mathbf{k}$. This likelihood can be maximized with respect to μ and $\mathbf{k}$ (and optionally $\mathbf{g}$) given the data, providing an estimate of the filters from neural responses to complex stimuli. In contrast to maximally informative approaches, the functional form of the nonlinearity f is explicitly assumed in likelihood-based methods like GLM. For specific forms of f, such as $f(z) = \log(1+\exp(z))$ or $f(z) = \exp(z)$, the likelihood optimization problem for the filter parameters is convex and gradient ascent is guaranteed to find a unique global maximum.

The disadvantage of this approach is that if the chosen nonlinearity f is different from the true function f' used by the neuron, the filters inferred by maximizing likelihood in Eq (12) could be biased. If we relax the stringent requirement for convexity, we can choose more general nonlinear functions for the model, for example by parametrizing the nonlinearity in a point-wise fashion and inferring it jointly with the filters. For this discussion however, we assume that f has been selected from the tractable set of nonlinearities guaranteed to yield a convex likelihood function.

How can we extend the GLM to cases where the neuron's response is more complex than a single linear projection of the stimulus? One idea is to perform a simple "kernel trick": we expand the stimulus clip $\mathbf{s}$ of dimension N into a larger space first, for instance by forming $\mathbf{s}^{\mathrm{T}}\mathbf{s}$ (of dimension $N\times N$) and then operate on this object with a filter, i.e., $\sum_{i,j=1}^{N}(s_i s_j)Q_{ij}$. Such a term can be added to the argument of f in the model exemplified in Eq (11). Specifically, we propose a generalization to quadratic dependence of the following form,

$$\lambda(t) = f\left(\mathbf{k}\cdot\mathbf{s}(t) + \mathbf{s}^{\mathrm{T}}(t)\mathbf{Q}\mathbf{s}(t) + \mathbf{g}\cdot\mathbf{y}(t_-) + \mu\right). \tag{13}$$

If we want to retain convexity, we cannot expand $\mathbf{Q}$ in its eigenbasis and learn its vectors by maximizing the likelihood, because the eigenvectors would appear quadratically. However, we can expand $\mathbf{Q}$ into a weighted sum of matrix basis functions, as in Eq (8), making the argument of f a linear function of basis coefficients α_μ,

$$\lambda(t) = f\left(\mathbf{k}\cdot\mathbf{s}(t) + \sum_{\mu=1}^{M}\left[\mathbf{s}^{\mathrm{T}}(t)\mathbf{B}^{(\mu)}\mathbf{s}(t)\right]\alpha_\mu + \mathbf{g}\cdot\mathbf{y}(t_-) + \mu\right). \tag{14}$$

Existing methods for inferring GLM parameters can be used to learn both the linear filter and the quadratic filter $\mathbf{Q}$ efficiently. After extracting $\mathbf{Q}$ it is possible to check if a few principal components account for most of its structure (equivalently, if $\mathbf{Q}$ is really of low rank). This consequently provides a way of extracting multiple filters with GLM analogous to diagonalizing the spike-triggered covariance matrix. We have implemented such a quadratic extension to the GLM and verified that it accurately recovers the quadrature pair of Gabor filters in a power model for a complex cell (not shown).

4 Connection between information theoretic and likelihood-based features

We now demonstrate analytically that under rather general assumptions the linear or quadratic filters obtained by maximizing mutual information match the filters inferred by maximizing the likelihood. We adapt a reasoning which has been used in the context of inferring protein-DNA sequence-specific interactions in Ref [19], to neural responses.

In what follows, x is still the projection of the stimulus $\mathbf{s}$ onto the linear ($x_t = \mathbf{k}\cdot\mathbf{s}_t$) or quadratic ($x_t = \mathbf{s}_t^{\mathrm{T}}\mathbf{Q}\mathbf{s}_t$) filter, with time discretized into bins of duration Δ indexed by subscript t. We

