Primer: The deconstruction of neuronal spike trains

JOHNATAN ALJADEFF[1,2], BENJAMIN J. LANSDELL[3],
ADRIENNE L. FAIRHALL[4] AND DAVID KLEINFELD[1,5]

[1] Department of Physics, University of California, San Diego, CA
[2] Department of Neurobiology, University of Chicago, IL
[3] Department of Applied Mathematics, University of Washington, Seattle, WA
[4] Department of Physiology and Biophysics, University of Washington, Seattle, WA
[5] Section of Neurobiology, University of California, San Diego, CA

Abstract

As information flows through the brain, neuronal firing progresses from encoding the world as sensed by the animal to driving the motor output of subsequent behavior. One of the more tractable goals of quantitative neuroscience is to develop predictive models that relate the sensory or motor streams with neuronal firing. Here we review and contrast analytical tools used to accomplish this task. We focus on classes of models in which the external variable is compared with one or more feature vectors to extract a low-dimensional representation, the history of spiking is potentially incorporated, and these factors are nonlinearly transformed to predict the occurrences of spikes. We illustrate these techniques in application to datasets of different degrees of complexity. In particular, we address the fitting of models in the presence of strong correlations in the sensory stream, as occurs in natural external stimuli and with sensation generated by self-motion.

Contents

1 Introduction

Advances in experimental design, measurement techniques, and computational analysis allow us unprecedented access to the dynamics of neural activity in brain areas that transform sensory input into behavior. One can address, for example, the representation of external stimuli by neurons in sensory pathways, the integration of information across modalities and across time, the transformations that occur during decision-making, and the representation of dynamic motor commands. While new methods are emerging with the potential to elucidate complex internal representations and transformations (Cunningham and Yu, 2014), here we will focus on established techniques within the rubric of *neuroinformatics* that summarize the relationship between sensory input or motor output and the spiking of neurons. These techniques have provided insight into neural function in a relatively large number of experimental paradigms. We discuss these methods in detail, illustrate their application to experimental data, and contrast and discuss the interpretation, reliability, and utility of the results obtained with different methods.

The methods that we will consider aim to establish input/output relationships that capture how spiking activity, generally at the single neuron level, is related to external variables: either sensory signals or motor output. These models focus on describing the statistical nature of this relationship without any direct attempt to establish mechanisms. The neural computation is parsed into several components (Fig. 1A). The first includes linear feature vectors that extract a low-dimensional description of the stimulus that drives firing. In some models, terms that capture dependence on the history of firing as well as the history of firing by other neurons in the network are incorporated. Finally, a nonlinear function describe the probability of firing as a function of these variables. The form of these components can reveal properties of the system that test theoretical concepts, such as information maximization. For example, changes in the feature under different stimulus conditions can reveal the system's ability to adapt to or cancel out correlations in the input (Hosoya et al., 2005; Sharpee et al., 2006), while changes in the nonlinearity reveal how the system can adapt its dynamic range as the intensity or variability of the stimulus changes (Fairhall et al., 2001; Wark et al., 2007; Dáaz-Quesada and Maravall, 2008).

While representing neuronal spiking through a predictive statistical model is only a limited aspect of neural computation, it is a fundamental first step in establishing function and guiding predictions as to the structure of neural circuitry. The key to any predictive model of a complex input/output relationship is *dimensionality reduction*, i.e., a simplification of the number of relevant variables that are needed to describe the stimulus. Here our primary goal is to present current methods for fitting descriptive models for single neurons and to directly compare and contrast them using different kinds of data. With the growing importance of multi-neuronal recording, it will also be necessary to seek lower-dimensional representations of network activity. While we will largely focus on methods to reduce the representation of external variables in order to predict firing, we will further point toward methods that yield a reduced description of oth external and neural variables in models of network activity.

We have chosen three datasets for analysis here as illustrative examples. The first set consists of multi-electrode array recordings from salamander retinal ganglion cells that have been presented with a long spatiotemporal white noise stimulus. This preparation has been a paradigmatic one

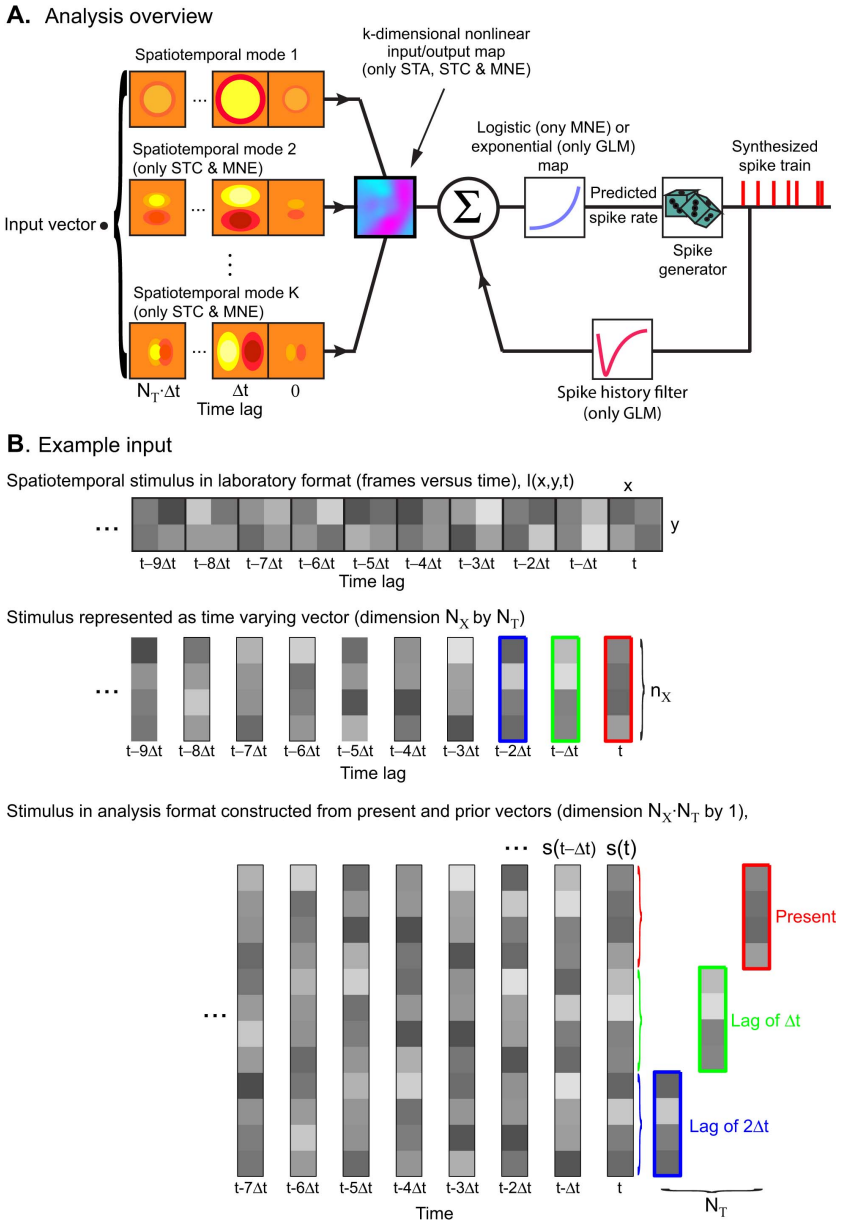

Figure 1: **Schematic for the generation of spike trains from stimuli from different classes of models. (A)** A LN model consists of a number of processing steps that transform the input stimulus into a predicted rate spike. Here we illustrate the *types* of processing stages included in the computational methods we consider. Most generally, the stimulus is projected onto one or more features and may then be passed through a nonlinear function. The result of this may be summed with a term that depends on the spike history and passed through a further nonlinear function. Finally there is a stochastic spike generation mechanism that yields a spike train. Note that none of the methods we considered have all model components; one should choose among the methods depending on the nature of the stimulus, the type of response and on the need for parsimony. For example, only the generalized linear model includes the influence of the spike history. **(B)** Here we illustrate how an example visual stimulus is reformatted as a time-dependent vector. Each stimulus frame has two spatial coordinates and a total of n_X pixels. First, the frame presented in each time point is unwrapped to give a N_X dimensional vector. Then, if the model we will construct depends on the stimulus at N_T time points, the final stimulus is a vector in which the spatial component is copied across consecutive time points to form $N = N_X \times N_T$ components.

in which many iterations of predictive modeling have been first successfully applied (Chichilnisky, 2001; Touryan et al., 2002; Rust et al., 2005; Pillow et al., 2008). The second and third set involve more challenging cases: the relationship between single unit recordings of thalamic neurons of alert, freely whisking rats and the recorded vibrissa self-generated motion (Moore et al., 2015a); and the relationship between unit recordings from motor cortex of monkeys and the recorded position and grip strength of the hand as monkeys use a joystick to manipulate a robotic arm (Engelhard et al., 2013). The data from behaving animals is more representative of typical and future experiments and allows us to discuss several important issues, including smaller data size and the highly correlated and non-repeated external variables that are generated by natural stimulus statistics and self-motion.

With each statistical model considered, it is important to ensure that one is capturing the trend of interest, and not simply structure that is specific to the training set. Thus one must always test the performance of a model with a portion of the data that was not used to build the model; typically 80% of the data is used to build the model and 20% for validation. By permuting the data among the fractions used to fit and to validate, one can build up a jack-knife estimate of the variance for the reliability of the fit.

We provide all of the code and spiking and stimulus data required to reproduce our results. Simple modification of this code will enable readers to extend the analysis methods we present to new datasets.

1.1 The linear/nonlinear modeling approach

We will focus on models broadly known as linear/nonlinear (LN) models. These have been successful in providing a phenomenological description for many neuronal input/output transformations and are constructed by correlating spikes with the external variable. Some models are nonparametric in the sense that both feature vectors and the nonlinear input/output response of the neuron are derived from the data. Other models are parametric, in that the mathematical form of the nonlinearity is fixed. While the external variable, as emphasized in the Introduction, could be either a sensory drive or a motor output, we will use the term *stimulus* for convenience from now on. Note, however, that while for sensory drive one considers only the stimulus history, in motor coding applications one would also consider motor outputs that extend into the future.

We express the neuron's response $r(t)$ at time t as a function of the recent stimulus $\mathbf{s}(t')$ (with $t' < t$) and, also, potentially its own previous spiking activity:

$$r(t) = f\left(r\left(t' < t\right), \mathbf{s}\left(t' < t\right)\right). \tag{1}$$

The stimulus vector $\mathbf{s}(t' < t)$ might, for example, represent the intensity of a full-field flicker or the pixels of a movie, the spectro-temporal power of a sound, the position of an animal's whiskers, and so on. The choice of this initial stimulus representation is an important step on its own and could in principle involve a nonlinear transformation, e.g., the phase in a whisk cycle, a case we will discuss later. The function $f(\cdot)$ generally represents a nonlinear dependence of the response on the stimulus, i.e., the response $r(t)$ is equivalent to the conditional probability of spiking and given by

$$r(t) = p(\text{spike}(t)|r\left(t' < t\right), \mathbf{s}\left(t' < t\right)). \tag{2}$$

In this form, the response $r(t)$ is generally taken to be the *expected* firing rate of a random process, which is assumed here to be Poisson. We will denote the spike counts observed on a single trial as $n(t)$.

If the neuron's response does not depend on its own history but only on the stimulus, the function $f(\cdot)$ can be expanded as a Volterra series (Marmarelis and Marmarelis, 1978; Chichilnisky, 2001), i.e.,

$$\begin{aligned} r(t) &= f\left(\mathbf{s}\left(t' < t\right)\right) \\ &= f_1(t' < t) * \mathbf{s}(t' < t) + f_2(t', t'' < t) * \mathbf{s}(t')\mathbf{s}(t'') + \ldots, \end{aligned} \tag{3}$$

where the functions $f_1(\cdot)$, $f_2(\cdot)$, etc. are kernels, analogous to the coefficients of a Taylor series, that are convolved with increasing powers of the stimulus. The Volterra series approach has been applied to a few examples in neuroscience, such as complex cells in primary vision (Szulborski and Palmer, 1990), limb position in walking in insects (Vidal-Gadea and Belanger, 2009), and single-neuron firing (Powers and Binder, 1996). To successfully implement this method, the amount of data needed to fit the kernels increases exponentially with the order of the expansion. Furthermore, capturing realistic nonlinearities including, e.g., saturation, typically requires expansions to more than first or second order.

The LN model is a powerful alternative approach that allows one to approximate the input/output relation (Eq. 1) using a plausible amount of experimental data. This differs from the Volterra series in that no attempt is made to approximate the nonlinearity in successive orders. Rather, the nonlinearity is an explicit component of the model and is arbitrary in some formulations and constrained to have a specific functional form in others. The stimulus $\mathbf{s}(t' < t)$ is in general high dimensional. It may consist of a sequence of successive instantaneous snapshots, e.g., frames of a movie, each with N_X *spatial* pixels or an auditory waveform with N_X frequency bands. With each "frame" discretized in time at sampling rate Δt (Fig. 1), there is some timescale $T = \Delta t N_T$ beyond which the influence of the stimulus on future spiking can be assumed to go to zero, defining the number of relevant frames as N_T. Then the total number of components defining the stimulus, or dimensionality of the stimulus space, denoted N, is given by

$$N = N_X \times N_T. \tag{4}$$

The key strategy of the LN approach is two-fold. First, to find a simplified description of this high- dimensional stimulus that captures its relevance to neuronal firing in terms of one to a few N-dimensional feature vectors that span the stimulus space. Second, to fit the response as a nonlinear function of those few components. As a matter of implementation, the relation $f(\cdot)$ in Eq. (1) is divided into two parts. First, the full stimulus is processed by a set of linear filters defined by feature vectors. These filters take the N-dimensional stimulus and extract from it certain components, i.e., linear combinations or *dimensions* of the stimulus, analogous to the Volterra approach, and possibly also spike history. Second, there is a nonlinear stage, which we will denote as $g(\cdot)$, that acts upon those components to predict the associated firing rate. The LN family of models makes two important assumptions about the system's input/output transformation. One is that the number of stimulus components or dimensions that are relevant to the neuron's response, K, is much less than the maximum stimulus dimensionality N. All methods then necessarily include a dimensionality reduction step, whose goal is to find these relevant K vectors, which we will call *features* and denote by $\boldsymbol{\phi}_i$, each of size N with $i = 1, \ldots, K$, in terms of which the input/output transformation can be

written:

$$r(t) = g\left(z_1, z_2, \ldots, z_K, r(t' < t)\right) \tag{5}$$

where

$$z_i = \boldsymbol{\phi}_i \cdot \mathbf{s}(t' < t) \tag{6}$$

is the projection of the stimulus on the i^{th} feature. The features $\boldsymbol{\phi}_i$ span a low-dimensional subspace within the full stimulus space, and the response of the system is approximated to depend *only* on variations of the stimulus within that subspace. The second assumption is that the nonlinear transformation, $g(\cdot)$ above, is taken to be *stationary in time*, i.e., $g(\cdot)$ has time dependence only through the stimulus and the history of the neuron's spiking response that, like the stimulus, may be high-dimensional.

The non-linearity $g(\cdot)$ can be determined nonparametrically using the probabilistic interpretation of Eq. (1) given in Eq. (2). Considering for now only dependence on the stimulus, we can use Bayes' rule, i.e.,

$$p\left(\text{spike}|\mathbf{s}(t)\right) = \frac{p\left(\mathbf{s}(t)|\text{spike}\right)\ p(\text{spike})}{p\left(\mathbf{s}(t)\right)} \tag{7}$$

to determine an input/output relation in terms of the reduced variables defined above (Eq. 6):

$$\begin{aligned} g(z_1, z_2, \ldots, z_K) &= p(\text{spike}|z_1, z_2, \ldots, z_K) \\ &= \frac{p(z_1, z_2, \ldots, z_K|\text{spike})\ p(\text{spike})}{p(z_1, z_2, \ldots, z_K)}. \end{aligned} \tag{8}$$

The probability distributions on the right hand side can be found from the data, i.e.,

- $p\left(z_1, z_2, \ldots, z_K\right)$, the *prior*, is the probability distribution of all stimuli in the experiment, projected on K stimulus features;
- $p\left(z_1, z_2, \ldots, z_K|\text{spike}\right)$, the *spike-conditional distribution*, is the probability distribution of the stimuli, projected onto the K features, conditioned on the occurrence of one or more spikes;
- $p(\text{spike})$, the mean firing rate over the entire stimulus presentation.

The *prior* distribution, $p(z_1, z_2, \ldots, z_K)$, and the *spike-conditional* distribution, $p(z_1, z_2, \ldots, z_K|\text{spike})$, are estimated by binning the K-dimensional stimulus space along each one of its directions. If we discretize each of the K stimulus components into N_B bins, the total number of bins is $(N_B)^K$, which grows exponentially with K. An accurate estimate of g requires some minimal number of samples *in each bin*, so the appropriate number of bins into which to divide the stimulus space will be determined by the duration of the experiment and the typical spike rate. With multiple dimensions, and when incorporating the history of activity, this direct method becomes prohibitive.

2 Nonparametric models

We will begin with methods that apply to situations in which systems are well driven by stimuli that approximate white noise. *White* means that the value of the stimulus at one point or time is unrelated to its value at any other point or time, that is, there are no correlations in the input. This means that all frequencies are represented in a spectral analysis of the stimulus up to the Nyquist frequency, which is simply $1/2\Delta t$; the choice of Δt may be guided by the 5 to 100 ms time-scale of neuronal responses.

An example of such a stimulus is a visual checkerboard stimulus with a total of N_X pixels whose luminance values are each chosen randomly from a Bernoulli distribution, i.e., a binary distribution with two choices, relative to an average intensity (Fig. 1B). The input that drives the cell may be viewed as a matrix of pixels in space and time, denoted $I(x,t)$.

To define a stimulus sample at time t, we select n_T frames of the input to form a matrix, i.e.,

$$\begin{pmatrix} \uparrow \\ N_X \text{ spatial positions} \\ \downarrow \end{pmatrix} \begin{pmatrix} I(1,t-N_T) & \cdots & I(1,t-1) \\ \vdots & \ddots & \vdots \\ I(N_X,t-N_T) & \cdots & I(N_X,t-1) \end{pmatrix} \\ \longleftrightarrow \\ N_T \text{ time points} \tag{9}$$

where $(\cdots)$ labels the component. In general, we wish to consider each stimulus sample as a vector in a high-dimensional space; thus one reorganizes each stimulus sample from this matrix format to an $N = N_X \times N_T$ (Eq. 4) vector that indexes the N_T frames that go back in time by $N_T \Delta t$ (Fig. 1B):

$$\boldsymbol{s}(t) = \begin{pmatrix} I(1,t-N_T) \\ \vdots \\ I(N_X,t-N_T) \\ \vdots \\ I(1,t-1) \\ \vdots \\ I(N_X,t-1) \end{pmatrix}. \tag{10}$$

2.1 Spike Triggered Average (STA)

The goal of the dimensionality reduction step is to identify a small number of stimulus features that most strongly modulate the neuron's probability to fire. Dimensionality reduction can be understood geometrically by considering each presented stimulus $\mathbf{s}(t)$ as a point in the N-dimensional space. Each location in this space is associated with a spiking probability, or firing rate $r(\mathbf{s})$, that is given by the nonlinearity evaluated at that location. A given experiment will sample a cloud of points in this N-dimensional space, with a geometry which is set by the stimulus design (all dots in Fig. 2A). The *spike-triggering* stimuli are a smaller cloud, or subset, of these stimuli (red dots in Fig. 2A). Dimensionality reduction seeks to find the stimulus subspace which captures the interesting geometrical structure of this spike-triggering ensemble.

The simplest assumption is that a cell's response is modulated by a single linear combination of the stimulus parameters, i.e., K is one. The single most effective dimension is in general the centroid of the points in this high-dimensional stimulus space that are associated with a spike. This is the *spike-triggered average*, denoted $\boldsymbol{\phi}_{\rm sta}$, the feature obtained by averaging together the stimuli that precede spikes (De Boer and Kuyper, 1968; Podvigin et al., 1974; Eckhorn and Pöpel, 1981; Chichilnisky, 2001), i.e.,

$$\boldsymbol{\phi}_{\rm sta} = \frac{1}{n_{\rm T}} \sum_t n(t) \left(\mathbf{s}(t) - \bar{\mathbf{s}} \right), \tag{11}$$

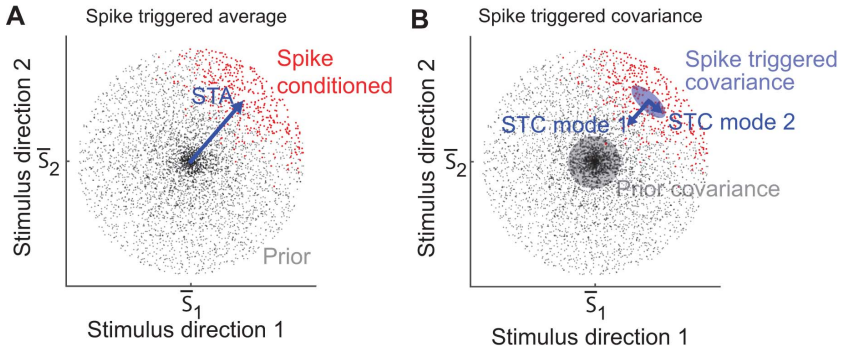

Figure 2: **Schematic of stimulus samples plotted in two arbitrary directions in stimulus space as gray and red dots; only red dots lead to a spike. (A)** The STA is a vector that points to the mean of the spike-triggered stimuli (red dots). **(B)** The covariance of the spike-triggered stimuli captures the coordinates of variation of the cloud. The covariance of the stimulus, i.e., the prior covariance $\mathbf{C}_\mathrm{p}$, forms one set of vectors and the covariance of the spike-triggered stimuli, $\mathbf{C}_\mathrm{s}$, forms a second set. The two dominant vectors comprising their difference, i.e., $\mathbf{\Delta C} = \mathbf{C}_\mathrm{s}$ - $\mathbf{C}_\mathrm{p}$, yield the dominant STC two modes.

where $\bar{\mathbf{s}}$ is the average stimulus, i.e.,

$$\bar{\mathbf{s}} = \frac{1}{n_\mathrm{T}} \sum_t \mathbf{s}(t), \tag{12}$$

$n(t)$ is the number of spikes at time t, and n_T is the total number of spikes. As for the case of the stimuli (Fig. 1B), the STA is organized as a vector of length N that indexes the N_T frames back in time from $t = \Delta t$ to $t = N_T \Delta t$ (Eq. 4), i.e.,

$$\phi_\mathrm{sta} = \begin{pmatrix} \phi_\mathrm{sta}[1] \\ \vdots \\ \phi_\mathrm{sta}[N_X] \\ \vdots \\ \phi_\mathrm{sta}[2N_X] \\ \vdots \\ \phi_\mathrm{sta}[N_T \times N_X] \end{pmatrix}. \tag{13}$$

For a Gaussian stimulus, the prior distribution of stimulus values projected onto the STA, $p\left(\boldsymbol{\phi}_\mathrm{sta} \cdot \mathbf{s}(t)\right)$, is also Gaussian. Often in experiments, however, the stimulus is binary, so that the stimulus in each pixel or time takes one of two values. If the stimulus has a large number of components, the central limit theorem ensures that these projections, as the weighted sum over many random values, will also have a Gaussian distribution whose form can either be computed analytically from the statistics used to construct the stimuli or accurately fit from data.