collect all parameters that determine the filter into a vector θ_1. Given an x_t, y_t spikes are generated according to some conditional probability distribution $\pi(y_t|x_t)$. This probability distribution is assumed to be Poisson with mean given by $f(x_t)$ in the case of GLM but we take a different approach. We discretize x_t into $x = 1, \ldots, K$ bins and parameterize $\pi(y_t|x_t)$, a $Y_{\max} \times K$ matrix, by a set of parameters θ_2. Apart from assuming a cutoff value for the number of spikes per bin $Y_{\max}$ (which can always be chosen large enough to assign an arbitrarily low probability to observing $> Y_{\max}$ spikes in any real dataset) and a particular discretization of the projection variable x, we leave the probabilistic relationship $\pi(y|x)$ between the projection and spike count completely unconstrained. The transformation from the stimulus to the spikes is then a Markov chain, fully specified by $\theta = \{\theta_1, \theta_2\}$,

$$\mathbf{s}_t \xrightarrow[\mathbf{k} \text{ or } \mathbf{Q}]{\theta_1} x_t \xrightarrow[\pi]{\theta_2} y_t. \tag{15}$$

The likelihood of the spike train $\{y_t\}$ given the stimulus $\mathbf{s}$ is $P(\{y_t\}|\mathbf{s}) = \prod_{t=1}^{T} \pi(y_t|x_t)$, where T is the total number of time bins in the dataset. With x discretized into K bins, any dataset can be summarized by the count matrix $c_{yx} = \sum_{t=1}^{T} \delta(y, y_t)\delta(x, x_t)$, where δ is the Kronecker delta; note that $c_{yx} = T\tilde{p}(y, x)$, where $\tilde{p}$ is simply the empirical distribution in the data of observing y spikes jointly with projection x. In terms of c, the likelihood of the observed spike train is $P(\{y_t\}|\mathbf{s}) = \prod_{y=0}^{Y_{\max}} \prod_{x=1}^{K} \pi(y|x)^{c_{yx}}$. Assuming that x is adequately discretized and that π is Poisson with mean $f(x)$, we will recover the GLM likelihood of Eq (12).

Suppose that we are only interested in inferring the filter (parametrized by θ_1), but neither want to infer the filter-to-spike mapping π (parameterized by θ_2), nor make any further assumptions about its structure. In that case we can integrate the likelihood over θ_2 with some prior $P_p(\theta_2)$ such that

$$P(\{y_t\}|\mathbf{s}) = \int d\theta_2 \; P_p(\theta_2) \prod_{y,x} \pi(y|x)^{c_{yx}}. \tag{16}$$

This resulting likelihood, called the *model averaged likelihood*, is now only a function of θ_1. The prior $P_p(\theta_2)$ can take many forms, but since we discretized x, thereby making $\pi(y|x)$ into a (conditional probability) matrix, the simplest choice for the prior is the so-called *uniform prior*. In this case we take θ_2 to be directly the entries in $\pi(y|x)$ matrix and choose $P(\theta_2)$ to be uniform over all valid matrices π, such that the matrix entries are positive and the normalization constraint, $\sum_x \pi(y|x) = 1$ for every x, is enforced.

For any choice of prior we can write Eq (16) as

$$P(\{y_t\}|\mathbf{s}) = \int d\theta_2 \; P_p(\theta_2) \exp\left[T \sum_{y,x} \tilde{p}(y, x) \log \pi(y|x)\right], \tag{17}$$

which, after some algebra, can be reorganized into

$$P(\{y_t\}|\mathbf{s}) = \int d\theta_2 P_p(\theta_2) \exp\left[T\left\{\tilde{I}(y; x) - \tilde{S}(y) - \langle D_{KL}(\tilde{p}(y|x) \;||\; \pi(y|x))\rangle_{\tilde{p}(x)}\right\}\right]. \tag{18}$$

Here $\tilde{I}(y; x) = \sum_{y,x} \tilde{p}(y, x) \log \frac{\tilde{p}(y,x)}{\tilde{p}(y)\tilde{p}(x)}$ is the empirical mutual information between spike counts y and projection x, $\tilde{S}(y)$ is likewise the empirical spike count entropy, and the "correction" term in brackets measures the average (Kullback-Leibler) divergence between the empirical and model conditional distributions. Importantly, only this correction term is a function of the π and thus of θ_2, and is affected by the prior $P_p(\theta_2)$ which is being integrated over; the other terms can be pulled outside of the integral. We can therefore write the per-timebin log likelihood as

$$\mathcal{L} = \frac{1}{T} \log P(\{y_t\}|\mathbf{s}) = \tilde{I}(y; x) - \tilde{S}(y) - \Lambda, \tag{19}$$

where the correction is

$$\Lambda = -\frac{1}{T} \log \int d\theta_2 \; P_p(\theta_2) e^{-T\langle D_{KL}(\tilde{p}(y|x) \;||\; \pi(y|x))\rangle_{\tilde{p}(x)}}. \tag{20}$$