The nonlinearity can be estimated directly using Bayes' rule, as in Eqs. (7, 8), i.e.,

$$\begin{aligned} p(\mathrm{spike}|\mathbf{s}(t)) &= p\left(\mathrm{spike}|\boldsymbol{\phi}_\mathrm{sta} \cdot \mathbf{s}(t)\right) \\ &= \frac{p\left(\boldsymbol{\phi}_\mathrm{sta} \cdot \mathbf{s}(t)|\mathrm{spike}\right) p(\mathrm{spike})}{p\left(\boldsymbol{\phi}_\mathrm{sta} \cdot \mathbf{s}(t)\right)}. \end{aligned} \tag{14}$$

The conditional histogram defining $p(\boldsymbol{\phi}_\mathrm{sta} \cdot \mathbf{s}(t)|\mathrm{spike})$ is generally not Gaussian and is often under-sampled in the tails of the distributions. Thus when computing this ratio of histograms, it can be

helpful to fit the nonlinearity using a parametric model. If no functional form is assumed, one can simply apply a smoothing spline to the conditional distribution. Further, it may be necessary to reduce the *a priori* stimulus dimension by concatenating pixels or downsampling in time if the spiking data is too sparse leading to under-sampled probability distributions.

2.1.1 Calculating the feature for retinal ganglion cells.

We consider the case of a binary checkerboard stimulus used to drive spiking in retinal ganglion cells (Fig. 3), the third-order sensory neurons that output visual information from the retina. In this experiment, the pixel values were chosen from a binary distribution (Fig. 3A). We applied the above formalism to the stimulus set reorganized as three consecutive frames for a stimulus dimension of $N = 14^2 \times 3 = 588$. We varied the number of bins used to discretize the stimulus to get reasonably smooth features. The STA feature was computed according to Eqs. (12, 11) for each of 53 retinal ganglion units; an example is shown in Fig. 4. One should aim to choose the dimensionality of the stimulus, i.e., the product of N_T and N_X, such that the stimulus-induced variation in spiking is well captured and returns to zero at the temporal and spatial boundaries of the STA. Further, the features of the STA should be well resolved but also well averaged given the data size.

Here we tried both $N_T = 3$ time points with a larger patch size of 14×14 pixels (Fig. 4A,B), and $N_T = 6$ frames with a smaller patch of 10×10 pixels (Fig. 4C,D). The key feature is a central spot of excitation that rises and falls over three frames (Fig. 4A,C) and is thus best captured in the configuration with smaller spatial extent and 6 temporal points (Fig. 4C). Thus the STA provides a readily computed one-dimensional description of the cell; in this case the feature is a transient spot of light. We return to this point when we extend the description through a covariance analysis.

For this dataset, the large number of frames and spikes permits the prior stimulus distribution $p\left(\boldsymbol{\phi}_{\mathrm{sta}} \cdot \mathbf{s}(t)\right)$ and the conditional stimulus distribution $p(\boldsymbol{\phi}_{\mathrm{sta}} \cdot \mathbf{s}(t)|\mathrm{spike})$ to be well sampled. The prior is consistent with a Gaussian, as can be expected for a projection on any direction for a white noise stimulus (Fig. 4B,D). The application of Bayes' rule (Eq. 14) yields a monotonic nonlinearity. The nonlinearity, which is proportional to the ratio of these two distributions, was smoothed by estimating it on a downsampled set of bins (Fig. 4B,D).

2.1.2 Interpreting the feature.

The STA procedure (Eqs. 11 and 12) has a strong theoretical basis. It has been shown (Chichilnisky, 2001; Paninski, 2003) that $\boldsymbol{\phi}_{\mathrm{sta}}$ is an unbiased estimator of the feature if the spike-triggering stimuli have a non-zero mean when projected onto any vector, i.e., the cloud of spike-triggering stimuli is offset from the origin, and if the distribution of spike-triggering stimuli has finite variance. In the limit of infinite data, the STA feature is guaranteed to correctly recover the dependence of the neuron's response on this single feature. Geometrically, the vector $\boldsymbol{\phi}_{\mathrm{sta}}$ points from the origin exactly to the center of the cloud for a sufficiently large dataset. This is independent of the nonlinearity of the cell's response. However, this theorem does not guarantee that if the cell's response depends only on the projection of the stimulus onto one vector, that vector must be $\boldsymbol{\phi}_{\mathrm{sta}}$. For example, the spike-triggering cloud of stimuli points might be symmetric, such that the average lies at the origin, but

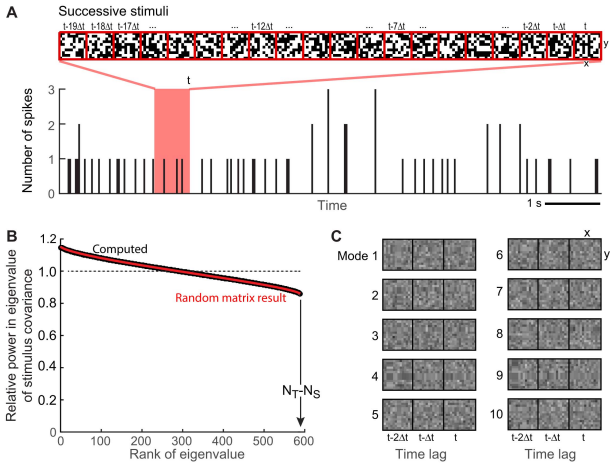

Figure 3: **Spike responses from salamander retinal ganglion cell 3 for a visual checkerboard stimulus, used to illustrate the methods with a "white noise" stimulus. (A)** Each pixel in the checkerboard was refreshed each Δt = 33.33 ms with a random value and the spikes recorded within the same interval. **(B)** We constructed the covariance matrix of the stimulus (Eq. 16) and plotted its spectrum (black).The eigenvalues are all close to the variance of a single pixel, σ^2 = 1, for the checkerboard stimulus. We compared this spectrum to the theoretical prediction given by the Marchenko-Pastur distribution with geometrical parameter $\gamma = n_{\mathrm{T}}/N$ (number of samples divided by number of dimensions). **(C)** The hallmark of white noise is that there is no structure in the stimulus, and indeed the largest eigenvectors of the stimulus covariance matrix (Eq. 16) contain no spatial or temporal structure. **Methods:** The dataset consists of 50 time series of spike arrival times simultaneously recorded from 53 retinal ganglion cells of retinae that had been isolated from larval tiger salamander (*Ambystoma tigrinum*) and laid upon a square array of planar electrodes (Segev et al., 2004). The pitch of the array was 30 μm and the spiking output of each cell, which includes spikes in both the soma and the axon, was observed on several electrodes. Using a template distributed across multiple electrodes enables one to accurately identify spikes as arising from a single retinal ganglion cell. Visual stimuli were a 14^2 = 196 square pixel array that was displayed on a cathode ray tube monitor at a frame rate of 30 Hz (Segev et al., 2006). Each pixel was randomly selected to be bright or dark relative to a mean value on each successive frame, i.e., the amplitude of each pixel was distributed bimodally, and was spectrally white up to the Nyquist frequency of 15 Hz. The image from the monitor was conjugate with the plane of the retina and the magnification was such that visual space was divided into 50 μm squares on the retina, which allowed many squares to fit inside the receptive field center of each ganglion cell with a cut-off of 200 cm^{-1} in spatial frequency. Each time series was 60 to 120 minutes long and contained between 1,000 and 10,000 spikes, but samples fewer than 2^{12} of the 2^{288} potential patterns. For some of the analyses, we extracted the 10^2 = 100 pixel central region for 2^{100} potential patterns.

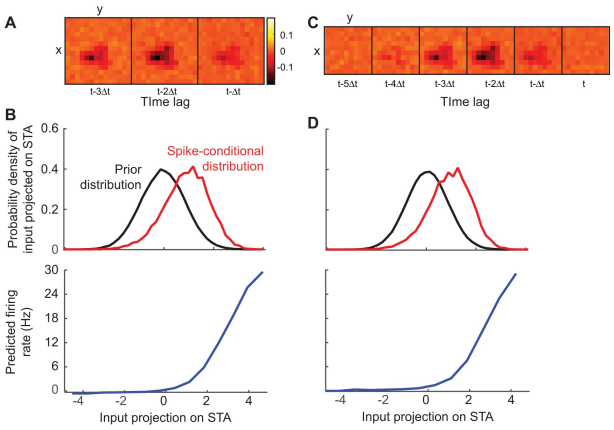

Figure 4: **The spike triggered average, ϕ_{sta}, for the responses of retinal ganglion cell 3.** We considered two stimulus representations. **(A,B)** A short sequence where we retain three stimulus frames in the past ($N_T = 3$) and the frame was $N_X = 14 \times 14 = 196$ pixels. **(C,D)** A long sequence where $N_T = 6$ but the frame was cropped such that $N_X = 10 \times 10 = 100$. In the case of the short sequence we chose the optimal lag for which the cell's response is maximally modulated by the stimulus, where for the long sequence we chose the first six frames into the past. We computed the STA for both representations, panels A and C, respectively. We then computed the prior distribution (black), the distribution of projections of *all* stimuli on the STA feature, and the spike-conditional distribution (red), the distribution of projections of stimuli associated with a spike on the STA feature. Clearly the spike-conditional distribution is shifted compared to the prior. Finally, we use Bayes' rule (Eq. 14) to obtain the input/output nonlinearity (blue), which is proportional to the probability of a spike given the value of the stimulus projected on the STA feature, i.e., $p\left(\text{spike}|\boldsymbol{\phi}_{\text{sta}} \cdot \mathbf{s}(t)\right)$

.

the shape is nonetheless very different from the cloud consisting of all stimuli, i.e., the prior distribution $p\left(\boldsymbol{\phi}_{\rm sta} \cdot \mathbf{s}(t)\right)$. The spike-triggered covariance, discussed next, is designed to make use of this additional information.

2.2 Spike-Triggered Covariance (STC)

While generally the spike-triggered average is the best solution to reduce the stimulus to a single dimension, the probability of a spike may be modulated along more than one direction in a stimulus space, as has been shown for many types of neurons across different sensory systems (Brenner et al., 2000; Fairhall et al., 2006; Slee et al., 2005; Fox et al., 2010; Maravall et al., 2007). Further, there may be a symmetry in the response, such as sensitivity to both ON or OFF visual inputs for a retinal ganglion cell, or invariance to phase in the whisk cycle for a vibrissa cortical cell, that causes the $\boldsymbol{\phi}_{\rm sta}$ to be close to zero. Thus, our next step is to generalize the notion of *feature* to a K-dimensional model of the form:

$$p(\text{spike}|\mathbf{s}(t)) = p\left(z_i, z_2, \ldots, z_K\right), \tag{15}$$

where, as a reminder, $z_i = \boldsymbol{\phi}_i \cdot \mathbf{s}(t)$ is the projection of the stimulus at time t on the ith identified feature vector $\boldsymbol{\phi}_i$. To find these K relevant dimensions, we will make use of the second order statistics of the spike-triggering stimuli.

Let us first consider the the second-order statistics of the stimulus itself, which are captured by its covariance matrix, also referred to as the prior covariance:

$$\mathbf{C}_{\rm p} = \frac{1}{n-1} \sum_t \left(\mathbf{s}(t) - \bar{\mathbf{s}}\right) \left(\mathbf{s}(t) - \bar{\mathbf{s}}\right)^\top \tag{16}$$

where $\top$ means transpose and we assume averaging over m stimulus samples indexed by t. The covariance matrix can be diagonalized into its eigenvalues, denoted λ_i, and corresponding eigenvectors, denoted $\mathbf{v}_i$, as in principal component analysis (PCA), i.e.,

$$\mathbf{C}_{\rm p} = \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \tag{17}$$

where the eigenvectors of $\mathbf{C}_{\rm p}$ are space-time patterns in the present case. The eigenvectors define a new basis set to represent directions in stimulus space that are ordered according to the variance of the stimulus in that direction, which is given by the corresponding eigenvalue.

For a Gaussian white noise stimulus, all eigenvalues of the covariance of the prior are equal and $\mathbf{C}_{\rm p}$ is a diagonal matrix with $\mathbf{C}_{\rm p} = \sigma^2 \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. The constant σ^2 is the variance of the distribution of pixel amplitudes. In practice, the use of a finite amount of data to compute the prior covariance (Eq. 17) effects the spectrum slightly but in a predictable way; the spectrum of eigenvalues of the stimulus covariance matrix is close to flat (black dots in Fig. 3B) in agreement with the analytical spectrum calculated for the same stimulus dimension and same number of samples (red dots in Fig. 3B). Although we could have computed the prior without finite size limitations, it is instructive to see the effect. The dominant eigenvectors, shown in a space-time format, appear featureless as they should (Fig. 3C).

Our goal is to find the directions in stimulus space in which the variances of the *spike-triggering* stimuli differ relative to the prior distribution of stimuli. These can be found through the covariance

difference matrix (Van Steveninck and Bialek, 1988; Agüera y Arcas and Fairhall, 2003; Bialek and van Steveninck, 2005; Aljadeff et al., 2013), denoted $\mathbf{\Delta C}$, where:

$$\mathbf{\Delta C} = \mathbf{C}_{\rm s} - \mathbf{C}_{\rm p}, \tag{18}$$

and the spike-triggered covariance matrix $\mathbf{C}_{\rm s}$, is computed relative to the spike triggered average (Eq. 11) and given by

$$\mathbf{C}_{\rm s} = \frac{1}{n_{\rm T}-1}\sum_t n(t)\left(\mathbf{s}(t)-\boldsymbol{\phi}_{\rm sta}\right)\left(\mathbf{s}(t)-\boldsymbol{\phi}_{\rm sta}\right)^\top. \tag{19}$$

The prior covariance matrix $\mathbf{C}_{\rm p}$ is given by Eq. (16) and we recall that $n_{\rm T}$ is the total number of spikes.

The matrix $\mathbf{\Delta C}$ (Eq. 18) may be expanded in terms of its eigenvalues, λ_i, and eigenvectors, $\boldsymbol{\phi}_{{\rm stc},i}$, i.e.,

$$\mathbf{\Delta C} = \sum_{i=1}^{N}\lambda_i \boldsymbol{\phi}_{{\rm stc},i}\boldsymbol{\phi}_{{\rm stc},i}^\top. \tag{20}$$

As $\mathbf{\Delta C}$ is a symmetric matrix, the eigenvalues are real numbers and the corresponding eigenvectors form a orthogonal normalized basis of the N-dimensional stimulus space, such that $\boldsymbol{\phi}_{{\rm stc},i}\cdot\boldsymbol{\phi}_{{\rm stc},j}=0$ for $i\neq j$ and $\boldsymbol{\phi}_{{\rm stc},i}\cdot\boldsymbol{\phi}_{{\rm stc},i}=1$. Positive eigenvalues correspond to directions in the stimulus space along which the variance of the spike-triggered distribution is larger than the prior distribution, and negative eigenvalues correspond to smaller variance. This analysis is illustrated in two dimension in Figure 2B. The dominant STC vectors, STC modes 1 and 2 respectively, are found by subtracting the eigenvectors of the prior covariance matrix (gray area Fig. 2B) from those of the spike-triggered covariance matrix (blue area in Fig. 2B).

Some eigenvalues will emerge from the background simply because of noise from finite sampling. To determine which K of the N eigenvectors of $\mathbf{\Delta C}$ are significant for the cell's input/output transformations, the eigenvalues λ_i are compared to a null distribution of eigenvalues obtained randomly from the same stimulus. We compute, for a large number of repetitions, a spike-triggered covariance matrix using randomly chosen spike times, $t_{\rm r}$, to select the same number of stimulus samples at random, i.e.,

$$\mathbf{C}_{\rm r} = \frac{1}{n_{\rm T}-1}\sum_{t_{\rm r}} n(t_{\rm r})\left(\mathbf{s}(t_{\rm r})-\boldsymbol{\phi}_{\rm sta}\right)\left(\mathbf{s}(t_{\rm r})-\boldsymbol{\phi}_{\rm sta}\right)^\top. \tag{21}$$

The corresponding matrix of covariance differences $\mathbf{\Delta C}_{\rm r} = \mathbf{C}_{\rm r} - \mathbf{C}_{\rm p}$ and its eigenvalues are computed for each random choice. The eigenvalues of all matrices $\mathbf{\Delta C}_{\rm r}$ form a so-called *null distribution*. Eigenvalues of $\mathbf{\Delta C}$ (Eq. 18) computed from the real spike train that lie outside the desired confidence interval of the null distribution are said to be significant. Note that one might wish to preserve any structure resulting from temporal correlations in the spike train, e.g., a tendency to spike in bursts. If such structure exists, one can compute the matrix $\mathbf{C}_{\rm r}$ (Eq. 21) using spike trains shifted by a random time lag with periodic boundary conditions such that the end of the spike train is wrapped around to the beginning.

The STC features, $\boldsymbol{\phi}_{{\rm stc},i}$, are the corresponding significant eigenvectors of the covariance difference matrix. If there is a nonzero STA, $\boldsymbol{\phi}_{\rm sta}$ will tend to be the most informative direction in stimulus space. Thus a higher dimensional model of the stimuli that lead to spiking includes the STA and the significant STC features. Examination of these features will give insight into the underlying feature selectivity of the neurons. However, for the purpose of predicting spikes, we must work in a basis where all features are orthogonal. As the STC feature vectors are not generally orthogonal to the STA,

one should project out the STA from each eigenvector used, recalling that the STC features remain orthogonal to one another. The new features are denoted as $\phi^{\perp}_{\mathrm{stc},i}$ where:

$$\phi^{\perp}_{\mathrm{stc},i} = \phi_{\mathrm{stc},i} - \frac{\phi_{\mathrm{stc},i} \cdot \phi_{\mathrm{sta}}}{||\phi_{\mathrm{sta}}||^2}\phi_{\mathrm{sta}},\ i = 1, \ldots K-1. \tag{22}$$

It is convenient to normalize these feature vectors such that the norm of each of them is equal to 1: $\phi^{\perp}_{\mathrm{stc},i} \cdot \phi^{\perp}_{\mathrm{stc},i} = 1$.

For the case of white noise, where the variance of the stimulus is equal along every direction, the eigenvalues of the prior covariance matrix, $\mathbf{C}_{\mathrm{p}}$, are essentially all equal and the STC features can be computed directly from $\mathbf{C}_{\mathrm{s}}$. However, if the variance along some directions of the stimulus is larger than others, as for the case for correlated noise, the significance threshold for each eigenvalue of $\mathbf{C}_{\mathrm{s}}$ is different. In this case, subtracting the prior covariance allows one to test whether the variance of the spike triggered distribution is different from that of the prior along each direction.

The STA and the set of orthogonalized STC vectors are then used to compute a multidimensional nonlinear function by computing the joint histogram of the K values of the spike-triggering stimuli projected onto the feature vectors and applying Bayes' rule (Eq. 7). The function $p(\mathrm{spike}|\mathbf{s}(t))$ acts as a multidimensional look-up table to determine the spike rate of the cell in terms of the overlap for the stimulus with each of the feature vectors.

2.2.1 Calculating the features for data from retina

The STC features were computed according to Eq. (18) for a set of retinal ganglion units; results for the same representative unit used for the STA features (Figs. 4 and 5A) are shown in Fig. 5. There are four STC features (Fig. 5A) that are statistically significant (Fig. 5B). The first STC feature appears as a spatial bump with a 0.93 overlap with the STA feature. Thus the dominant STC stimulus dimension is oriented in almost the same direction as the STA. The second STC feature is spatially bimodal and the third and fourth STC features have higher frequency spatial oscillations; all of these second-order features are nearly orthogonal to the STA and indicate space-time patterns beyond a "bump" that will drive the neuron to fire.

We complete the model by calculating the nonlinearity (Eq. 7). We first project out the component along the STA feature from the STC features (Eq. 22) to find the orthogonal components. The first STC feature has such a high overlap with the STA feature that the projection essentially leaves only noise. The second STC feature is essentially unchanged by the projection. As there are too few spikes to consider fitting more than a two dimensional nonlinearity, the nonlinearity is computed as a function of two variables, i.e., $p(\mathrm{spike}|\phi_{\mathrm{sta}} \cdot \mathbf{s}, \phi^{\perp}_{\mathrm{stc},2} \cdot \mathbf{s})$ (Fig. 5C). As a check on this calculation, we recover the previous result for the nonlinearity with respect to the STA alone by projecting along the STC axis (Fig. 5C). The corresponding nonlinearity for the STC mode is bowl-shaped, increasing at large negative as well as positive values of the overlap of the stimulus with $\phi^{\perp}_{\mathrm{stc},2}$. Such a nonlinearity can arise if the neuron is sensitive to a feature independent of its sign, e.g., responds equally to 'ON' or 'OFF' inputs, as in some retinal ganglion cells (Fairhall et al., 2006; Gollisch and Meister, 2008).

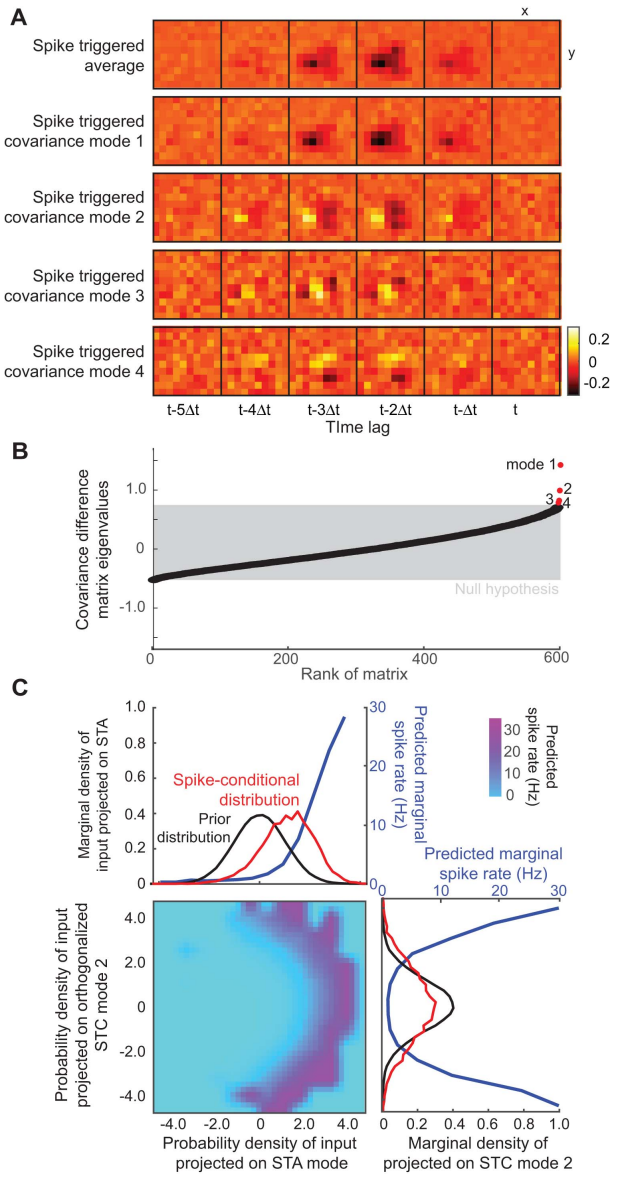

Figure 5: **The spike triggered covariance features for the response of retinal ganglion cell 13. (A)** The two significant STC feature vectors, in addition to the STA feature for comparison, using the stimulus representation with $N_T = 6$ and $N_X = 100$. The feature vector $\phi_{\mathrm{stc},1}$ has 0.930 overlap with ϕ_{sta}, while $\phi_{\mathrm{stc},2}$ through $\phi_{\mathrm{stc},4}$, have only a 0.195, 0.109, and a 0.052 overlap, respectively. **(B)** The significance of each candidate STC feature, i.e., eigenvectors of $\mathbf{\Delta C}$ (Eq. 18) were determined by comparing the corresponding eigenvalue (red and black) to the null distribution (gray shaded area). **(D)** The nonlinearity in the space spanned by the STA and the second orthogonalized STC feature, after the STA feature was projected out (Eq. 22), $\phi^{\perp}_{\mathrm{stc},2}$, completes the construction of the spiking model. The nonlinearity is a scaled version of the bivariate probability $p(\mathrm{spike}|\mathbf{s}(t))$, which is found by invoking Bayes' rule (Eq. 7). The marginals of this distribution give nonlinearities with respect to the STA (top) and second STC features (right) alone.

2.2.2 Interpreting the features

For a sufficiently large dataset, the significant STC features are guaranteed to span the entire subspace where the variance of the spike-triggered stimulus ensemble is not equal to the variance of the prior stimulus distribution (Paninski, 2003). In contrast to the corresponding result for the STA feature, for the STC feature this guarantee only holds when the stimulus distribution is Gaussian or under certain restrictions on the form of the nonlinearity (Paninski, 2003). Even when it is difficult to obtain an accurate model for the nonlinearity, the relevant STC features help to develop an understanding of the processing the system performs on its inputs. For example, in the retina, STC analysis can reveal potentially separate ON and OFF inputs to an ON/OFF cell (Fairhall et al., 2006), as noted above, and can capture spatial or temporal phase invariance, such as that exhibited by complex cells, by spanning the stimulus space with two complementary filters that can add in quadrature (Touryan et al., 2002; Fairhall et al., 2006; Rust et al., 2005; Schwartz et al., 2006; Maravall et al., 2007). Because the spectral decomposition of the symmetric matrix $\mathbf{\Delta C}$ always returns orthogonal components, the STC features cannot in general be interpreted as stimulus subunits that independently modulate the cell's response (McFarland et al., 2013). Instead, the features span a basis that includes relevant stimulus components, which may be found by rotation or other methods (Hong et al., 2008; Kaardal et al., 2013; Ramirez et al., 2014), strengthening the potential link between the functional model and underlying properties of the neural circuit. In the case of single neuron dynamics, the sampled STC feature vectors can in some cases be shown to correspond to a rotation of eigenvectors derived from subthreshold neuronal dynamics (Hong et al., 2008).

2.2.3 Relation to Principal Component Analysis (PCA)

While in PCA one usually selects eigenvectors that point in the directions of maximum variance, stimulus dimensions that are relevant to triggering a spike have a variance which may be either decreased or increased relative to the background. Consider for example a *filter-and-fire* type neuron (Agüera y Arcas and Fairhall, 2003; Paninski, 2006), where the neuron extracts a single component of the stimulus and fires when the projection of the stimulus onto that feature is larger than some threshold. The variance of the spike-triggered stimuli in the direction of that filter will therefore be reduced relative to the background. On the other hand, for a neuron such as an ON/OFF neuron in the retina (Fairhall et al., 2006), the neuron is driven to fire by an upward or by a downward fluctuation in light level. While the STA feature may then be close to zero, the eigenvalue for the eigenmode describing that feature will be positive, as spike-triggering stimuli will have both large positive and large negative values.

2.3 Natural stimuli and correlations

Our development so far has focused on methods that work well for white noise inputs, yet neurons in intermediate and late stages of sensory processing, for example, areas V2 or V4 in the visual pathway, are often not responsive to such stimuli. Rather, robust responses from these cells often require drive by highly structured stimuli, such as correlated moving stimuli that are typical of the statistics of the natural sensory environment (Simoncelli and Olshausen, 2001). In this case, the methods discussed above may be inappropriate or at the very least can be expected to yield suboptimal models.

Therefore, much attention has been given to developing methods that are appropriate to analyzing neuronal responses to *natural stimuli* or stimuli with statistics that match those of the natural sensory environment (David and Gallant, 2005; Sharpee, 2013).

Another facet of coding natural scene statistics is that animals self-modulate the structure of incoming stimuli through *active sensing*. While one could, for example, sample the natural scene statistics of a forest environment by computing the spatial and temporal correlations recorded by a stationary video camera (Ruderman and Bialek, 1994; van Hateren and van der Schaaf, 1998), an animal navigating through the forest experiences very different statistics because of its body motion (Lee and Kalmus, 1980) and saccadic eye movements (Rao et al., 2002; Nandy and Tjan, 2012). It is desirable to characterize the response properties of groups of neurons to the type of inputs driving them in a scenario that is as close to real as possible, but as we will see below, analysis of responses to such stimuli presents considerable challenges.

2.3.1 Spectral whitening for correlated stimuli

Our calculation of features so far has been limited to the case of white noise stimuli with a variance that is equal, or nearly equal, in all stimulus dimensions. This led to a covariance matrix for these stimuli, $\mathbf{C}^{\rm p}$, whose eigenvalue spectrum was nearly flat (Fig. 3B). Yet stimuli in the natural sensory environment have statistics that differ markedly, with spatiotemporal correlations and non-Gaussian structure (Ruderman and Bialek, 1994; Simoncelli and Olshausen, 2001). While the complex higher-order moments of natural inputs may be relevant for neural responses and will not be captured by first- and second-order moments (see for example Pasupathy and Connor (2002)), we can still address the issue of correlation. A correlated stimulus has a prior covariance matrix $\mathbf{C}^{\rm p}$ that contains significant off-diagonal components and whose eigenvalue spectrum is far from flat.

The STA feature and the eigenvectors of $\mathbf{\Delta C}$, i.e., the STC features, will be filtered by the correlations within the stimulus (Bialek and van Steveninck, 2005). There are two ways to correct for this. First, the features may be calculated as above, and the effect of correlations removed by dividing by the prior covariance matrix (Eq. 16). This process is referred to as decorrelation or whitening, and we denote the whitened features as $\hat{\boldsymbol{\phi}}_{\rm sta}$ and $\hat{\boldsymbol{\phi}}_{{\rm stc},i}$, where:

$$\hat{\boldsymbol{\phi}}_{\rm sta} = \mathbf{C}_{\rm p}^{-1}\boldsymbol{\phi}_{\rm sta} \tag{23}$$

$$\hat{\boldsymbol{\phi}}_{{\rm stc},i} = \mathbf{C}_{\rm p}^{-1}\boldsymbol{\phi}_{{\rm stc},i}, \qquad i = 1, \ldots, K-1, \tag{24}$$

where $\boldsymbol{\phi}_{\rm sta}$ and $\boldsymbol{\phi}_{{\rm stc},i}$ are the estimates defined by Eqs. (11, 20), respectively. The matrix $\mathbf{C}_{\rm p}^{-1}$ has the same eigenvectors as $\mathbf{C}_{\rm p}$ (Eq. 17) but the eigenvalues are inverted, i.e.,

$$\mathbf{C}_{\rm p}^{-1} = \sum_{i=1}^{K} \lambda_i^{-1}\mathbf{v}_i\mathbf{v}_i^{\top}. \tag{25}$$

Recall that $\mathbf{C}_{\rm p}$ and thus $\mathbf{C}_{\rm p}^{-1}$ are close to the identity matrix for white noise.

Equivalently, one can also first decorrelate or pre-whiten the stimulus itself by dividing by the prior covariance matrix (Theunissen et al., 2001; Schwartz et al., 2006):

$$\hat{\mathbf{s}}(t) = \mathbf{C}_{\rm p}^{-\frac{1}{2}}\mathbf{s}(t), \tag{26}$$

and then proceed with the STA and STC analysis as defined by Eqs. 11 and 20, but with $\mathbf{s}(t)$ replacing $\hat{\mathbf{s}}(t)$. Similar to the inverse of the covariance of the prior (Eq. 25), the matrix $\mathbf{C}_{\mathrm{p}}^{-\frac{1}{2}}$ is defined by

$$\mathbf{C}_{\mathrm{p}}^{-\frac{1}{2}} = \sum_{i=1}^{k} \lambda_i^{-\frac{1}{2}} \mathbf{v}_i \mathbf{v}_i^{\top}. \tag{27}$$

The decorrelation procedure is also applied when producing the null eigenvalue distribution used to determine the significance of the STC features (Eq. 21).

2.3.2 Regularization

The whitening procedure is usually numerically unstable as it tends to amplify noise. This is because decorrelation attempts to equalize the variance in all directions. Yet the eigenvector decomposition of the stimulus prior covariance matrix, $\mathbf{C}_{\mathrm{p}}$, includes directions in the stimulus space that have very low variance, i.e., small values of λ_i that are also likely to be poorly sampled. Unchecked, this leads to dividing the feature vectors or stimulus by small but noisy eigenvalues that amplify the noise in these components. This is especially a problem when there is a big difference between the large and small eigenvalues of $\mathbf{C}_{\mathrm{p}}$. To surmount this problem, we replace $\mathbf{C}_{\mathrm{p}}^{-1}$ with the *pseudoinverse* of $\mathbf{C}_{\mathrm{p}}$, which allows one to discard the small eigenvalue modes. The pseudoinverse of order L and pseudo square-root inverse of order L, with the eigenvalues λ_i arranged in decreasing order and $L < N$, are respectively defined as:

$$\mathbf{C}_{\mathrm{p};L}^{-1} = \sum_{i=1}^{L} \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^{\top} \tag{28}$$

$$\mathbf{C}_{\mathrm{p};L}^{-\frac{1}{2}} = \sum_{i=1}^{L} \lambda_i^{-\frac{1}{2}} \mathbf{v}_i \mathbf{v}_i^{\top}. \tag{29}$$

Multiplying by the pseudoinverse is equivalent to projecting out components of the stimulus along directions $\mathbf{v}_i$ that correspond to *small* λ_i before multiplying by the inverse.

The order of the pseudoinverse, L, is a regularization parameter that allows one to decide how small the variance along a certain direction of the stimulus space has to be in order to decide that one cannot accurately estimate the component of the feature in that direction. If we are able to construct a full spiking model of features and nonlinearity, we may choose the value of L as the one that yields a model that gives the best predictions for a test dataset; this is the course we followed.

2.3.3 Features from thalamic spiking during whisking in rat.

As an animal probes its environment, it presumably encodes and separates information about the motion of its sensors from the sensor's responses to external stimuli (Nelson and MacIver, 2006; Kleinfeld et al., 2006; Schroeder et al., 2010; Prescott et al., 2011). Rat whisking provides an excellent example of such active sensing in which spiking is tied to the motion of the vibrissae, i.e., long hairs that the rat sweeps through space as it interrogates the region about its head (Fig. 6A). Whisking consists of an underlying 6 to 10 Hz rhythm whose envelope and set-point are slowly modulated over time. It is often convenient to characterize vibrissa position in terms of phase in the whisk cycle as opposed to absolute angle (Curtis and Kleinfeld, 2009) (Fig. 6A), as many neurons have a preferred phase for

spiking (Fig. 6B). In our dataset, we include records of spiking from seven neurons along the primary sensory pathway in thalamus along with vibrissa position as the rats whisked in air (Moore et al., 2015a) (Fig. 6C). To ensure that the mean firing rate is stationary over the time-course of each behavioral epoch, we decomposed the whisking stimulus using a Hilbert transform (Hill et al., 2011a) and removed shifts in the set-point of the motion (compare red and blue traces of the full reconstruction in Fig. 6C), and then reconstructed the stimulus as changes in angle with respect to the set-point (Fig. 6D).

To analyze these data, we choose a 300 ms window with a 2 ms sampling period so that the stimulus $\mathbf{s}(t)$ is a $N_T = 150$ dimension vector in time. Here, because we consider only a single whisker, $N_X = 1$ and $N = N_T$. The prior covariance (Eq. 17) has eigenvalues that fall off dramatically by a few orders of magnitude (Fig. 6E), contrasting with the nearly flat spectrum of white noise (Fig. 3B). The dominant eigenvectors appear as sines and cosines at the whisking frequency (modes 1 and 2 in Fig. 6E), with higher order modes corresponding to variations in amplitude (modes 3 to 6) and higher harmonics (mode 7 and 8). The power in modes higher than about 60 is negligible. This spectral decomposition illustrates the high degree of correlation of the stimulus and the considerable variation in the sampling of each stimulus dimension, seen from the amplitude fall-off in high frequencies. Lastly, we observed that while the inter-whisk interval shows a peak at the whisking frequency, the inter-spike interval for a representative neuron appears largely exponential despite the presence of a strong rhythmic component in the stimulus (Fig. 6C).

We first consider the case of the feature vectors without whitening. We computed the STA feature (Eq. 11) (Fig. 7A) and the three significant STC features (Eq. 18) (Fig. 7A,B) for neurons in vibrissa thalamus. The STA feature appears as a decaying sine wave (Fig. 7A) and the dominant STC feature appears as a phase-shifted version of the $\boldsymbol{\phi}_{\mathrm{sta}}$ (gray, Fig. 7A). The overlap of $\boldsymbol{\phi}_{\mathrm{stc},1}$ with $\boldsymbol{\phi}_{\mathrm{sta}}$ is small, -0.06. Thus the dominant unwhitened STC feature could be safely orthogonalized relative to the unwhitened STA feature (Eq. 22) and used to construct a nonlinear input/output surface for this cell (not shown).

We repeated the above analysis with a whitened stimulus. The stimulus was decorrelated using an order L pseudoinverse (Eq. 25), where L was varied between 2 and 40. For each value of L we computed a predictive model, as described below, and chose $L = 3$ as providing the best predictability. We show the decorrelated (Eq. 23) and regularized (Eq. 28) STA feature (Fig. 7A) and the one significant STC feature (Fig. 7A,B). Here, the STA and the STC feature are very similar to those for the unwhitened case even though the analysis was restricted to a three dimensional subspace spanned by the leading eigenvalues of $\mathbf{\Delta C}$ (Egs. 18) after whitening . We then constructed the nonlinear input/output surface for the cell using these feature vectors (Fig. 7C). The nonlinearity with respect to the STA feature alone appears as a saturating curve with a shut-down for extremely high inputs.

2.4 Maximally informative dimensions

So far the model features and nonlinearity have been nonparametric, determined only by data. Another method in the same spirit is that of *maximally informative dimensions* (Sharpee et al., 2003, 2004; Rowekamp and Sharpee, 2011), an alternative means to find spike-triggering features and an arbitrary nonlinearity. Rather than using a geometrical approach, this method instead implements a search to locate a feature that maximizes the information that the spikes contain about this feature, i.e., the

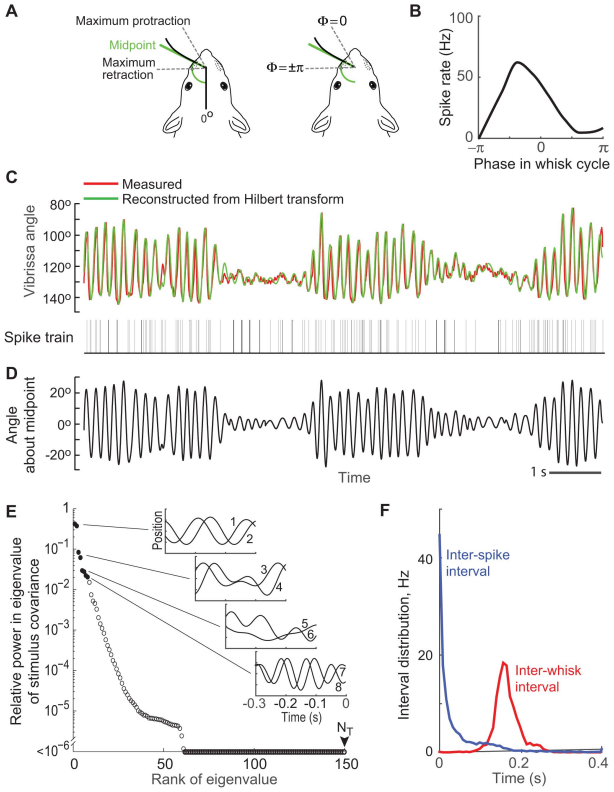

Figure 6: **Spike responses from thalamic cell 57 in response to whisking in air. (A)** The coordinate systems used to describe the whisk cycle. The left is absolute angle, θ_{whisk} and the right is phase, $\Phi(t)$, which are related by $\theta_{\text{whisk}}(t) = \theta_{\text{protract}} - \theta_{\text{amp}}\,(1 - \cos(\Phi(t)))$. **(B)** The spike rate as a function of phase in the whisk cycle; the peak defines the preferred phase Φ_o. **(C)** A typical whisk, the stimulus, and spikes in the vibrissa area of ventral posterior medial thalamus. We show raw whisking data and, as a check, the data after the components θ_{protract}, θ_{amp}, and $\Phi(t)$ were found by the Hilbert transform and the whisk reconstructed. **(D)** Reconstructed whisk, leaving out slowly varying mid-point θ_{protract} - θ_{amp}. The self-motion stimulus is taken as the vibrissa position up to 300ms in the past with Δt = 2 ms time bins, so that $N_X = 1$, $N_T = 150$, and thus $N = 150$. **(E)** The spectrum of the covariance matrix of the self-motion (Eq. 16). Note the highly structured dominant modes. **(F)** The inter-whisk and inter-spike intervals. **Methods:** The whisking dataset is used to illustrate our methods with a stimulus that contains strong temporal correlations. It consists of seven sets of spike arrival times, each recorded from a single unit in the vibrissa region of ventral posterior medial thalamus of awake, head-restrained rats (Moore et al., 2015a). The animals were motivated to whisk by the smell of their home-cage. Spiking data were obtained with quartz pipets using juxtacellular recording (Moore et al., 2015b); this method ensures that the spiking events originate from a single cell. The anterior-to-posterior angle of the vibrissae as a function of time was recorded simultaneously using high-speed videography. Each time-series contained 4 to 14 trials, each 10 s in length, with between 1,300 and 3,500 spikes per time series. The correlation time of the whisking, which serves as the stimulus for encoding by neurons in thalamus, is nominally 0.2 s (Hill et al., 2011a). We found that the cells' response was strongly modulated by the whisker dynamics only when the amplitude θ_{amp} was relatively high; therefore we constructed the models and tested their predictions only for periods when $\theta_{\text{amp}} \geq 10^\circ$.