It is necessary to show that as the amount of data T grows, the correction Λ decreases for a given choice of prior distribution $P_p(\theta_2)$, and for the choice of uniform prior this is possible analytically

[19]. Intuitively, it is clear that as $T \to \infty$, the empirical distribution $\tilde{p}(y|x)$ converges to the true underlying distribution $p(y|x)$, and the integral becomes dominated by the extremal point θ_2^*, such that, in the saddle point approximation,

$$\Lambda(T \to \infty) \sim \langle D_{KL}(p(y|x) \;||\; \pi^*(y|x)) \rangle_{p(x)}. \tag{21}$$

The distribution $\pi^*(y|x)$ is the closest distribution to $p(y|x)$ in the space over which the prior $P_p(\theta_2)$ is nonzero. As long as the prior assigns a non-zero probability to any (normalized) distribution, the divergence in Λ will decrease and Λ will vanish as T grows. The case in which Λ does not decay occurs when the prior completely excludes certain distributions by assigning zero probability, while the data $p(y|x)$ precisely favors those excluded distributions.

Returning to the per-timebin log likelihood $\mathcal{L}$ in Eq (19), as we decrease the time bin Δ, we enter a regime where there is only 0 or 1 spike per bin, i.e., $y \in \{0, 1\}$. Then the empirical information per time bin $\tilde{I}(y; x)$ can be written as,

$$\tilde{I}(y; x) = \tilde{p}(y = 0) D_{KL}\left(\tilde{p}(x|y = 0)||\tilde{p}(x)\right) + \tilde{p}(y = 1) D_{KL}\left(\tilde{p}(x|y = 1)||\tilde{p}(x)\right), \tag{22}$$

that is,

$$\tilde{I}(y; x) = \tilde{p}(\text{silence})\tilde{I}_{\text{silence}} + \tilde{p}(\text{spike})\tilde{I}_{\text{spike}}. \tag{23}$$

If the information in the spike train is dominated by the information carried in spikes [20], then the likelihood from Eq (19) becomes

$$\mathcal{L} = \tilde{p}(\text{spike})\tilde{I}_{\text{spike}} + \ldots, \tag{24}$$

where $\ldots$ are terms that either do not depend of the filter parameters θ_1 (i.e. entropy of the spike counts $\tilde{S}(y)$), or vanish as the size of dataset grows (Λ). The identity in Eq (24) is the sought-after connection between the inference using information maximization and the likelihood-based approach. In the limit of small time-bins, maximizing the information per spike I_{spike} (in maximally informative approaches, as in [8] and Section 2 of this paper), on right-hand side of the identity, is the same as maximizing the *model averaged likelihood* $\mathcal{L}$ of Eq (19), on the left-hand side of the identity.

5 Discussion

While powerful conceptually, the notion that neurons respond to multiple projections of the stimulus onto orthogonal filters is typically difficult to turn into a tractable inference procedure when the number of filters is larger than two. To address this concern, we proposed an alternative neural model where the neuron can be jointly sensitive to linear and quadratic features in the stimulus. Instead of being described by multiple linear filters, the neuron's sensitivity is described by a single linear and a single quadratic filter. We motivated this choice by showing that a number of neural phenomena previously described in isolation can be seen as instances of quadratic stimulus sensitivity. We then presented two inference methods for such quadratic stimulus dependence: one based on information maximization and the other based on maximizing likelihood in a class of generalized linear models. With information maximization, no assumptions are made about how projections onto the linear and quadratic filters (probabilistically) map into spiking and silence. This approach yields unbiased filter estimates under any stimulus ensemble, but requires optimization in a possibly rugged information landscape. Alternatively, with a proper choice of nonlinearity and filter basis, likelihood inference within the GLM class can be extended to quadratic stimulus dependence while retaining the convexity of the likelihood. Lastly we demonstrated that the information-maximization approach and maximum likelihood inference will generically yield consistent filter estimates when **(i)** the timebins are made so small that each timebin contains either zero or a single spike; **(ii)** in the likelihood picture no single nonlinearity is assumed a priori, but the inference of the filters is done while simultaneously averaging over all possible choices for the nonlinearity with an uninformative (e.g. uniform) prior. Phase invariance, adaptation to local contrast or sensitivity to signal envelope are widespread features of sensory neuron responses [21, 22, 23]. The methods presented here will help us systematically elucidate sensitivity to these higher-order statistical features from responses of sensory neurons to natural stimuli.