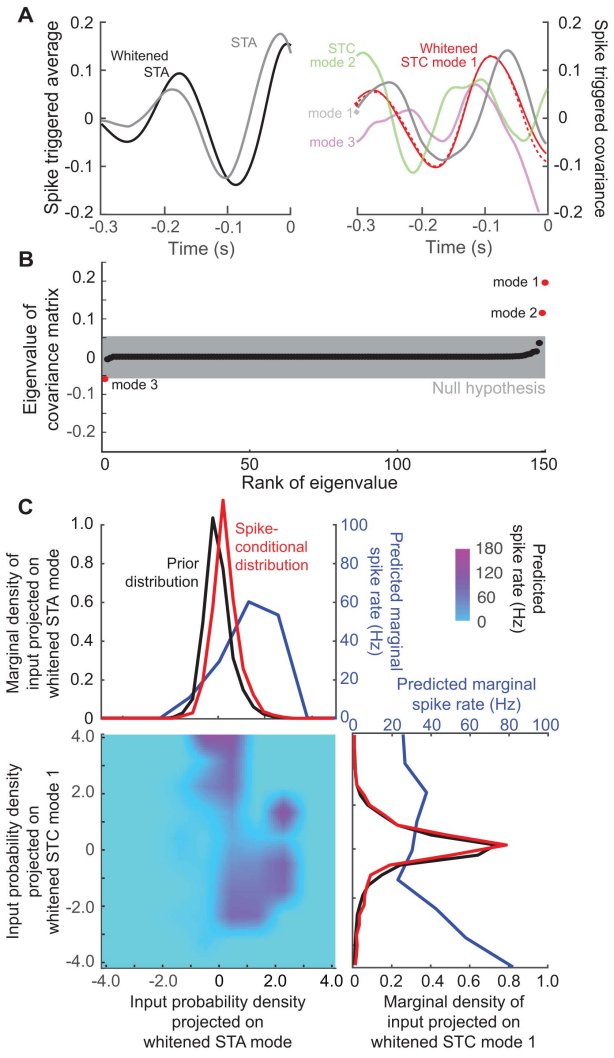

Figure 7: **The spike-triggered average and spike-triggered covariance feature vectors for the response of thalamic cell 57 in the rat vibrissa system. (A)** The STA feature and the same feature computed for the whitened stimulus, along with the leading STC features calculated with and without whitening. The dashed curves are after projecting out the STA vector from the STC modes. **(B)** Comparing the eigenvalues of $\mathbf{\Delta C}$, without whitening, to the null eigenvalue distribution computed from randomly shifted spike trains demonstrates the statistical significance of the leading STC eigenvectors; red denotes significant eigenvectors and black not significant. For the case of $\mathbf{\Delta C}$ with whitening, regularization led to only three eigenvectors of which one was significant. **(C)** A two dimensional model of the nonlinearity for $\hat{\phi}_{\mathrm{sta}}$ and the leading STC feature $\hat{\phi}_{\mathrm{stc},1}$, both computed after whitening. We further plot the two marginals.

mutual information between stimulus and spikes. To understand this approach, we return to the definition of the nonlinearity based on Bayes' rule (Eq. 7), which we will recall just for a single feature and the corresponding projection of the stimulus, i.e., $z_1 = \phi_1 \cdot \mathbf{s}$, so that:

$$r(t) \sim \frac{p(z_1|\text{spike})}{p(z_1)}. \tag{30}$$

One wishes to find a feature ϕ_1 such that this function varies strongly with z_1. If it is constant, the observation of a spike gives no information about the presence of the feature in the input, and conversely that feature is not predictive of the occurrence of a spike. The mutual information between spike and stimulus will be maximized when the two distributions, $p(z_1|\text{spike})$ and $p(z_1)$, are as different as possible. This can be measured through the Kullback-Leibler divergence. In this approach, one searches for the direction that maximizes the divergence between the distribution of all stimuli, projected onto ϕ_1, and the spike-conditional distribution of these projections. Unlike the STC procedure, this approach requires no assumptions about the structure of the stimulus space and has been applied to derive features from natural images. It can also be extended to multiple features. In general, however, this method is computationally expensive and prone to local minima, so we do not implement this analysis here; the code can be downloaded from `http://cnl-t.salk.edu/Code/`.

3 Models with constrained nonlinearities

The ability to find nonparametric stimulus features and nonlinearity can be severely constrained by data size. As we have seen, with realistic amounts of data, such models are often under-sampled, particularly if one wants to incorporate dependence on multiple features and other factors such as the history of spiking and, potentially, network effects. The methods we will discuss next instead make specific assumptions about the form of the nonlinearity that simplify the fitting problem.

In this approach, one poses a so-called *noise model* for the responses given the stimulus and the choice of model parameters and then estimates the parameters of the model that best account for the data. Once the noise model is specified the likelihood of a given set of parameters given the data can be computed. Maximization of the likelihood function then provides an estimate of the model that best accounts for the data. This maximization can be achieved reliably when the likelihood is convex. A convex function, one whose curvature does not change sign, can have no local minima or maxima, thus maximization can be performed using local gradient information and ascending the likelihood function to a unique peak. There are many convex optimization algorithms available, for instance the conjugate gradient ascent algorithm (Malouf, 2002).

An important consideration in fitting these models is that, even in cases in which the solution is unique due to convexity, the model may be accounting for variation that is specific to the data used for the fit. This is a phenomenon known as over-fitting and it manifests as a decrease in predictability of the model on novel datasets relative to the quality of the fit obtained in the training data. To ensure that the model is not simply capturing noise terms specific to the training set, a comparison between performance on test and training data is, for all approaches, a critical validation step. To minimize overfitting, one can increase the tolerance of the fitting function such that the gradient ascent stops when the model parameters have not yet reached the global minimum. Alternatively, one can partition the data into different random choices of training and test sets, known as jackknife resampling,

Table 1: Moments for MNE models

Moment	0	1	2
Element	scalar	$[i]$-th component of a vector	$[i,j]$-th component of a matrix
Symbol	$\langle r(t)\rangle$	$\langle r(t)\ \mathbf{s}[i](t)\rangle$	$\langle r(t)\ \mathbf{s}[i](t)\ \mathbf{s}[j](t)\rangle$
Data	n_{T}/N	$\frac{1}{N}\sum_t n(t)\mathbf{s}[i](t)$	$\frac{1}{N}\sum_t n(t)\mathbf{s}[i](t)\ \mathbf{s}[j](t)$
Model	p(spike)	$\sum_{t_{\mathrm{s}}} \mathbf{s}[i](t_{\mathrm{s}})p(\mathrm{spike})p(\mathrm{spike}\|\mathbf{s}(\mathbf{t}_{\mathrm{s}}))$	$\sum_{t_{\mathrm{s}}} \mathbf{s}[i](t_{\mathrm{s}})\mathbf{s}[j](t_{\mathrm{s}})p(\mathrm{spike})p(\mathrm{spike}\|\mathbf{s}(\mathbf{t}_{\mathrm{s}}))$

where n_{T} is as before the total number of spikes and t_{s} are the spike times.

and run the optimization repeatedly on these different partitions. The resulting parameters may then be averaged over the repetitions; the variability of the estimates may also be quantified.

3.1 Maximum Noise Entropy (MNE) method

A theoretically principled way to specify a noise model is by assuming a conditional probability distribution of stimuli and responses, $p(\mathrm{spike}|\mathbf{s})$, that is as agnostic as possible about the relationship between input and output, while remaining consistent with well-defined measurements on the data. This can be done by assuming that the variability in the response is described by a *maximum entropy distribution*; that is, a distribution that has the maximum possible variability given the stimulus and constraints set by measurements of the data. In this approach, called the Maximum Noise Entropy method, we compute moments of the measured spiking response with respect to the stimulus and equate these with the same moments calculated with the joint probability distribution from the model (Table 1). A full list of moments across the N dimensions of the stimulus space contains complete information about the neuronal response. However, as in other approaches, it is typically difficult to go beyond two moments.

The functional form of the maximal noise entropy joint distribution, with constraints to second order, (Globerson et al., 2009; Fitzgerald et al., 2011b,a) is given by

$$p(\mathrm{spike}|\mathbf{s}) = \frac{1}{1+\exp\left\{a+\mathbf{h}\cdot\mathbf{s}+\mathbf{s}^{\top}\mathbf{J}\mathbf{s}\right\}}. \tag{31}$$

The parameters of the model are a, a scalar needed to satisfy the zeroth order constraint; $\mathbf{h}$, an N-component vector needed to satisfy the first order constraints and $\mathbf{J}$, an $N \times N$ symmetric matrix needed to satisfy the second order constraints. The distribution (Eq. 31) is found through a convex minimization procedure that evaluates the moments (Model in Table 1) with the constraint $\sum_{t_{\mathrm{s}}} p(\mathrm{spike}|\mathbf{s}(\mathbf{t}_{\mathrm{s}})) = 1$. There is no need to use a spectrally white stimulus with MNE.

3.1.1 Interpreting the model.

How does the MNE model (Eq. 31) ensure the maximal variability in the spike rate? Consider the maximum entropy distribution (Eq. 31) without any constraints, i.e., a = $\mathbf{h}$ = $\mathbf{J}$ = 0. The probability of a spike given a stimulus then is $p(\mathrm{spike}|\mathbf{s})$ = 1/2 and can be thought of as the *least structured* spiking model. At every time bin the neuron will fire or not fire with equal probability. The next simplest model is the one where the probability of a spike is independent of the stimulus $p(\mathrm{spike}|\mathbf{s})$ = $p(\mathrm{spike})$, but the overall firing rate is constrained to be the experimentally measured rate r_0. Now the goal of the fitting procedure is to find a such that r_0 = $p(\mathrm{spike}|\mathbf{s})$ = $1/(1+e^a)$, which yields a = $\log(1/r_0 - 1)$.

In general, when there are multiple parameters and spiking depends on the stimulus, a numerical fitting procedure is required to fit the value of the constraints computed from the data and return the value of the parameters for the second order model (Eq. 31). The zeroth order term, $\langle r(t) \rangle$, has no stimulus dependence and, as explained above, enforces that the average firing rate of the MNE model will equal that of the neuron. The parameters $\mathbf{h}$ and $\mathbf{J}$ act as linear feature vectors analogous to $\boldsymbol{\phi}_{\rm sta}$ and the $\boldsymbol{\phi}_{{\rm stc},i}$:

- Setting $\mathbf{J} = 0$, equivalent to choosing a first order MNE model, results in the model

$$p(\text{spike}|\mathbf{s}) = \frac{1}{1 + \exp\{a + \mathbf{h} \cdot \mathbf{s}\}}. \tag{32}$$

  This is equivalent to a STA model with feature $\boldsymbol{\phi}_{\rm sta} \approx \mathbf{h}$ and a sigmoidal nonlinearity.

- The matrix $\mathbf{J}$ can be decomposed in terms of its eigenvalues λ_i and eigenvectors, denoted $\mathbf{u}_i$, with $i = 1, \ldots, K$, i.e.,

$$\mathbf{J} = \sum_{i=1}^{K} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top. \tag{33}$$

  Defining the projection of a stimulus vector onto an eigenvector as $z_i = \mathbf{u}_i \cdot \mathbf{s}$ allows us to rewrite the quadratic term in Eq. (31) as:

$$\mathbf{s}^\top \mathbf{J} \mathbf{s} = \sum_{i=1}^{K} \lambda_i (\mathbf{s} \cdot \mathbf{u}_i)(\mathbf{u}_i \cdot \mathbf{s}) = \sum_{i=1}^{K} \lambda_i z_i^2. \tag{34}$$

  Therefore, the eigenvectors of $\mathbf{J}$ with large eigenvalues, in absolute value, can be viewed as analogues of the STC features $\boldsymbol{\phi}_{{\rm stc},i}$ with a quadratic-sigmoidal nonlinearity. The match is not exact as the $\boldsymbol{\phi}_{{\rm stc},i}$ were calculated from the covariance difference matrix (Eq. 18), which is taken relative to $\boldsymbol{\phi}_{\rm sta}$. Similarly to the STC method, the eigenvectors of $\mathbf{J}$ are orthogonal to each other and thus we may not interpret these spatiotemporal vectors as independent receptive fields that drive the cell's response.

In the STC approach, the significance of a given feature was determined by comparing the corresponding eigenvalue of $\boldsymbol{\Delta}\mathbf{C}$ (Eq. 18) to the null distribution constructed using shuffled spike trains. Here, because the model parameters are estimated using a gradient ascent algorithm, we cannot construct a null model using shuffled spike trains. It is still possible, however, to estimate which of the eigenvalues of $\mathbf{J}$ correspond to features, denoted $\mathbf{u}_i$, that significantly modulate the spiking output. We accomplish this by shuffling the entries of $\mathbf{J}$ and computing the eigenvalues of the shuffled matrix. Note that the shuffled matrix must remain symmetric and the diagonal and off-diagonal elements should be shuffled separately. Eigenvalues of the matrix $\mathbf{J}$ obtained from the real data are said to be *significant* only if they exceed the range calculated using this shuffling procedure, since the shuffled matrix represents a set of features with the same statistics as the components of the MNE models, but without the structure.

3.1.2 Features from thalamic spiking during whisking in rat.

We applied the MNE procedure to the datasets obtained from thalamic recordings while rats whisked in air, as shown for a representative unit in Fig. 8. As expected, the calculated first-order feature, $\mathbf{h}$, closely approximated the STA feature (Fig. 8A). We found that the top nine of $N = 150$ eigenvectors

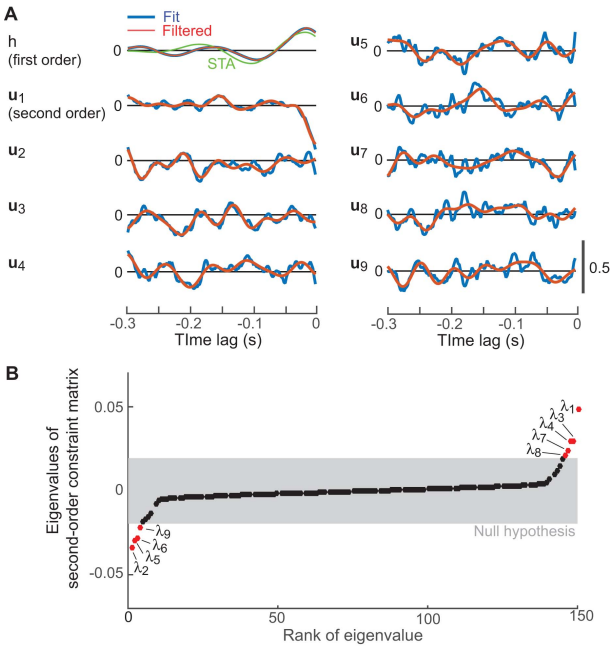

Figure 8: **The dominant features calculated by the Maximum Noise Entropy method for example thalamic cell 57.** (**A**) We fit a MNE model to the spike train with the same stimulus representation, with $N = 150$, and plot the first feature, i.e., $\mathbf{h}$, and statistically significant second-order feature vectors, i.e., eigenvalues of J (Eq. 3.1.1). We also plot the STA feature next to the first order mode for comparison. (**B**) The number of significant second order features was found by comparing the eigenvalues of $\mathbf{J}$ to a null distribution.

of $\mathbf{J}$ were statistically significant (Fig. 8B). The dominant feature, $\mathbf{u}_1$, makes a substantial contribution at short times, like the dominant STC feature $\phi_{\mathrm{stc},1}$ (Fig. 7A), but decays much more rapidly than the STC feature. The higher order features calculated from $\mathbf{J}$, i.e., $\mathbf{u}_2$ through $\mathbf{u}_9$, correspond to variations in the stimulus from whisk to whisk and have no clear interpretation.

A number of practical matters arise. First, in its raw form, the fitting procedure can generate high frequency components. In the present case, we filtered the significant features by removing the components orthogonal to the first 15 principal components of the stimulus. Second, we use the full matrix $\mathbf{J}$ that was found by the fitting procedure to generate the predictions using this model (discussed later under Model Evaluation). Removing the insignificant eigenvectors often leads to poor predictions because the average spike rate predicted from the model no longer exactly matches the zeroth moment, i.e., the average firing rate, since the projections onto the insignificant eigenvalues of $\mathbf{J}$ do not sum exactly to zero. Second, while potential issues with over-fitting are always an issue, they did not arise with this dataset, possibly because of the rapid fall-off of the eigenvalues for the covariance of the stimulus matrix (Fig. 6E). We return to the issue of overfitting when we discuss validation of the models and note that the MNE method was particularly susceptible to overfitting for white noise stimuli.

3.2 Separability

The feature vectors in the first two models we discussed, namely STA and STC, are computed directly from the spike-triggered and prior stimulus distributions, and do not require a fitting procedure to be

applied. As such they do not suffer significantly if the the stimulus space is expanded, for example by assuming that the spiking depends on the stimulus history further back into the past. However, if the cell's response is found to be modulated by a large number of features, e.g., multiple STC modes, the number of spikes will severely limit how many of these can be incorporated in a predictive model. In the second-order MNE model, on the other hand, the number of parameters scales as the dimensionality of the stimulus squared, i.e., N^2. Therefore it may suffer from overfitting as a large number of spikes is required to accurately fit the parameters.

Here we discuss two forms of *separability*, which can be thought of as approximations that two or more of the model components act independently. If these approximations are accurate for a given cell, they may greatly reduce the number of spikes needed to fit the model or help prevent overfitting.

3.2.1 Separability of a feature vector

Many stimuli, such as the checkerboard presented for the retinal studies, consist of both spatial and temporal components. Yet only a small number of these $N_X \times N_T$ components (Eq. 9) are likely to be significant. The spatiotemporal features ϕ_{i} may, in general, be expanded in a series of outer products of spatial modes and temporal modes (Golomb et al., 1994). We define these as $\phi i_{\mathrm{i}}^{X,d}$ and $\phi_{\mathrm{i}}^{T,d}$, respectively, where d labels the mode.

We express ϕ_{i} in the same form of a matrix for the space time stimulus (Eq. 9), i.e.,

$$\phi_{\mathrm{i}}(x,t) = \begin{pmatrix} \phi_{\mathrm{i}}(1,1) & \cdots & \phi_{\mathrm{i}}(1,N_T) \\ \vdots & \ddots & \vdots \\ \phi_{\mathrm{i}}(N_X,1) & \cdots & \phi_{\mathrm{i}}(N_X,N_T) \end{pmatrix} \tag{35}$$

(N_X spatial positions; N_T time points)

$$= \sum_{d=1}^{\min(\mathrm{N_X},\mathrm{N_T})} \lambda_d \begin{pmatrix} \phi_{\mathrm{i}}^{X,d}(1) \\ \vdots \\ \phi_{\mathrm{i}}^{X,d}(N_X) \end{pmatrix} \begin{pmatrix} \phi_{\mathrm{i}}^{T,d}(1) & \cdots & \phi_{\mathrm{i}}^{T,d}(N_T) \end{pmatrix} \tag{36}$$

where λ_d is the weight of the d^{th} mode of the feature, also referred to as the singular value in singular value decomposition.

A great simplification occurs if the dependence on spatial components and temporal components is *separable*. In this case, the spatiotemporal features are well approximated by the product of a single, i.e., d = 1, spatial and temporal contribution. This corresponds to a single spatial pattern that is modulated equally at all pixels by a single function of time. This assumption reduces the number of parameters one needs to estimate, per feature, from $N_X \times N_T$ to $N_X + N_T$.

3.2.2 Separability of the nonlinearity

Another important form of separability relates to the nonlinear function $g(\cdot)$. While the nonlinearity $g(\cdot)$ can be any positive function of the K stimulus components z_i, the amount of data required to fit $g(\cdot)$ over multiple dimensions is prohibitive. It is possible to get around this data requirement

by making assumptions about $g(\cdot)$. First, one might assume that the nonlinearity is separable with respect to its linear filters (Slee et al., 2005). Under this assumption, $g(\cdot)$ can be written as:

$$g\left(z_1, \ldots, z_K\right) = g_1(z_1) \times \cdots \times g_K(z_K). \tag{37}$$

This approximation is equivalent to assuming that the joint conditional probability distribution over the projections of the stimulus on the filters, $p(z_1, z_2, \ldots, z_K|\text{spike})$, is equal to the product of the marginal distributions, $p(z_1|\text{spike}) \ldots\ p(z_K|\text{spike})$. The validity and quality of this approximation can be quantified using mutual information (Adelman et al., 2003; Fairhall et al., 2006), which is a measure of the difference between joint and independent distributions.

Beyond the enormous reduction in the number of spikes sufficient to accurately fit the model, a separable model that makes reasonably good predictions can help us interpret the model and potentially relate it to circuit and biophysical properties of the system. A successful separable model implies that the cell is driven by processes that are, to a good approximation, independent. These could be, for example, inputs from parallel pathways such as separate dendrites or subunits, or the effects of feed-forward versus feedback processing. A specific example of a model whose typical application generally assumes that different factors influencing the firing of the neuron contribute independently and multiplicatively is the *generalized linear model*.

3.3 Generalized Linear Models (GLM)

While the models so far only consider stimulus dependence, the biophysical dynamics of the neuron or local circuit properties might alter the ability of the cell to respond to stimuli as a function of its recent history of activity. For example, all neurons have a relative refractory period that could prevent them from spiking immediately after a previous spike, even if the stimulus at that time is one that normally strongly drives the cell (Berry and Meister, 1998). Further, projection neurons have a tendency to emit bursts of spikes, such that the probability of a spike will be increased if the cell has recently spiked (Magee, 2003). These effects, and other more general dependencies, can be incorporated in the framework of a generalized linear model (GLM) (Nelder and Wedderburn, 1972; Brown et al., 1998).

Generalized linear models are a flexible extension of standard linear models that allow one to incorporate non-linear dependencies on any chosen set of variables, including the cell's own spiking history. They gain this ability to incorporate a richer set of inputs by taking an explicit form for the nonlinear function $g(\cdot)$ to reduce demands on data. A GLM is characterized by the choice of $g(\cdot)$ and by a noise model that specifies the distribution of spiking, required to be within a class of distributions known as the exponential family. This includes many appropriate probability distributions, e.g., binomial, normal, and Poisson. As in previous approaches, we choose a Poisson process, for which the probability of counting n spikes in a time bin of width Δt at time t is determined by the predicted firing rate $r(t)$ averaged over that time bin, i.e.,

$$p(n \,\text{spikes between}\, t - \Delta t \,\text{and}\, t) = \frac{(r(t)\Delta t)^n(t)}{n(t)!} e^{-r(t)\Delta t}. \tag{38}$$

The firing probability is taken to be a function $g(\cdot)$ of a linear combination of the stimulus, the recent spiking of the cell, and potentially other factors (Fig. 1A). In its simplest form, the spike rate is given

by

$$r(t) = g\left(a + \sum_{t'<t} \boldsymbol{\phi}_{\mathrm{glm}}(t') \cdot \mathbf{s}(t') + \sum_{t'<t} \boldsymbol{\psi}(t') n(t')\right). \tag{39}$$

where the parameter a sets the overall level of the firing rate, the sum $\sum_{t'<t} \boldsymbol{\phi}_{\mathrm{glm}}(t') \cdot \mathbf{s}(t')$ is the familiar projection of the stimulus onto the spatiotemporal feature $\boldsymbol{\phi}_{\mathrm{glm}}(t)$, and we have now included a temporal *spike history filter*, denoted $\boldsymbol{\psi}(t)$, which is a N_{h} dimensional vector that weights the recent activity of the neuron. Together we refer to the set of parameters for the GLM as Θ.

As before, $r(t)$ is a function of the stimulus and depends on all parameters, Θ, of the model. The task is to determine the optimal value of Θ given the specific observed sequence of spike counts. This is done by maximizing the *likelihood*, i.e., the probability of the data given the parameters viewed as a function of the parameters, $\mathcal{L}(\Theta) = P(n(t)|\Theta)$, over choices of Θ.

When the nonlinearity $g(\cdot)$ is both convex and log-concave, the likelihood function will itself be a convex function. This means that the likelihood $\mathcal{L}(\Theta)$ has a single, global optimum that can obtained through any convex optimization routine. Fortunately nonlinearities that satisfy this property include common choices like the exponential and the piecewise linear-exponential function (Paninski, 2004). Thus we maximize the log-likelihood, which for Poisson spiking is

$$\log \mathcal{L}(\Theta) = \sum_t \log\left(r\left(\mathbf{s}(t)|\Theta\right) \Delta t\right) - \sum_t \left(r(t)\Delta t\right). \tag{40}$$

where $r\left(\mathbf{s}(t)|\Theta\right)$ is the predicted firing rate. With this, the computational fitting problem we solve is simply

$$\underset{\Theta}{\operatorname{argmax}} \left(\log \mathcal{L}(\Theta)\right), \tag{41}$$

which can be maximized through a convex optimization routine of choice.

3.3.1 Overfitting and regularization

As for other methods, the fitted model may best fit the training data but not generalize to test datasets. In a likelihood framework, overfitting is simple to understand: one can always improve the log-likelihood simply by adding more parameters. Indeed, if the number of parameters encompassed by Θ is the same as the dimensionality of $n(t)$ we can construct a model that fits the observed data *exactly*. But this is not the aim of constructing a model. Rather, we seek to find a model that captures trends in the data that are common across different samples, rather than details of individual fluctuations.

Overfitting arises either as a result of insufficient training data relative to the number of parameters being estimated or from the model containing more parameters than are needed to describe the relationship under consideration. As discussed with respect to natural stimuli, correlations in the input reduce its effective dimensionality of the data and thus the number of parameters required in the model. A common effect in GLMs that are fit to slowly-varying stimuli is the presence of high frequency components in the feature vector, as occurred for the MNE model, since such fast variations projected onto the slowly-varying stimulus cancel out and minimally effect the predicted spike trains and log-likelihood (Eqs. 40 and 41). While their effect on the log-likelihood may be minimal, they obstruct interpretation of the feature vectors $\boldsymbol{\phi}_{\mathrm{glm}}$. Such overfitting can be avoided by penalizing models that are over-parameterized by adding a penalty term $Q(\Theta)$ to the quantity we are maximizing:

$$\underset{\Theta}{\operatorname{argmax}} \log \mathcal{L}(\Theta) - Q(\Theta). \tag{42}$$

For instance, to avoid overfitting we might choose the term $Q(\Theta)$ to be large for models that contain a large number of non-zero parameters. The simple choice,

$$\underset{\Theta}{\operatorname{argmax}} \log \mathcal{L}(\Theta) - N_{\Theta}, \tag{43}$$

where N_{Θ} is the number of parameters of the model, is known as the Akaike Information Criterion (Akaike, 1973; Boisbunon et al., 2014). This and related criteria provide a simple, principled means to choose between competing models of differing numbers of parameters and may be used to determine the optimal stimulus and history filter sizes (Shoham et al., 2005).

Penalty terms may be interpreted as representing *prior* knowledge relevant to the estimation problem. In particular, if one has a prior distribution on the space of parameter estimates, $p_{\Theta}(\Theta)$, one can use Bayes' rule to find an estimate that maximizes the *a posteriori* probability, denoted $\mathbf{\Theta}_{MAP}$, where

$$\begin{aligned} \mathbf{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}} \log p(\Theta|\mathbf{s};\mathbf{r}) \\ &= \underset{\Theta}{\operatorname{argmax}} \left(\mathcal{L}(\Theta;\mathbf{s},\mathbf{r}) + \log p_{\Theta}(\Theta)\right). \end{aligned} \tag{44}$$

Then we can identify the penalty term as the negative logarithm of the prior, i.e., $Q(\Theta) = -\log p_{\Theta}(\Theta)$. For instance, if one expects the feature vector to be smooth, one might apply a Gaussian prior:

$$Q(\Theta) = \lambda \Theta^{\top} \mathbf{D} \Theta. \tag{45}$$

The function $Q(\Theta)$ will penalize feature vectors that are not smooth or that vary excessively when $\mathbf{D}$ is chosen to be a second-derivative operator (Linden et al., 2003). The weight λ is often chosen to maximize the model's performance on data withheld from the optimization procedure.

Finally, a very simple heuristic that sometimes mimics the effect of these regularization methods to avoid overfitting is early stopping. Here we simply limit the number of iterations in the fitting process to effectively stop the fitting before the unique solution is found. This approach assumes that solutions near the optimal one for the training data are good and also lead to generalization. This involves monitoring the form of the solution at each step of the optimization and choosing the number of iterates that recovers a reasonable solution.

3.3.2 Choice of basis

For completeness, overfitting can be avoided by forbidding rather than just penalizing models that are over-parameterized. This is achieved by reducing the number of parameters of the model to a value known through experience to be reasonable. While we have discussed previously the simple expedient of downsampling or truncating the data, more generally one can project the stimulus into a subspace that captures important properties of the data; the basis vectors for this subspace then define the number of parameters of the stimulus feature vector. One natural choice is to use the leading principal components of the stimulus (Eq. 16) as the basis set. In the case of the spike-history filter, one can choose basis functions that are appropriate to capture the expected biophysics of the neuron, such as refractoriness or burstiness. A common set of basis functions for representing spike history filters is a 'raised cosine' basis:

$$k_i(t) = \begin{cases} \frac{1}{2}\left(\cos\left[a\log(t-\psi_i) - \phi_i\right] + 1\right), & \phi_i - \pi < a\log(t-\psi_i) < \phi_i + \pi; \\ 0, & \text{otherwise;} \end{cases} \tag{46}$$

that describes a sequence of bumps whose peaks are tightly spaced near the time of the spike, and become increasingly sparse for earlier times. In this way the basis is well resolved where the spike history filter changes most rapidly (Pillow et al., 2008).

3.3.3 Stability

Despite much theory surrounding their application (Paninski, 2004), correctly specifying a GLM using appropriate timescales and basis functions remains as much an art as a science. Particular care must be taken in correctly parameterizing spike history filters. One approach is to initially fit the model with no special basis functions, examine the resulting filters, and then choose a parameterization of a reduced basis, e.g., raised cosines or exponentials that allows for the form obtained in the full dimensional case. While this involves fitting a full dimensional model, a lower dimensional model is ultimately obtained that is less likely to be overfit.

Unfortunately nothing guarantees that the maximum likelihood estimate of a GLM will be stable. Unstable models diverge when used to simulate novel spike trains. While such models may still provide insight from the form of their feature vectors, they are not able to simulate spike trains on novel stimulus datasets, the essence of model validation. If unstable GLMs are encountered, one should first check that the parameterization of the spike history filter accurately characterizes the neuron's refractory period. In this regard, improper spike sorting that leads to the presence of spike intervals that are less than the refractory period (Hill et al., 2011b) can cause misestimation of the spike history term and lead to instability.

3.3.4 Features from retina and thalamic cells.

We fit GLMs for the set of retinal ganglion cells stimulated with white noise (Fig. 3) and thalamic neurons stimulated by self-motion of the vibrissae (Fig. 6). We consider the white noise case first. A sequence of delta functions, i.e., independent pixels, was used as the basis functions for the stimulus feature vector and raised cosines were used as basis of the spike history filter (Eq. 46). For the same representative neuron used previously, the feature vector $\boldsymbol{\phi}_{\mathrm{glm}}$ corresponds to a transient spot of illumination that is similar to the STA feature yet slightly delayed in time (Fig. 9A). This shift is presumably the effect of the spike history dependence, which leads to increased firing rate approximately 40 ms after the previous spike, a time scale similar to the stimulus refresh, Δt. Since the effect of the spike history filter is exponentiated, we plot both the result of the fit (black line, Fig. 9B) and the exponent of the filter (gray line, Fig. 9B) to illustrate the effect that this component of the model has on spiking.

The GLM fit in the case of correlated noise gives a less intuitive, but thus perhaps more interesting, result. Here the twelve leading terms of a PCA of the stimulus were used as the basis functions for the stimulus feature vector and raised cosines were used as the basis of the spike history filter (Eq. 46). Here the feature vector $\boldsymbol{\phi}_{\mathrm{glm}}$ oscillates for one cycle then returns to baseline very quickly, before even a single whisk is completed (Fig. 9C), and thus is quite different than the $\boldsymbol{\phi}_{\mathrm{sta}}$. Further, while the spike history shows a significant excitatory component, this component is extremely short lived (Fig. 9D). The GLM analysis therefore suggests that the thalamic cell is very responsive to instantaneous

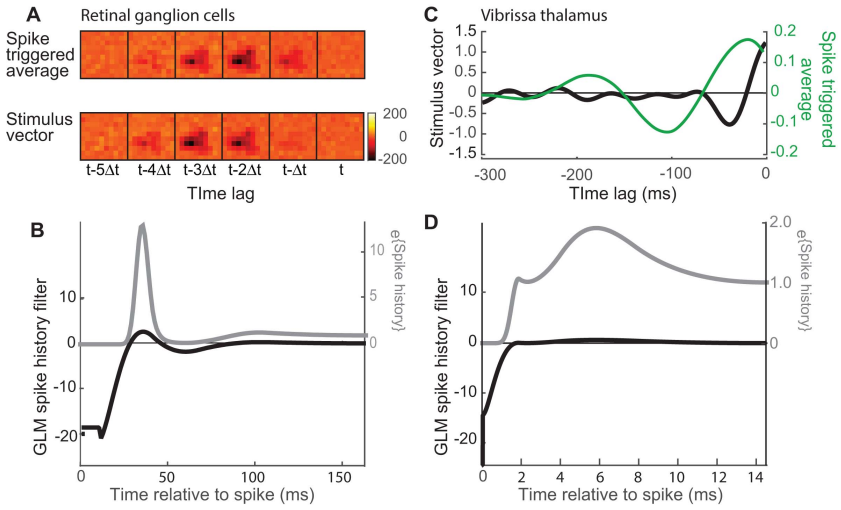

Figure 9: **The fit of the generalized linear model for the responses of retina ganglion cell 3 and thalamic vibrissa cell 57. (A,C)** The stimulus feature ϕ_{glm} compared with the previously calculated STA feature. **(B,D)** The spike history filters ψ (black curves). We also plot the exponent of the filter (gray), as it is exponentiated in the model, to better illustrate the effect this component of the model has on spiking.

changes in position of the vibrissae but has little dependence on the history of the stimulus or past spiking beyond around 80 ms, corresponding to about half of a whisk.

4 Model evaluation

We now consider how well each of the models performs in predicting the spike rate for a fraction of the stimuli, designated the test set, that has the same statistical properties as the training set but is otherwise novel. In every case, 80% of the data was used as the training set for fitting the model, and the remaining 20% was reserved for testing. A number of measures are available to test the quality of the model in predicting spikes. The most direct and intuitive is the root mean square of the difference between the recorded firing rate $r_{\mathrm{s}}(t)$ and that predicted by the model. Ideally this would be computed for responses to a repeated but rich stimulus so that one could estimate the intrinsic variability of the neuronal spiking response. However, here and generally for natural stimuli, one only has a single presentation of the stimulus, or the relationship between the external variable and the spike train may be inherently non-repeatable, as during behavior when the stimulus is under the animal's control.

4.1 Log-likelihood

In this case, one can compare the log-likelihood of the data given the model for different models. For Poisson spiking, (Eq. 38), this is

$$\log \mathcal{L}\left(\phi_i\right) = \sum_t \left(n_{\mathrm{s}}(t) \log \left(\sum_i \phi_i \cdot \mathbf{s}(t) \Delta t \right) - \sum_i \phi_i \cdot \mathbf{s}(t) \Delta t - \log\left(n_{\mathrm{s}}(t)!\right) \right). \tag{47}$$

Typically, the log-likelihood estimate has a common large offset that depends only on the firing rate and a small range of variation of the term $\log\left(\sum_i \phi_i \cdot \mathbf{s}(t)\right)$ among different models because of the

logarithmic compression. To estimate a lower bound on the log-likelihood, we replace the calculated rate with the measured rate to form a null hypothesis, i.e.,

$$\log \mathcal{L}\left(\text{null}\right) = \sum_t \left(n_{\rm s}(t)\left(\log\left(n_{\rm s}(t)\right) - 1\right) - \log\left(n_{\rm s}(t)!\right)\right). \tag{48}$$

4.2 Spectral coherence

A complementary metric for the fidelity of the predicted spike trains is the spectral *coherence* between the predicted and measured responses. This measure can distinguish the performance of different models across different frequency bands, each of which may have particular behavioral relevance. We define $\tilde{r}(f)$ and $\tilde{r}_{\rm s}(f)$ as the Fourier transform of the predicted and measured rates, respectively. The spectral coherence, denoted $\tilde{C}(f)$, is:

$$\tilde{C}(f) = \frac{\langle \tilde{r}^\star(f)\tilde{r}_{\rm s}^\star(f)\rangle}{\sqrt{\langle|\tilde{r}(f)|^2\rangle\,\langle|\tilde{r}_{\rm s}(f)|^2\rangle}}. \tag{49}$$

where the multi-taper method is used for averaging, $\langle\ldots\rangle$, over a spectral bandwidth that is larger than the Raleigh frequency $1/(N_T\Delta t)$ (Thomson, 1982; Kleinfeld and Mitra, 2011). The magnitude of the coherence reports the tendency of two signals to track each other within a spectral band and is normalized by the power in either signal. The phase of the coherence reports the relative lag or lead of the two signals. There are no assumptions on the nature of the signals. The confidence level is determined by a jack-knife procedure (Thomson, 1982).

Spectral coherence may be viewed in analogy to the Pearson correlation coefficient in linear regression, i.e., to the extent that real and imaginary parts of both $\tilde{r}(f)$ and $\tilde{r}_{\rm s}(f)$ may be considered as Gaussian variables, $\tilde{C}(f)$ forms part of the regression coefficient. The expected value of the predicted rate given the observed rate is:

$$\mathcal{E}\left(\tilde{r}(f)|\tilde{r}_{\rm s}(f)\right) = \tilde{b}(f)\tilde{r}_{\rm s}(f) \tag{50}$$

where the coefficient $\tilde{b}(f)$ is:

$$\tilde{b}(f) = \tilde{C}(f)\sqrt{\frac{\langle|\tilde{r}(f)|^2\rangle}{\langle|\tilde{r}_{\rm s}(f)|^2\rangle}}. \tag{51}$$

The variance of the expectation, denoted $\mathcal{V}\left(\tilde{r}(f)|\tilde{r}_{\rm s}(f)\right)$, is given by

$$\mathcal{V}\left(\tilde{r}(f)|\tilde{r}_{\rm s}(f)\right) = \left(1 - |\tilde{C}(f)|^2\right)\langle|\tilde{r}(f)|^2\rangle \tag{52}$$

and, of course, goes to zero when measured and predicted signals are the same.

4.3 Validation of models with white noise stimuli

The predictions with the STA model, the STC plus STA model, and the GLM capture the gross variations in spike rate (Fig. 10A,B). The GLM yields representative spike trains, as opposed to rates, so that we computed predicted rates by averaging over many spike trains computed by repeatedly presenting the same stimulus to the same model. In these predictions, many spikes are unaccounted for, while the spike probability also indicates spikes when none occur. Interestingly, the STA plus STC model has the highest value of the log-likelihood (Eq. 47) while the GLM has the lowest, lower even than the STA (Fig. 10C). This may imply overfitting of the training data with the GLM, as models that

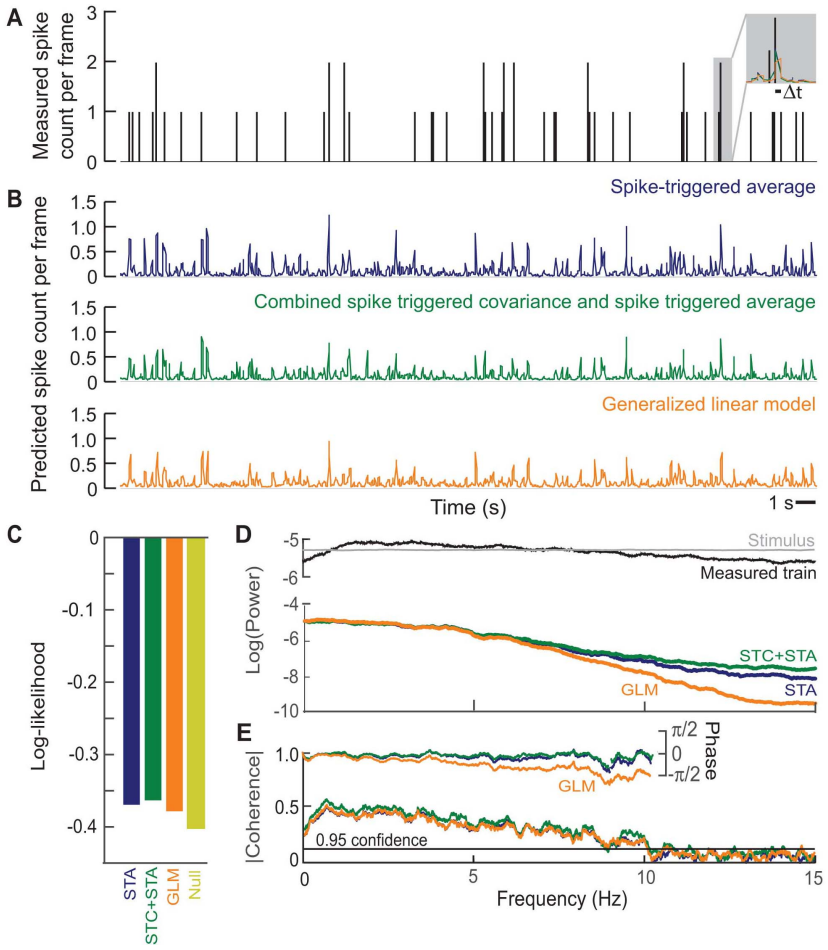

Figure 10: **Summary of the performance of model predictions for the retinal ganglion cell 13; three methods, STA, STC plus STA, and GLM, are compared. (A)** A part of the spike train cut out from the test set for illustration purposes. **Insert** Expanded temporal scale to highlight the slight delay inherent with the GLM. **(B)** The predicted spike count per frame obtained by computing the probability of a spike corresponding to each stimulus frame (top, STA; middle, STC; bottom, GLM). Note that to generate a prediction from the GLM at time t we need the history of the spike train up to that point $t' < t$, which is not deterministic due to the Poisson variability. Thus, the trace presented here (orange) is the average spike count over 500 simulations of the GLM on the test set. **(C)** The log-likelihood (Eq. 47) of each model given the test set, which quantifies the quality of the prediction. We also include the log-likelihood for the null condition (Eq. 48). **(D)** The spectral power of individual pixels in the stimulus (black) and the recorded spike train (gray), as well as those of the predicted spike trains. The mean value has been removed, so that the initial data point represents an average over the spectral half-bandwidth. Spectra were computed with a half-bandwidth of 0.087 Hz as an average over 159 spectral estimators for 920 s of data **(E)** The phase and magnitude of the spectral coherence between the recorded and predicted spike train for each method. Coherence was computed with a half-bandwidth of 0.065 Hz as an average over 119 spectral estimators.

involve more parameters have a "higher" log-likelihood. All models, however, perform better than the null expectation (Eq. 48) (Fig. 10C).