References

[1] E de Boer & P Kuyper (1968) Triggered correlation. *IEEE Trans Biomed Eng* **15:** 169179.

[2] EP Simoncelli, L Paninski, J Pillow & O Schwartz (2004) Characterization of neural responses with stochastic stimuli. In Gazzaniga M (ed), The Cognitive Neurosciences, 3rd ed. MIT Press, Cambridge, MA.

[3] RR de Ruyter van Steveninck & W Bialek (1988) Real-time performance of a movement sensitive in the blowfly visual system: Information transfer in short spike sequences. *Proc Roy Soc Lond B* **234:** 379414.

[4] W Bialek & RR de Ruyter van Steveninck (2005) Features and dimensions: Motion estimation in fly vision. *arxiv.org:q-bio/0505003*.

[5] O Schwartz, EJ Chichilnisky & E Simoncelli (2002) Characterizing neural gain control using spike triggered covariance. *NIPS* **14:** 269–276.

[6] AL Fairhall, CA Burlingame, R Narasimhan, RA Harris, JL Puchalla & MJ Berry 2nd (2006) Selectivity for multiple stimulus features in retinal ganglion cells. *J Neurophysiol* **96:** 2724–38.

[7] M Maravall, RS Petersen, AL Fairhall, E Arabzadeh & ME Diamond (2007) Shifts in coding properties and maintainance of information transmission during adaptation in barrel cortex. *PLoS Biol* **5:** e19.

[8] TO Sharpee, NC Rust & W Bialek (2004) Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput* **16:** 223–50.

[9] TO Sharpee, H Sugihara, AV Kurgansky, SP Rebrik, MP Stryker & KD Miller (2006) Adaptive filtering enhances information transmission in visual cortex. *Nature* **439:** 936–42.

[10] A Recio-Spinoso, AN Temchin, P van Dijk, YH Fan & MA Ruggero (2005) Wiener-kernel analysis of responses to noise of Chinchilla auditory-nerve fibers. *J Neurophys* **93:** 3615–34.

[11] W Truccolo, UT Eden, MR Fellows, JP Donoghue & EN Brown (2004) A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. *J Neurophysiol* **93:** 1074–89.

[12] L Paninski (2004) Maximum likelihood estimation of cascade point-process neural encoding models. *Network Comp Neural Syst* **15:** 243–62.

[13] C Zetzsche & U Nuding (2005) Nonlinear and higher-order approaches to the encoding of natural scenes. *Network* **16:** 191–221.

[14] Y Karklin & MS Lewicki (2009) Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* **457:** 83–6.

[15] G Tkačik, JS Prentice, JD Victor & V Balasubramanian (2010) Local statistics in natural scenes predict the saliency of synthetic textures. *Proc Nat'l Acad Sci USA* **107:** 18149–54.

[16] A Anonymous. In Press.

[17] JW Pillow, J Shlens, L Paninski, A Sher, AM Litke, EJ Chichilnisky & EP Simoncelli (2008) Spatio-temporal correlations and visual signalling in a complete neural population. *Nature* **454:** 995–9.

[18] JW Pillow (2007) Likelihood-based approaches to modeling the neural code. In *Bayesian Brain: Probabilistic Approaches to Neural Coding,* eds K Doya, S Ishii, A Pouget & R Rao, pg. 53–70. MIT Press.

[19] JB Kinney, G Tkačik & CG Callan Jr (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc Nat'l Acad Sci USA* **104:** 501–506.

[20] N Brenner, RR de Ruyter van Steveninck & W Bialek (2000) Adaptive rescaling maximizes information transmission. *Neuron* **26:** 695–702.

[21] DH Hubel & TH Wiesel (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Physiol* **28:** 229–289.

[22] J Touryan, B Lau & Y Dan (2002) Isolation of relevant visual features from random stimuli for cortical complex cells. *J Neurosci* **22:** 10811–8.

[23] SA Baccus & M Meister (2004) Retina versus cortex; contrast adaptation in parallel visual pathways. *Neuron* **42:** 5–7.