Greater insight into fitting of the models is provided by a spectral decomposition. First, the spectral power of the stimulus is constant, by design (Fig. 10D), and the power of the spike train decreases only weakly with increasing frequency, consistent with a Poisson process. The spectral power for the spike rates predicted from three models, STA, STC plus STA, and GLM, show a rather strong frequency dependence. The coherence is substantially below $|\tilde{C}(f)|$ = 1 at all frequencies yet it is statistically significant (Fig. 10E). Consistent with expectations from the log-likelihood (Fig. 10C), the STC plus STA model has an approximately 5% improved coherence at all frequencies (Fig. 10E). The GLM yielded the worst predictions. Further, while the phase for the STA and STC plus STA models is close to zero, which implies that the predicted spikes arrive at the correct time, the phase is a decreasing function of frequency for the GLM model (Fig. 10E). This implies that the predicted spikes arrive with a brief time delay, as noted earlier (Fig. 10A), that is estimated to be $(1/2\pi)(\Delta\mathrm{phase}/\Delta f)$ = -25 ms or less than Δt (inset Fig. 10A).

4.3.1 Synopsis

For the white noise stimulus and this particular set of retinal ganglion cells, the data appear to be adequately modeled by the single STA feature and the accompanying nonlinearity (Fig. 4C,D). The coherence shows an improvement with the STC plus STA model (Fig. 10E). The GLM gives the worst predictions by all measures, and the predicted spikes occur with a shift in timing compared to the test data. Time delays relative to reverse correlation approaches have been seen in past implementations of the GLM as well (Mease et al., 2014).

Not surprisingly, the MNE model, with a large number of parameters, was susceptible to over-fitting. The parameters from fitting the stimulus set with $n = 600$ (Figs. 4C, 5A, and 10A) led to a stable calculation of the linear feature, $\mathbf{h}$, and three statistically significant second-order features (Eq. 31), but the model gave poor predictions. The log-likelihood metric was *lower* for the MNE model than for the null hypothesis (Fig. 10C) and the spectral coherence was relatively small. To reduce overfitting, we truncated the stimulus from six to two frames to reduce the number of stimulus components to $n = 200$. This procedure led to a linear feature and a single, statistically significant second-order feature. The log-likelihood for this reduced model was now greater than that of the null hypothesis, although still less than that for all other models. Similarly, truncation of the stimulus led to an increase in the spectral coherence at all frequencies, although the coherence was still lower than that achieved with the other models.

4.4 Validation of models with correlated noise from self-motion

We now consider the case of models for whisking cells in thalamus (Figs. 6 and 11). Here, the underlying stimulus is highly correlated and strongly rhythmic (Fig. 11A) with a broad spectral peak at the fundamental and harmonic frequencies of whisking (Fig. 11D); recall that the stimulus has its slowly varying midpoint removed (Fig. 6D). Despite the structure in the stimulus, the spectrum of the spike train of our example thalamic cell, (Fig. 11D) was largely featureless.

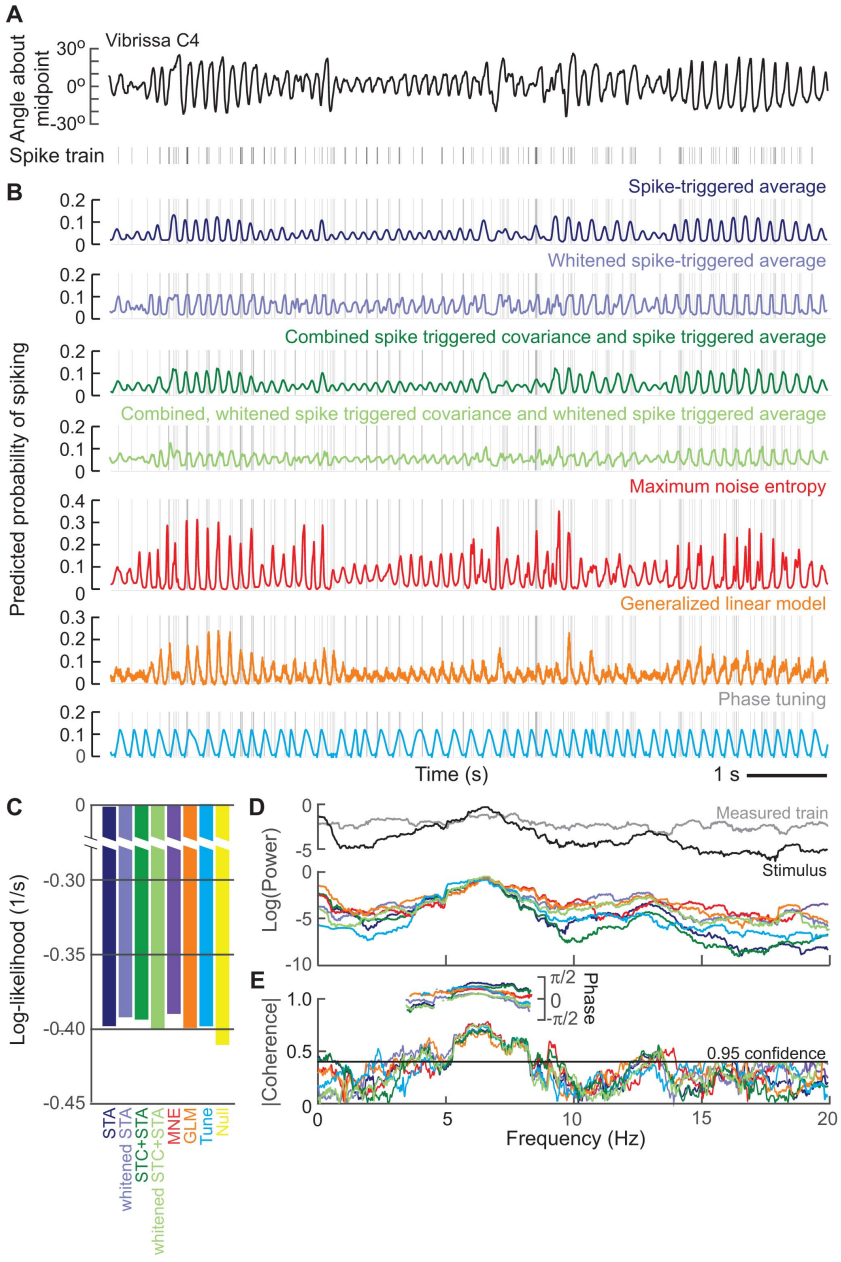

Figure 11: **Summary of the performance of predicted spike trains for thalamic neuron 3. Seven means of analysis are compared, i.e., STA and STC plus STA, STA and STC plus STA after whitening of the stimulus, MNE, GLM, and a phase tuning curve model.** (**A**) The stimulus corresponds to vibrissa position with slowly carrying changes in the set-point removed. (**B**) The predicted probability of spiking per 2 ms time bin obtained by computing by each model and the corresponding stimulus. Note that to generate a prediction from the GLM at time t we need the history of the spike train up to that point $t' < t$, which is not deterministic due to the Poisson variability. Thus, the trace presented here (orange) is the average spike count over 500 simulations of the GLM on the test set. (**C**) The log-likelihood (Eq. 47) of each model given the test set, which quantifies the quality of the prediction. (**D**) The power spectra of individual pixels in the stimulus (black) and the recorded spike train (gray), as well as those of the predicted spike trains. Spectra were computed with a half-bandwidth of 0.6 Hz as an average over 23 spectral estimators.(**E**) The phase and magnitude of the coherence between the recorded and predicted spike train for each method (Eq. 49). Coherence was computed with a half-bandwidth of 1.2 Hz as an average over 49 spectral estimators.

We first ask if whitening the stimulus does indeed lead to an improved prediction. We computed the predicted rate from the feature vector for the STA model, i.e., $\phi_{\rm sta}$, and the feature vector after whitening $\hat{\phi}_{\rm sta}$ (Figs. 7A and 11B). As expected, the log-likelihood is greater after whitening (Fig. 11C). The spectral power for the $\hat{\phi}_{\rm sta}$ is greater at the harmonics, but not the fundamental, compared to the nonwhitened feature vector (Fig. 11D). Interestingly, whitening increases the coherence between the predicted and the measured rates at the whisking frequency, with $|\tilde{C}(f)|$ increasing from 0.70 to 0.75, as well as at other frequencies (Fig. 11E). The exception is that the coherence below about 1 Hz, where variations in the envelope of the whisk may be coded, is better for the nonwhitened STA feature vector.

We further computed the predicted rate for the feature vector for the STC model, i.e, $\phi_{\rm stc,2}$, and the feature vector after whitening, i.e., $\hat{\phi}_{\rm stc,2}$ (Figs. 7A and 11B). Unlike for the STA, the log-likelihood for the STC plus STA model is diminished after whitening (Fig. 11C). As for the STA alone, whitening increases the coherence between the predicted and the measured rates at the whisking frequency, with $|\tilde{C}(f)|$ increasing from 0.70 to 0.75, as well as at other frequencies, with the exception that the coherence below about 1 Hz is better for the nonwhitened STC plus STA feature vectors (Fig. 11E).

Across all models, the best predictability at the whisking frequency occurred with the whitened STA and the MNE models, albeit only by 5 to 10% compared with the STC plus STA model and the GLM. All of the models exhibited a slight phase advance at the whisking frequency. This corresponds to a time shift of approximately $(1/2\pi)(\Delta\text{phase}/f_{whisk})$ = 20 ms, which is worrisome, although short compared to the approximately 160 ms period of a whisk. All told, none of the models was clearly "best" at all frequencies, although the MNE model appeared to be strongly coherent with the measured train at all but the lowest frequencies (Fig. 11E).

It has been shown that whisking may be characterized in terms of a rapidly varying phase (Hill et al., 2011a), denoted $\Phi(t)$. If the firing of neurons is sensitive to phase in the whisk cycle independent of frequency, then a linear feature vector will be a poor representation. We therefore constructed an additional model in which we first applied a nonlinear transformation, the Hilbert transform (Hill et al., 2011a), to the stimulus to extract $\Phi(t)$. We then used Bayes' rule to construct a phase tuning model to compare with the LN approaches (Fig. 6B):

$$p(\text{spike}|\Phi) = \frac{p(\Phi|\text{spike})p(\text{spike})}{p(\Phi)}. \tag{53}$$

The phase model achieves the same high level of coherence at the whisking frequency as the whitened STA and MNE models (Fig. 11D). This suggests that the feature vectors are largely acting as broadband filters. Of course, the tuning model performs badly for frequencies away from the approximately 6 Hz whisking peak (Fig. 11E).

Finally, we consider two additional thalamic neurons that had extreme response properties (Fig. 12). The first is a neuron that tended to spike with respect to changes in the amplitude of whisking (Fig. 12A-C). Here the whitened STA and STC plus STA models did well, the MNE model was most impressive with greatest coherence over the broadest frequency range, and the GLM did poorly (Fig. 12D). On the other hand, we consider a neuron that responds almost solely to the phase of whisking (Fig. 12E-G). All models performed well at the whisking frequency; the phase model performs particularly well (Fig. 12H), and here too the MNE model has higher coherence with the measured rate at both lower and higher frequencies.

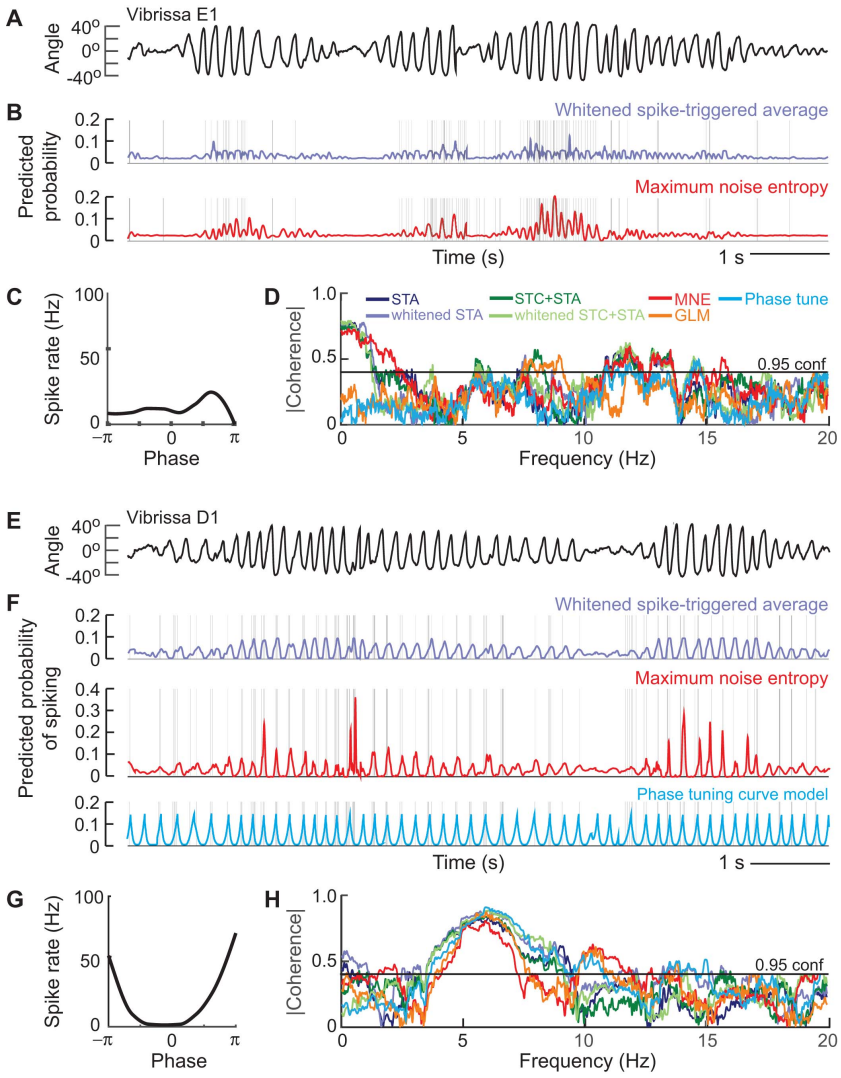

Figure 12: **Summary of the performance of predicted spike trains for two additional thalamic cells, units 88 and 99.** (**A-D**) The whisking stimulus (panel A) and predicted spike probabilities (panel B) for a cell with weak phase tuning (panel C). Yet this cell was strongly modulated by the amplitude of whisking, which changes on a slow time-scale, approximately 1 s, compared with changes in phase. The predicted rate is shown for two models that perform about best, i.e., STA after whitening of the stimulus and MNE. The phase tuning model performs poorly as it ignores the amplitude (panel D). (**E-H**) The whisking stimulus (panel E) and predicted spike probabilities (panel F) for a cell with particularly strong phase tuning (panel G). The predicted rate is shown for three models that perform about best, i.e., STA after whitening of the stimulus, MNE, and the phase tuning model. Here the coherence between the predictions and the measurements in the whisking frequency band is near 1.0 for all models (panel H).

4.4.1 Synopsis

This analysis suggests that for stimuli of this type, a metric for "goodness of fit" based on spectral decomposition offers far more insight than a scalar measure based on maximum likelihood. This may be particularly helpful when certain frequencies may have ethological significance. As for the "best" method with the thalamus data, we were impressed with the results obtained with the MNE model, which fits well over a broad range of frequencies. This stands in contrast to the difficulties in using MNE with the white noise data.

5 Network GLMs

The GLM framework can be readily extended to network implementations of M neurons (Truccolo et al., 2005; Pillow et al., 2008). Each neuron is considered to be driven by a filtered stimulus, its own spiking history and also the filtered activity of the rest of the neurons. If $\boldsymbol{\psi}_{ij}(t)$ $(i, j = 1, \ldots, M)$ is the filter acting on the spiking history of neuron j driving neuron i, then the model for the i^{th} neuron is:

$$r_i(t) = \exp\left\{ a_i + \sum_{t'<t} \boldsymbol{\phi}_i(t') \cdot \mathbf{s}(t') + \sum_{j=1}^{M} \sum_{t'<t} \psi_{ij}(t') n_j(t') \right\}. \tag{54}$$

The incorporation of such network filters have been shown to improve the capability of the model to account for correlations between neurons in a retinal population (Pillow et al., 2008). While it is tempting to interpret the network filters as capturing, for example, synaptic or dendritic filtering of direct interneuronal connections, these terms cannot be taken to imply that two neurons are anatomically connected. For example, correlations might arise from a common input that is not taken into account through the stimulus filter (Kulkarni and Paninski, 2007; Pillow et al., 2008; Archer et al., 2014).

Prior work found coupling terms, $\boldsymbol{\psi}_{ij}(t)$ in a network GLM (Eq. 54), that could be interpreted as functional interaction kernels between cells (Pillow et al., 2008). In that study, model validation of each neuron was done using the stimulus and the recorded activity of the remainder of the cells. This procedure is equivalent to fitting a single-cell model where the stimulus is expanded to include the spiking history of the rest of the network, i.e., the $n_j(t)$. As a practical matter, this procedure has value when one is interested in the precise timing of coupling between cells, e.g., to find whether neurons are anatomically connected (Gerhard et al., 2013). Yet, in our opinion, expanding the stimulus to encompass the spiking history of the rest of the network stands in contrast to validation of a true network GLM, for which the spike histories are based solely on simulations and the only external variable is the stimulus. We use the full network approach in our validation procedure.

5.1 Application to cortical data during a monkey reach task

We present an example of a network GLM based on nine simultaneous recordings from monkey primary motor cortex in which the monkey performs a grip and reach motor task (Engelhard et al., 2013). The model consists of feature vectors that relate to hand motion, as measured by a cursor trajectory and grip force (Fig. 13 A), that were modeled with Gaussian bump' basis functions. Since

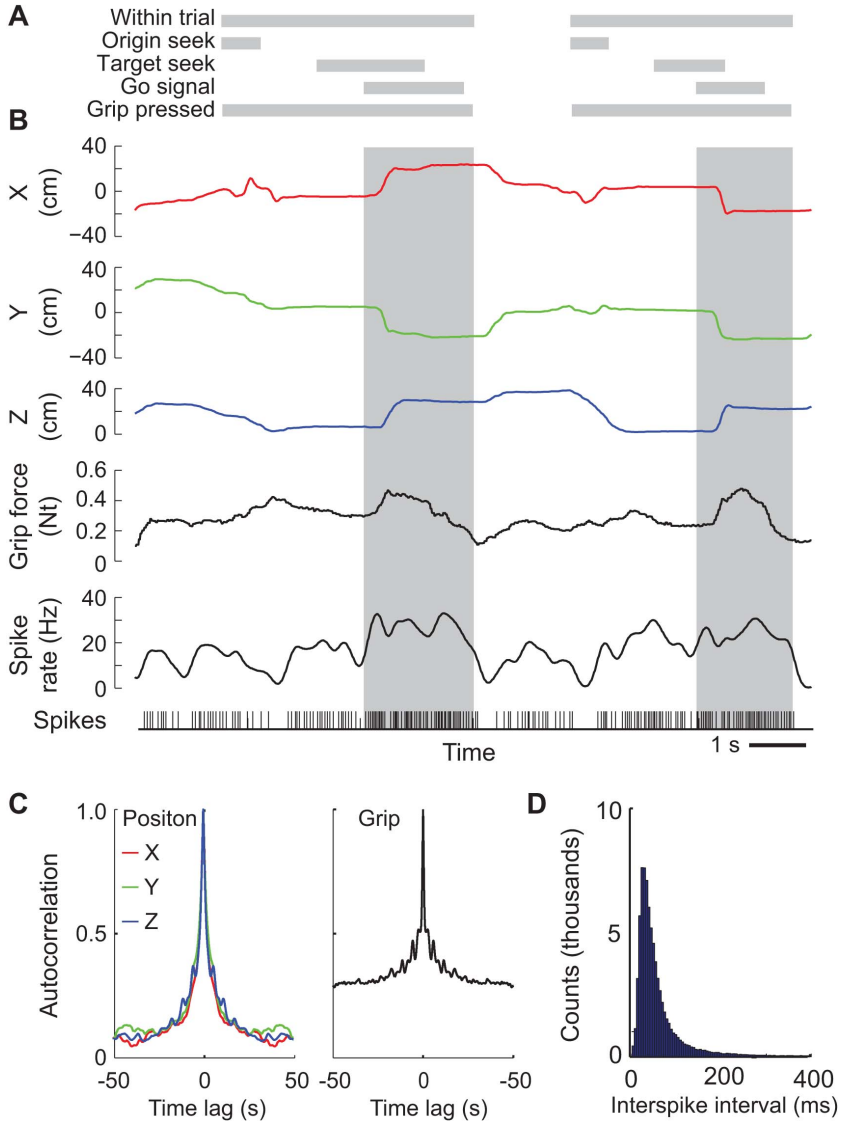

Figure 13: **Summary of the three dimensional monkey-based reach task with spike data from unit 36.** Analysis is based on one approximately 90 minute recording while performing the task both cursor motion and grip force are recorded. **A** Grip-and-reach task involves first moving the cursor to a central position, followed by gripping the handle with sufficient force. Once gripping at the center, after a variable wait time, a target appears randomly in one of 8 locations. Following another wait of a variable time, the cue at the origin disappears, acting as a go signal, after which the monkey may perform the reach movement. Grip on the handle has to be maintained through the duration of the trial. A successful trial requires reaching the target within a set time limit. Once the target is reached, the monkey needs to hold the cursor at the target for 700 ms, and to release its grip on the handle. Following a successful trial, the monkey receives a reward, and after an inter-trial period the next trial begins. **B** Measured cursor position and grip force. **C** Stimulus auto-correlation. **D** Distribution of interspike intervals shows a clear refractory period. **Methods:** Spikes were recorded from single isolated units in the contralateral cortex to the task arm using an intracortical multi electrode array (Blackrock Utah array) implanted in the arm region of M1. Spiking data were binned into millisecond intervals, while both cursor data and grip force are sampled at 100Hz. Of the isolated units, we selected those which showed no evidence of contamination based on inspection of the interspike interval distribution. Analysis was performed from the time of the Go signal until the grip was released; see gray band in panel B.

motor neurons encode future motor outputs, the 'stimulus' filter encodes both causal and acausal relationships, in that it is applied to past and future measurements of cursor and grip, relative to the current time-bin (Fig. 13B). Similar choices with GLMs have been previously applied to neurons in motor cortex (Shoham et al., 2005; Truccolo et al., 2005; Saleh et al., 2012). Lastly, we used raised cosine basis functions (Eq. 46) for the spike history filters and the coupling filters for the histories other neurons in the network.

The cursor position and grip data varies over hundreds of milliseconds to seconds (Fig. 13C), while the spike history data varies on the order of milliseconds. Capturing effects on these separate time scales within the same model requires some care, as the data is non-Gaussian and highly temporally correlated. As noted previously, such correlation can result in uninterpretable high frequencies in the feature vectors. This requires some form of regularization. The approach adopted here is to use only a limited number of basis vectors that sparsely sample the stimulus at regular intervals, with the interval size on the order of the stimulus auto-correlation time scale.

The fitting was performed only on data within the movement phases of the trials, excluding the hold periods (Fig. 13A). In order to avoid unnecessary coupling terms, a group least absolute shrinkage and selection operator (LASSO) (Yuan and Lin, 2006) penalty is applied to the sets of parameters representing of connections between neurons. This takes the form

$$\underset{\Theta}{\operatorname{argmax}} \log \mathcal{L}(\Theta) - \lambda \sum_{i \neq j}^{I} \|\Theta_{i,j}\|^2. \tag{55}$$

where $\{\Theta_{i,j}\}$ are the parameters representing the coupling from neuron j to neuron i. A similar penalty is applied in prior work (Pillow et al., 2008). The penalized likelihood is still convex, which ensures global convergence.

5.2 Validation

As in the previous cases, the model is validated by splitting the data into a training set representing 80% of the total data. A test set representing 20% of the total data, or 4 minutes of recording, is used for validation. We take a value λ =100 in our network analysis (Eq. 55); smaller values decreased the log-likelihood while larger values reduced all coupling terms to near zero. We then calculated the predicted rate for the models, used in log-likelihood estimate (Eq. 47), by averaging repeated simulations of the GLM given the same stimulus.

With respect to the particular example of cell 36, we find that history filter is the same for the coupled and uncoupled cases (Fig. 14A), coupling terms are present on a variety of time scales (Fig. 14B), and the stimulus feature vectors are altered in magnitude by the coupling (Fig. 14C). Interestingly, for all cells in the network, the log-likelihood of the model evaluated for the observed spike train shows overall a negligible difference between the coupled and uncoupled models (Fig. 14D). This is consistent with studies of coupled GLMs applied to retina data (Pillow et al., 2008), in which the addition of coupling terms yields no observable benefit to predicting the *average* rate given the same stimulus.

As seen for the case of the retina and thalamus datasets, more information can be gleaned from the coherence between the predicted rate and the observed spike train. Significant spectral power in

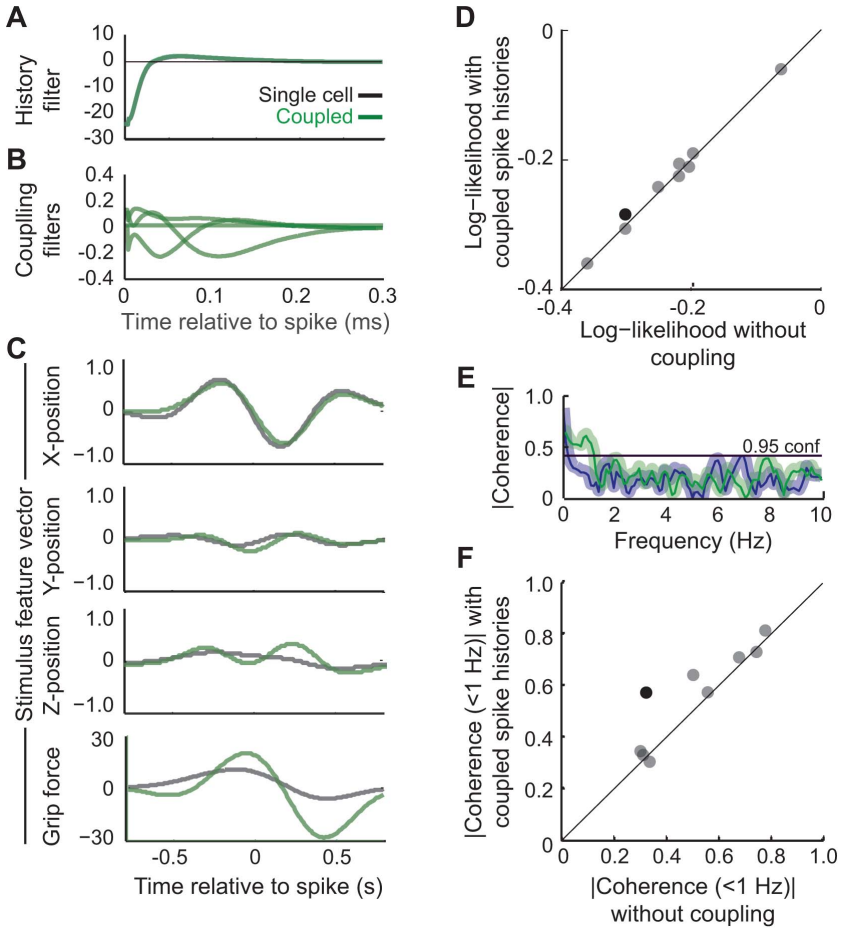

Figure 14: **Network GLM features and validation for monkey reach data using the interval between the start of the Go signal and the end of the trial.** **A** Spike history filter for sample unit 36, one of nine concurrently record units in our analysis. The nine were chosen as those, out of 45 units, with no extra spikes in the refractory period of the inter-spike interval. Green curve shows result for the coupled model (λ = 100 in 55) and black curve shows the filter in the absence of coupling between units; in this case the two curves are indistinguishable. **B** Spike history filters from eight neighboring cells for the coupled model (λ = 100).The coupling terms are nonzero for three neighbors. **C** Feature vectors calculated for the network, i.e., coupled, (green) and single cell, i.e., uncoupled GLM (black). **D** Scatter plot of Log-likelihood between predicted spike rate and observed spike train for the coupled and uncoupled model. The black dot refers to the data in panels A to C. **E**. The spectral coherence, calculated as an average over all trials with a 0.5 Hz bandwidth. The band is one SE. **F** Scatter plot of the coherence between predicted spike rate and observed spike train for the coupled and uncoupled model. The black dot refers to the data in panels A to C and E.

both coupled and uncoupled cases only occurred for low frequencies, i.e., 0 to 1 Hz, and the coupled model had a higher coherence in this range for some cells. This increase was particularly strong for our example cell (Fig. 14E). Thus network interactions through the spike history terms of neighboring cells improve the ability the predict the spike trains for some cells in this dataset.

5.3 Further network GLM methods

A priori, the coupling terms of the network GLM cannot be interpreted as representing direct or anatomical connectivity. Rather, they are best understood as representing functional interactions between the neurons modeled. Such measures of connectivity can still provide insight into anatomical connections in small networks (Gerhard et al., 2013) and population dynamics and encoding in large systems (Stevenson et al., 2012; Chen et al., 2009; Takahashi et al., 2012). In these cases it is useful to quantify the significance of a coupling term between neurons. A common approach is to employ an analysis based on *Granger causality* (Barnett and Seth, 2014). Granger causality is designed to determine when one variable is useful in predicting another (Granger, 1969): if a causal relationship between two processes exists, then the past values of one process should help to predict the future values of the other process. One can apply a variant of Granger causality to the network GLM (Eq. 54) to test the connection from neuron l to neuron i (Kim et al., 2011). More generally, the issue of disambiguating direct interactions from interactions that occur through unobserved, or latent, variables is an important one which is receiving increasing attention. (Pfau et al., 2013; Vidne et al., 2012; Okun et al., 2015).

6 Discussion

We have presented and analyzed a class of methods which summarize the response properties of neurons in terms of one or a few feature vectors and an associated nonlinear input/output function (Fig. 15). These methods provide a principled means to describe neuronal responses. They are a clear improvement over qualitatively described receptive fields, particularly through the inclusion of time dependence that is often suppressed in "classic" receptive field descriptions. However, the approaches we discussed are still phenomenological and it is fair to ask what has been gained.

First, these methods provide a largely automatic and objective means to determine neuronal feature vectors, allowing one to determine how responses "tile" stimulus space. Second, the models are predictive and can be applied to novel stimuli, both as a crucial test of the reliability of the fit of the model as well as a means to estimate the fraction of the cell's response that is modeled by one or a few features. Further, the ability to predict spikes from stimuli will likely play a critical role in neuroprosthetic devices to restore sensation, such as artificial cochleas (Brown and Balkany, 2007), retinas (Trenholm and Roska, 2014; Nirenberg and Pandarinath, 2012), semi-circular canals (Merfeld and Lewis, 2012) and even artificial proprioception (Tabot et al., 2013). Third, the general scheme for all methods can be mapped to feedforward circuitry (Fig. 1A), with the addition of lateral connections between neurons for the network GLM. These circuits may be implemented in terms of Perceptron models. The monotonic nonlinearities found for the STA models resemble those found with spiking neurons (Connors et al., 1982) and the imposed logistic nonlinearity with the MNE model allows for

	STA	STC + STA	MNE	GLM
Number of stimulus feature vectors	One	Unbounded (but typically two or three)	Unbounded	One
History dependence	No			Yes
Network interactions	No		Yes	
Fitting method	Averaging and binning	Matrix diagonalization and binning	Optimization	
Nonlinearity	Derived from Bayes' rule		Fixed as logistic	Fixed as exponential
Binning	Necessary but not problematic	Necessary but problematic for multiple dimensions	Not appropriate	
Convergence on training set	Guaranteed for elliptic distributions of stimuli with a finite second moment		Optimization always converges as fitting is convex	
Over fitting	Not a problem with appropriate binning	Not a problem as nonlinearity is smoothed by lack of data	Potential problem as number of parameters scales as square of stimulus dimension	Potential problem from features and spike history that occur on vastly different time scales
Pioneering publication	Eckhorn & Popel (1981) for averaging. de Ruyter van Steveninck & Bialek (1988) for Bayes' rule.	de Ruyter van Steveninck & Bialek (1988)	Fitzgerald, Sincich & Sharpee (2011) for single cells. Granot-Atedgi, Tkacik, Segev & Schneidman (2013) for networks.	Brown, Frank, Tang, Quirk & Wilson (1998) for single cells. Pillow, Shlens, Paninski, Scher, Litke, Chichilnisky & Simoncelli (2008) for networks

Figure 15: **Synopsis of the methods discussed in this primer.**

high-order features to be directly implemented with biological neurons as well. The parabolic nonlinearities found for some features with the STC model do not have a direct interpretation, yet may be formed by combining pairs of responses. For example, for complex cells in V1 visual cortex, features often appear as pairs that can be combined quadratically (Rust et al., 2005).

6.1 Model assessment

Generally, one would like to measure neural responses to repeated trials, allowing one to estimate the intrinsic variability in the responses and thus bound the expected precision of the model predictions. This results in a observed variance that is a continuous function of time and could be compared to a "model", in this case the observed mean rate; these values, of course, will depend on the smoothing scale applied to the data. Within early stages of the visual pathway, modeling based on repetitive trials capture 80 to 90% of the variance in macaque retina (Pillow et al., 2008), 80 to 90% of the variance in cat primary visual cortex (Touryan et al., 2002), and 94% of the variance in macaque primary visual cortex (Rust et al., 2005).

Here we dealt with the more general case of data that did not have repetitions. We therefore chose to evaluate the accuracy of each model's prediction in two different ways: the log-likelihood (Eq. 47) and the coherence (Eq. 49) between the test spike train and the prediction. The log-likelihood, applied to test data (Figs. 10C and 11C), is a natural choice as it is used as an objective function when fitting the MNE models and the GLM and can be used with spike trains as well as spike rates. However, we observed that it is not always satisfactory. It can, in some instances, be a shallow function that does not clearly discriminate between predictions from models that are rather distinct (Figs. 10C and 11C). Also, as a scalar quantity, the log-likelihood provides no insight into what aspect of the cell's response is or is not captured by the model.

Calculating the coherence between the responses and the predictions offers a complementary approach (Figs. 10E and 11E). Coherence has not been used directly as an objective function for model fitting. In contrast to the log-likelihood, it gives a normalized measure of the portion of the power of

the neuronal responses at a given frequency that is explained by each model. It also indicates timing errors via phase shifts (insert in Fig. 10A and Figs. 10E and 11E). Therefore, it provides information about what aspects of the spike rate are captured by the model and may provide insight into how the model can be improved. Lastly, the normalization allows one to compare results between cells in addition to comparing models of the same cell.

6.2 Caveats on whitening

The pre-whitening procedure for the STA and STC analysis is mathematically sound for random stimuli that have Gaussian statistics (Paninski, 2003) and a limited number of other distributions (Samengo and Gollisch, 2013). Even when this constraint does not strictly hold, our experience (Figs. 7) suggests that, despite no convergence guarantees, a whitened STA or STC plus STA model can give rather good predictions for responses to novel stimuli with natural statistics (Figs. 11). The pre-whitening procedure, however, does not always substantially improve predictions over using the raw stimulus. Since the latter simple approach is easier to construct and less computationally demanding than models specifically tailored for natural scenes, it is worthwhile to construct them and test their predictions.

An intermediate case between natural scenes and Gaussian white noise is when stimuli are drawn from a highly correlated Gaussian distribution, such that the variance along some dimensions is much greater than along others. Here the STC method is guaranteed to converge to the correct set of features, but the large ranges of variances may imply a slow convergence rate. This process can be improved through a modification of the STC method (Aljadeff et al., 2013).

6.3 Adaptation and dependence on stimulus statistics

One significant issue with the fitting of LN models is that the resulting model, including feature, spike history filter and nonlinearity, often depends on the mean, variance, and correlation structure of the stimulus that is used to probe the system. For many sensory systems, the changes that are observed in LN models for different stimulus ensembles (Fairhall, 2013) act to improve information transmission through the system, i.e., account for the presence of noise (Atick, 1992), match the dynamic range of the input/output to the range of stimuli (Brenner et al., 2000; Fairhall et al., 2001; Wark et al., 2007), or cancel out correlations in the input to produce a predictive code (Srinivasan et al., 1982; Hosoya et al., 2005; Sharpee et al., 2006). In some cases, the timescales under which these changes occur suggest that biophysical or circuit properties are altered through long timescale adaptation to different stimulus conditions (Hosoya et al., 2005; Sharpee et al., 2006). However, when the stimulus ensemble is changed abruptly, some corresponding changes in LN models follow close to instantaneously and need not require changes in any biophysical properties of the system (Rudd and Brown, 1997; Fairhall et al., 2001; Mease et al., 2013).

This effect can occur because different stimulus ensembles may drive the system through different parts of its nonlinear regime, and the response behavior is only approximated through the LN model. Thus the best reduced model describing responses for a particular stimulus ensemble will depend on how that ensemble drives the system, even without any changes in the system itself (Gaudry and Reinagel, 2007; Hong et al., 2008; Mease et al., 2014). In some cases these dependencies can be

predicted explicitly (Hong et al., 2008; Famulare and Fairhall, 2010) but more typically are simply empirically observed. The development of models that incorporate these dependencies on stimulus statistics would be of great value and would be able to generalize to a wider range of stimuli. One might have hoped, for example, that the GLM's dependence on the history of activity might take into account issues like spike frequency adaptation and allow one to separate out a common stimulus sensitivity along with a dependence on firing rate that could allow for greater generalization. However, GLMs fit for different stimulus statistics generally differ in all components (Mease et al., 2014) and do not generalize well to different ensembles. It is likely that incorporating features or dynamics acting over multiple timescales can provide sensitivity both to rapid fluctuations and slower-varying statistical properties of the stimulus. For example, a promising current alternative approach is the development of hybrid models that combine an LN model with a dynamical component modeling, for example, activity-dependent changes in kinetic parameters (Baccus and Meister, 2002).

6.4 Population dimensionality reduction

The potential role of correlation in neuronal firing is widely recognized (Cohen and Kohn, 2011). The network GLM is just one approach to deciphering how the activity of many neurons in a fully connected network jointly encodes external inputs/outputs and carries out internal dynamics. More generally, one might expect to be able to represent measured high-dimensional multi-neuronal activity in terms of a smaller number of spatially distributed activation patterns. One approach toward this goal is to project activity patterns into a low-dimensional space and reveal the dynamics occurring during computation (Cunningham and Yu, 2014). A natural starting point to determine this space is to apply PCA to the instantaneous firing patterns (Mazor and Laurent, 2005; Churchland et al., 2010b,a). The method of Gaussian process factor analysis (Yu et al., 2009) further adds some assumptions on the smoothness of the temporal evolution of firing patterns. Given these reduced descriptions of neural activity, typically one then "reverse correlates" on a generally arbitrary or experimenter-defined low-dimensional description of the stimulus or behavior to sort and analyze these patterns according to their external correlates (Churchland et al., 2010b,a). The second strategy aims to systematically model the multineuronal response distribution, $P(r_1, r_2, \ldots, r_n)$, and its correlations using maximum entropy approaches (Schneidman et al., 2006; Ganmor et al., 2011; Fairhall et al., 2012). In this case, one assumes, similar to the MNE approach (Eq. 31), that responses from multiple neurons are jointly distributed according to the maximum entropy distribution that satisfies constraints given by measurements of response correlations. Yet these methods do not yet provide a full input/output mapping. Hybrid maximum entropy models, where the first moment of the distribution depends on the response and the second on network interactions, have also been proposed (Granot-Atedgi et al., 2013).

6.5 Non-spiking data

Our focus has been confined to the relation of spikes, or more generally point processes, to the ongoing stimulus. Yet many neurological events are smoothly varying. At the macroscopic level this includes the flow of current in the extracellular space that is measured by field electrodes or by magnetoencephelography, while at the microscopic level this includes second messenger activation, such

as the intracellular concentration of calcium or cyclic AMP. Measurements of intracellular calcium are of particular importance as the technology to measure such signals with a high signal-to-noise ratio is pervasive throughout neuroscience (Svoboda et al., 1997; Grienberger and Konnerth, 2012) and the onset of the calcium signal can often be taken as a surrogate for an electrical spike (Lütcke et al., 2013). The methods we presented to compute the STA, STC, and MNE features can readily be used to compute feature vectors by replacing the variable for the number of spikes per sample time, $n_{\mathrm{s}}(t)$, by the intensity of the sampled signal (Ramirez et al., 2014). The challenge arises in computing the nonlinearity associated with the STA and STC methods. For the case of spiking, the procedures of spike detection and sorting provide a threshold between no spikes and one or more spikes, although this discrimination process has an associated uncertainty (Lewicki, 1998; Hill et al., 2011b). For an analog process like a change in intracellular calcium, one could simply regard the signal as a continuous signal and choose an appropriate noise model, e.g., Gaussian. Alternatively one can represent it as a point process by selecting a threshold level of detectability. Detection of calcium events, as well as their mapping to spikes, is a topic of ongoing research (Vogelstein et al., 2010).

6.6 Conclusion

We have presented, evaluated, and provided code for a number of methods, all established if not quite mainstream, that answer a simple question: "What makes a neuron fire?". We, along with a plethora of other practitioners, believe that these methods provide a convenient means to obtain insight into the responses of neurons typically obtained in a recording session. In so far as this has proven useful for measurements of single cells, the development of efficient and effective descriptive models becomes a necessity for simultaneous measurements across populations of neurons; thousands if not millions of neurons at once if the hopes for new electrical and optical probes bear out (Alivisatos et al., 2012). As yet, serious limitations apply. When real data does not satisfy certain constraints, such as Gaussian distributed stimulus inputs and monotonic input/output functions, that guarantee convergence for simpler methods, heuristics need to be used to keep fitting procedures from becoming numerically unstable. Even in the retina, LN models often fail to generalize to natural stimuli and do not capture more complex responses such as looming. Responses in neurons that are far downstream from the sensory periphery often have invariances that are very difficult to capture by these methods. In primary visual cortex, LN models have added substantially to the richness of previous descriptions yet leave much unexplained (Olshausen and Field, 2005). Further, real world stimuli may contain critical yet rare stimulus events (Khouri and Nelken, 2015), at least rare on the time-scale of typical physiological recordings. By their very nature, rare stimuli will not be captured by low-order statistics no matter how hard they drive a cell to spike. Despite these caveats, we are optimistic that continuing advances that extend these approaches will likely to become part of the standard canon of electrophysiology as recording techniques progress. But the application of spiking models is still an art form and, like much of electrophysiology (Kleinfeld and Griesbeck, 2005), is not yet an industrial process. *Fortitudine vincimus*.

7 Implementation

All calculations were performed using Matlab (The MathWorks, MA) running on a single processor computer. Annotated code is supplied that was used for all calculations and to generate the figures in the manuscript, along with all datasets, is supplied (download file from http://neurophysics.ucsd.edu/software.php): fifty three salamander retina sets, seven rat thalamic sets, and nine monkey cortex sets. We recommend that interested individuals first repeat the calculations that we used to generate the figures for this paper, then modify the code to analyze their own data.

The following commercial software from The MathWorks (`www.mathworks.com`) is required: Matlab, the Image Processing Toolbox, the Optimization Toolbox, the Signal Processing Toolbox, the Statistics Toolbox, and the Symbolic Math Toolbox. In addition, the following free software must be downloaded: Daniel Hill's code for the Hilbert transform (`neurophysics.ucsd.edu/software.php`), Partha Mitra's Chronux Toolbox (`www.chronux.org`), Jonathan Pillow's Generalized Linear Model (GLM) implementation for spike trains (`http://pillowlab.princeton.edu/code_GLM.html`), Mark Schmidt's L1-norm function `L1GeneralGroup_Auxiliary.m` downloaded at `https://www.cs.ubc.ca/~schmidtm/Software/thesis.html`, and the multi-dimensional histogram function `histcn.m` downloaded at `http://www.mathworks.com/matlabcentral/fileexchange/23897-n-dimensional-histogram/content/histcn.m`.

8 Acknowledgements

This Primer evolved from material presented at the "Methods in Computational Neuroscience" and "Neuroinformatics" summer schools at the Marine Biological Laboratory and the program on "Emerging Techniques in Neuroscience" at the Kavli Institute for Theoretical Physics. We thank Emery N. Brown, Kenneth Latimer, Partha P. Mitra, Jeffrey D. Moore, Rich Pang, Jonathan W. Pillow, Ryan J. Rowekamp and Tatyana O. Sharpee for valuable discussions, Michael J. Berry II and Ronen Segev for making their retina data available, Ben Engelhard and Eilon Vaadia for making their motor cortex data available, and Joel Kaardal for help with the computer code. This effort was supported by grants from the NIH (NS058668 and NS090595 to DK), the NSF (0928251 and EEC-1028725 to ALF and EAGER 2144GA to DK), the Allen Family Foundation (ALF), and the US-Israel Binational Science Foundation (855DBA to DK).

References

Thomas L Adelman, William Bialek, and Robert M Olberg. The information content of receptive fields. *Neuron*, 40(4):823–833, 2003.

Blaise Agüera y Arcas and Adrienne L Fairhall. What causes a neuron to spike? *Neural Computation*, 15(8):1789–1807, 2003.

H Akaike. *Information theory as an extension of the maximum likelihood principle*, chapter International Symposium on Information Theory, pages pp. 267–281. Akademiai Kiado, Budapest, 1973.

A Paul Alivisatos, Miyoung Chun, George M Church, Ralph J Greenspan, Michael L Roukes, and Rafael Yuste. The brain activity map project and the challenge of functional connectomics. *Neuron*, 74(6):970–974, 2012.

Johnatan Aljadeff, Ronen Segev, Michael J Berry II, and Tatyana O Sharpee. Spike Triggered Covariance in Strongly Correlated Gaussian Stimuli. *PLoS computational biology*, 9(9):e1003206, 2013.

Evan W Archer, Urs Koster, Jonathan W Pillow, and Jakob H Macke. Low-dimensional models of neural population activity in sensory cortical circuits. In *Advances in Neural Information Processing Systems*, pages 343–351, 2014.

Joseph J Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

Stephen A Baccus and Markus Meister. Fast and slow contrast adaptation in retinal circuitry. *Neuron*, 36(5):909–919, 2002.

Lionel Barnett and Anil K Seth. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of neuroscience methods*, 223:50–68, 2014.

Michael J Berry and Markus Meister. Refractoriness and neural precision. *The Journal of Neuroscience*, 18(6):2200–2211, 1998.

William Bialek and Rob R van Steveninck. Features and dimensions: Motion estimation in fly vision. *arXiv preprint q-bio/0505003*, 2005.

Aurélie Boisbunon, Stéphane Canu, Dominique Fourdrinier, William Strawderman, and Martin T. Wells. Akaike's information criterion, cp and estimators of loss for elliptically symmetric distributions. *International Statistical Review*, 82(3):422–439, 2014. ISSN 1751-5823. doi: 10.1111/insr.12052. URL `http://dx.doi.org/10.1111/insr.12052`.

Naama Brenner, William Bialek, and Rob de Ruyter Van Steveninck. Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702, 2000.

Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *The Journal of Neuroscience*, 18(18):7411–7425, 1998.

Kevin D Brown and Thomas J Balkany. Benefits of bilateral cochlear implantation: a review. *Current opinion in otolaryngology & head and neck surgery*, 15(5):315–318, 2007.

Zhe Chen, David F Putrino, Demba E Ba, Soumya Ghosh, Riccardo Barbieri, and Emery N Brown. A regularized point process generalized linear model for assessing the functional connectivity in the cat motor cortex. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 5006–5009. IEEE, 2009.

EJ Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, 2001.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Stephen I Ryu, and Krishna V Shenoy. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68(3):387–400, 2010a.

Mark M Churchland, Byron M Yu, John P Cunningham, Leo P Sugrue, Marlene R Cohen, Greg S Corrado, William T Newsome, Andrew M Clark, Paymon Hosseini, and Benjamin B Scott. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3): 369–378, 2010b.

Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811–819, 2011.

BW Connors, MJ Gutnick, and DA Prince. Electrophysiological properties of neocortical neurons in vitro. *Journal of Neurophysiology*, 48(6):1302–1320, 1982.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 2014.

John C Curtis and David Kleinfeld. Phase-to-rate transformations encode touch in cortical neurons of a scanning sensorimotor system. *Nature neuroscience*, 12(4):492–501, 2009.

Marta Dáaz-Quesada and Miguel Maravall. Intrinsic mechanisms for adaptive gain rescaling in barrel cortex. *The Journal of Neuroscience*, 28(3):696–710, 2008.

Stephen V David and Jack L Gallant. Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, 16(2-3):239–260, 2005.

Egbert De Boer and Paul Kuyper. Triggered Correlation. *Biomedical Engineering, IEEE Transactions on*, BME-15(3):169–179, 1968. ISSN 0018-9294.

Reinhard Eckhorn and Bertram Pöpel. Responses of cat retinal ganglion cells to the random motion of a spot stimulus. *Vision research*, 21(4):435–443, 1981.

Ben Engelhard, Nofar Ozeri, Zvi Israel, Hagai Bergman, and Eilon Vaadia. Inducing gamma oscillations and precise spike synchrony by operant conditioning via brain-machine interface. *Neuron*, 77 (2):361–375, 2013.

Adrienne Fairhall. *Adaptation and natural stimulus statistics*, chapter 26, pages 283–293. MIT Press, 2013.

Adrienne Fairhall, Eric Shea-Brown, and Andrea Barreiro. Information theoretic approaches to understanding circuit function. *Current opinion in neurobiology*, 22(4):653–659, 2012.

Adrienne L Fairhall, Geoffrey D Lewen, William Bialek, and Robert R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

Adrienne L Fairhall, C Andrew Burlingame, Ramesh Narasimhan, Robert A Harris, Jason L Puchalla, and Michael J Berry. Selectivity for multiple stimulus features in retinal ganglion cells. *Journal of neurophysiology*, 96(5):2724–2738, 2006.

Michael Famulare and Adrienne Fairhall. Feature selection in simple neurons: how coding depends on spiking dynamics. *Neural computation*, 22(3):581–598, 2010.

Jeffrey D Fitzgerald, Ryan J Rowekamp, Lawrence C Sincich, and Tatyana O Sharpee. Second order dimensionality reduction using minimum and maximum mutual information models. *PLoS computational biology*, 7(10):e1002249, 2011a.

Jeffrey D Fitzgerald, Lawrence C Sincich, and Tatyana O Sharpee. Minimal models of multidimensional computations. *PLoS computational biology*, 7(3):e1001111, 2011b.

Jessica L Fox, Adrienne L Fairhall, and Thomas L Daniel. Encoding properties of haltere neurons enable motion feature detection in a biological gyroscope. *Proceedings of the National Academy of Sciences*, 107(8):3840–3845, 2010.

Elad Ganmor, Ronen Segev, and Elad Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences*, 108(23):9679–9684, 2011.

Kate S Gaudry and Pamela Reinagel. Contrast adaptation in a nonadapting LGN model. *Journal of neurophysiology*, 98(3):1287–1296, 2007.

Felipe Gerhard, Tilman Kispersky, Gabrielle J. Gutierrez, Eve Marder, Mark Kramer, and Uri Eden. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. *PLoS Comput Biol*, 9(7):e1003138–, July 2013. doi: 10.1371/journal.pcbi.1003138. URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.1003138`.

Amir Globerson, Eran Stark, Eilon Vaadia, and Naftali Tishby. The minimum information principle and its application to neural code analysis. *Proceedings of the National Academy of Sciences*, 106(9): 3490–3495, 2009.

Tim Gollisch and Markus Meister. Modeling convergent ON and OFF pathways in the early visual system. *Biological cybernetics*, 99(4-5):263–278, 2008.

D Golomb, D Kleinfeld, RC Reid, RM Shapley, and BI Shraiman. On temporal codes and the spatiotemporal response of neurons in the lateral geniculate nucleus. *Journal of neurophysiology*, 72(6): 2990–3003, 1994.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Einat Granot-Atedgi, Gasper Tkacik, Ronen Segev, and Elad Schneidman. Stimulus-dependent maximum entropy models of neural population codes. *PLoS Comput. Biol*, 9(3):e1002922, 2013.

Christine Grienberger and Arthur Konnerth. Imaging calcium in neurons. *Neuron*, 73(5):862–885, 2012.

Daniel N Hill, John C Curtis, Jeffrey D Moore, and David Kleinfeld. Primary motor cortex reports efferent control of vibrissa motion on multiple timescales. *Neuron*, 72(2):344–356, 2011a.

Daniel N Hill, Samar B Mehta, and David Kleinfeld. Quality metrics to accompany spike sorting of extracellular signals. *The Journal of Neuroscience*, 31(24):8699–8705, 2011b.

Sungho Hong, Brian Nils Lundstrom, and Adrienne L Fairhall. Intrinsic gain modulation and adaptive neural coding. *PLoS Comput Biol*, 4(7):e1000119–e1000119, 2008.

Toshihiko Hosoya, Stephen A Baccus, and Markus Meister. Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77, 2005.

Joel Kaardal, Jeffrey D Fitzgerald, Michael J Berry, and Tatyana O Sharpee. Identifying functional bases for multidimensional neural computations. *Neural computation*, 25(7):1870–1890, 2013.

Leila Khouri and Israel Nelken. Detecting the unexpected. *Current opinion in neurobiology*, 35:142–147, 2015.

Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput Biol*, 7(3):e1001110, 2011.

David Kleinfeld and Oliver Griesbeck. From art to engineering? The rise of in vivo mammalian electrophysiology via genetically targeted labeling and nonlinear imaging. *PLoS biology*, 3(10):1685, 2005.

David Kleinfeld and Partha P Mitra. Applications of spectral methods in functional brain imaging. *Imaging: a laboratory manual*, 1:12.11–12.17, 2011.

David Kleinfeld, Ehud Ahissar, and Mathew E Diamond. Active sensation: insights from the rodent vibrissa sensorimotor system. *Current opinion in neurobiology*, 16(4):435–444, 2006.

Jayant E Kulkarni and Liam Paninski. Common-input models for multiple neural spike-train data. *Network: Computation in Neural Systems*, 18(4):375–407, 2007.

Denis N Lee and H Kalmus. The optic flow field: The foundation of vision [and discussion]. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):169–179, 1980.

Michael S Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, 1998.

Jennifer F Linden, Robert C Liu, Maneesh Sahani, Christoph E Schreiner, and Michael M Merzenich. Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *Journal of Neurophysiology*, 90(4):2660–2675, 2003.

Henry Lütcke, Felipe Gerhard, Friedemann Zenke, Wulfram Gerstner, and Fritjof Helmchen. Inference of neuronal network spike dynamics and topology from calcium imaging data. *Frontiers in neural circuits*, 7, 2013.

Jeffrey C Magee. A prominent role for intrinsic neuronal properties in temporal coding. *Trends in neurosciences*, 26(1):14–16, 2003.

Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.

Miguel Maravall, Rasmus S Petersen, Adrienne L Fairhall, Ehsan Arabzadeh, and Mathew E Diamond. Shifts in coding properties and maintenance of information transmission during adaptation in barrel cortex. *PLoS Biol*, 5(2):e19, 2007.

Panos Z Marmarelis and Vasilis Z Marmarelis. The White-Noise Method in System Identification. In *Analysis of physiological systems*, pages 131–180. Springer, 1978.

Ofer Mazor and Gilles Laurent. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron*, 48(4):661–673, 2005.

James M McFarland, Yuwei Cui, and Daniel A Butts. Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput Biol*, 9(7):e1003143, 2013.

Rebecca A Mease, Michael Famulare, Julijana Gjorgjieva, William J Moody, and Adrienne L Fairhall. Emergence of adaptive computation by single neurons in the developing cortex. *The Journal of Neuroscience*, 33(30):12154–12170, 2013.

Rebecca A Mease, SangWook Lee, Anna T Moritz, Randall K Powers, Marc D Binder, and Adrienne L Fairhall. Context-dependent coding in single neurons. *Journal of computational neuroscience*, 37(3): 459–480, 2014.

Daniel M Merfeld and Richard F Lewis. Replacing semicircular canal function with a vestibular implant. *Current opinion in otolaryngology & head and neck surgery*, 20(5):386–392, 2012.

J. D. Moore, N.M. Lindsay, M. Deschênes, and D. Kleinfeld. Vibrissa self-motion and touch are encoded along the same somatosensory pathway from brainstem through thalamus. *PLoS Biology*, in press, 2015a.

Jeffrey D Moore, Martin Deschênes, and David Kleinfeld. Juxtacellular Monitoring and Localization of Single Neurons within Sub-cortical Brain Structures of Alert, Head-restrained Rats. *Journal of visualized experiments: JoVE*, in press(98), 2015b.

Anirvan S Nandy and Bosco S Tjan. Saccade-confounded image statistics explain visual crowding. *Nature neuroscience*, 15(3):463–469, 2012.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384, 1972.

Mark E Nelson and Malcolm A MacIver. Sensory acquisition in active sensing systems. *Journal of Comparative Physiology A*, 192(6):573–586, 2006.

Sheila Nirenberg and Chethan Pandarinath. Retinal prosthetic strategy with the capacity to restore normal vision. *Proceedings of the National Academy of Sciences*, 109(37):15012–15017, 2012.

Michael Okun, Nicholas A Steinmetz, Lee Cossell, M Florencia Iacaruso, Ho Ko, Péter Barthó, Tirin Moore, Sonja B Hofer, Thomas D Mrsic-Flogel, and Matteo Carandini. Diverse coupling of neurons to populations in sensory cortex. *Nature*, 521(7553):511–515, 2015.

Bruno A Olshausen and David J Field. How close are we to understanding V1? *Neural computation*, 17(8):1665–1699, 2005.

Liam Paninski. Convergence properties of three spike-triggered analysis techniques. *Network: Computation in Neural Systems*, 14(3):437–464, 2003.

Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.

Liam Paninski. The spike-triggered average of the integrate-and-fire cell driven by gaussian white noise. *Neural computation*, 18(11):2592–2616, 2006.

Anitha Pasupathy and Charles E Connor. Population coding of shape in area V4. *Nature neuroscience*, 5(12):1332–1338, 2002.

David Pfau, Eftychios A Pnevmatikakis, and Liam Paninski. Robust learning of low-dimensional dynamics from large neural ensembles. In *Advances in Neural Information Processing Systems*, pages 2391–2399, 2013.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

NF Podvigin, AM Cooperman, and IV Tchueva. The space-time properties of excitation and inhibition and wave processes in cat's corpus geniculatum lateralis. *Biophysics*, 19:341–346, 1974.

RK Powers and MARC D Binder. Experimental evaluation of input-output models of motoneuron discharge. *Journal of neurophysiology*, 75(1):367–379, 1996.

Tony J Prescott, Mathew E Diamond, and Alan M Wing. Active touch sensing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1581):2989–2995, 2011.

Alejandro Ramirez, Eftychios A Pnevmatikakis, Josh Merel, Liam Paninski, Kenneth D Miller, and Randy M Bruno. Spatiotemporal receptive fields of barrel cortex revealed by reverse correlation of synaptic input. *Nature neuroscience*, 2014.

Rajesh PN Rao, Gregory J Zelinsky, Mary M Hayhoe, and Dana H Ballard. Eye movements in iconic visual search. *Vision research*, 42(11):1447–1463, 2002.

Ryan J Rowekamp and Tatyana O Sharpee. Analyzing multicomponent receptive fields from neural responses to natural stimuli. *Network: Computation in Neural Systems*, 22(1-4):45–73, 2011.

Michael E Rudd and Lawrence G Brown. Noise adaptation in integrate-and-fire neurons. *Neural Computation*, 9(5):1047–1069, 1997.

Daniel L Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical review letters*, 73(6):814, 1994.

Nicole C Rust, Odelia Schwartz, J Anthony Movshon, and Eero P Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

Maryam Saleh, Kazutaka Takahashi, and Nicholas G Hatsopoulos. Encoding of coordinated reach and grasp trajectories in primary motor cortex. *The Journal of Neuroscience*, 32(4):1220–1232, 2012.

Inés Samengo and Tim Gollisch. Spike-triggered covariance: geometric proof, symmetry properties, and extension beyond Gaussian stimuli. *Journal of computational neuroscience*, 34(1):137–161, 2013.

Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

Charles E Schroeder, Donald A Wilson, Thomas Radman, Helen Scharfman, and Peter Lakatos. Dynamics of active sensing and perceptual selection. *Current opinion in neurobiology*, 20(2):172–176, 2010.

Odelia Schwartz, Jonathan W Pillow, Nicole C Rust, and Eero P Simoncelli. Spike-triggered neural characterization. *Journal of Vision*, 6(4):13, 2006.

Ronen Segev, Joe Goodhouse, Jason Puchalla, and Michael J Berry. Recording spikes from a large fraction of the ganglion cells in a retinal patch. *Nature neuroscience*, 7(10):1155–1162, 2004.

Ronen Segev, Jason Puchalla, and Michael J Berry. Functional organization of ganglion cells in the salamander retina. *Journal of neurophysiology*, 95(4):2277–2292, 2006.

Tatyana Sharpee, Nicole C Rust, and William Bialek. Maximally informative dimensions: analyzing neural responses to natural signals. *Advances in Neural Information Processing Systems*, pages 277–284, 2003.

Tatyana Sharpee, Nicole C Rust, and William Bialek. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural computation*, 16(2):223–250, 2004.

Tatyana O Sharpee. Computational identification of receptive fields. *Annual review of neuroscience*, 36: 103–120, 2013.

Tatyana O Sharpee, Hiroki Sugihara, Andrei V Kurgansky, Sergei P Rebrik, Michael P Stryker, and Kenneth D Miller. Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439(7079):936–942, 2006.

Shy Shoham, Liam M Paninski, Matthew R Fellows, Nicholas G Hatsopoulos, John P Donoghue, and Richard Normann. Statistical encoding model for a primary motor cortical brain-machine interface. *Biomedical Engineering, IEEE Transactions on*, 52(7):1312–1322, 2005.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Sean J Slee, Matthew H Higgs, Adrienne L Fairhall, and William J Spain. Two-dimensional time coding in the auditory brainstem. *The Journal of Neuroscience*, 25(43):9978–9988, 2005.

Mandyam V Srinivasan, Simon B Laughlin, and Andreas Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B: Biological Sciences*, 216(1205): 427–459, 1982.

Ian H Stevenson, Brian M London, Emily R Oby, Nicholas A Sachs, Jacob Reimer, Bernhard Englitz, Stephen V David, Shihab A Shamma, Timothy J Blanche, and Kenji Mizuseki. Functional connectivity and tuning curves in populations of simultaneously recorded neurons. 2012.

Karel Svoboda, Winfried Denk, David Kleinfeld, and David W Tank. In vivo dendritic calcium dynamics in neocortical pyramidal neurons. *Nature*, 385(6612):161–165, 1997.

Robert G Szulborski and Larry A Palmer. The two-dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. *Vision research*, 30(2):249–254, 1990.

Gregg A Tabot, John F Dammann, Joshua A Berg, Francesco V Tenore, Jessica L Boback, R Jacob Vogelstein, and Sliman J Bensmaia. Restoring the sense of touch with a prosthetic hand through a brain interface. *Proceedings of the National Academy of Sciences*, 110(45):18279–18284, 2013.

Koichi Takahashi, Lorenzo Pesce, Jose Iriarte-Diaz, Matt Best, Sanggyun Kim, Todd P Coleman, Nicholas G Hatsopoulos, and Callum F Ross. Granger causality analysis of state dependent functional connectivity of neurons in orofacial motor cortex during chewing and swallowing. In *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*, pages 1067–1071. IEEE, 2012.

Frédéric E Theunissen, Stephen V David, Nandini C Singh, Anne Hsu, William E Vinje, and Jack L Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12(3):289–316, 2001.

David J Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.

Jon Touryan, Brian Lau, and Yang Dan. Isolation of relevant visual features from random stimuli for cortical complex cells. *The Journal of neuroscience*, 22(24):10811–10818, 2002.

Stuart Trenholm and Botond Roska. Cell-Type-Specific Electric Stimulation for Vision Restoration. *Neuron*, 83(1):1–2, 2014.

Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

J Hans van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.

R De Ruyter Van Steveninck and William Bialek. Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. *Proceedings of the Royal Society of London B: Biological Sciences*, 234(1277):379–414, 1988.

AG Vidal-Gadea and JH Belanger. Muscular anatomy of the legs of the forward walking crab, Libinia emarginata (Decapoda, Brachyura, Majoidea). *Arthropod structure & development*, 38(3):179–194, 2009.

Michael Vidne, Yashar Ahmadian, Jonathon Shlens, Jonathan W Pillow, Jayant Kulkarni, Alan M Litke, EJ Chichilnisky, Eero Simoncelli, and Liam Paninski. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of computational neuroscience*, 33(1): 97–121, 2012.

Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi, Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of neurophysiology*, 104(6):3691–3704, 2010.

Barry Wark, Brian Nils Lundstrom, and Adrienne Fairhall. Sensory adaptation. *Current opinion in neurobiology*, 17(4):423–429, 2007.

Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00532.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x`.