Neuron
Primer

Analysis of Neuronal Spike Trains, Deconstructed

Johnatan Aljadeff,[1,2,*] Benjamin J. Lansdell,[3] Adrienne L. Fairhall,[4,5] and David Kleinfeld[1,6,7,*]
[1]Department of Physics, University of California, San Diego, San Diego, CA 92093, USA
[2]Department of Neurobiology, University of Chicago, Chicago, IL 60637, USA
[3]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA
[4]Department of Physiology and Biophysics, University of Washington, Seattle, WA 98195, USA
[5]WRF UW Institute for Neuroengineering, University of Washington, Seattle, WA 98195, USA
[6]Section of Neurobiology, University of California, San Diego, La Jolla, CA 92093, USA
[7]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA
*Correspondence: aljadeff@uchicago.edu (J.A.), dk@physics.ucsd.edu (D.K.)
http://dx.doi.org/10.1016/j.neuron.2016.05.039

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 and other variables are 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 external variable, as occurs in natural sensory stimuli and in movement. Spectral correlation between predicted and measured spike trains is introduced to contrast the relative success of different methods.

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, as reviewed in 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 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 a description of the statistical nature of this relationship without any direct attempt to establish mechanisms; rather, they provide a compact representation of the components in a stimulus that cause a neuron to fire a spike.

Each of our methods is described by a model that relates the external input to a pattern of spiking (Box 1). A model has several stages (Figure 1A). The first stage includes linear feature vectors that extract a low-dimensional description of the stimulus that drives firing. In spike-triggered average (STA) models, a single feature is extracted from the input. These models have been very successful for neurons in the initial steps of sensory processing, such as retinal ganglion cells (Chichilnisky, 2001; Pillow et al., 2005) in vision or trigeminal cells in vibrissa touch (Jones et al., 2004; Campagner et al., 2016). When a single-feature vector is insufficient to fully describe the firing of the cells, additional features are included. These can be determined in a number of ways, including through spike-triggered covariance (STC) and maximum noise entropy (MNE) methods (Figure 1A). The second stage in these models is a static, nonlinear function that maps the strength of the feature in the time-varying input to an output firing rate; this nonlinear function can, for example, ensure that the predicted spike rate does not go below zero and that it saturates for very large inputs. This succession of linear feature selection followed by nonlinear firing rate prediction means that models of this type are generally known as linear/nonlinear (LN) cascade models. In addition to the stimulus dependence, the so-called generalized linear model (GLM) allows one to incorporate a dependence on the history of firing, as well as the history of firing by other neurons in the network, and potentially other stimulus or task parameters as well (Figure 1A).

The output of these models can be taken to be a time-dependent firing rate. As a final stage, however, one may wish to generate a spike train. To do so, one can assume a specific mathematical process that converts the rate into spikes on a probabilistic basis. This is called a noise model and is generally chosen to be a Poisson or a Bernoulli process, which are described by only a single parameter. In many cases these particular noise models provide a good approximation of spike

Box 1. Glossary of Model Terms

- Dimensionality reduction: in a neuroscience context, dimensionality reduction can be applied to stimulus inputs and to neural responses. Finding a reduced representation for data leads to fewer variables to specify each data point. The reduced representation for responses leads to a restricted range of spiking patterns.
- Feature vector: this is the mathematical representation of a template for a stimulus that is relevant to the neuron's response. When the overlap of the stimulus with this vector is large, i.e., a large inner product, the chance of observing a spike is significantly different from the baseline probability.
- Generalized linear model: a model for the output of a neuron that is a nonlinear function of a sum of inputs linearly filtered by a feature vector and the history of spiking filtered by a history filter.
- History filter: the weights that multiply the recent history of spiking output of the neuron to modulate the future responses.
- Linear/nonlinear model: a class of phenomenological models of neuronal spiking. "Linear" refers to the extraction of stimulus components by linear filtering with feature vectors. "Nonlinear" refers to the static relationship between the filtered stimulus components and the firing rate.
- Maximum entropy model: in a neuroscience context, these are probabilistic models of the associations between inputs and outputs. The probability distribution of producing a spike, given a stimulus as the input, is chosen as the one with the largest entropy subject to a set of predefined constraints that determine the family of associations that one wishes to model.
- Network model: a mathematical description of the joint activity of multiple units. The predicted responses depend on interactions of each neuron with the rest of the network elements.
- Nonlinear input/output function: this relates the filtered stimulus components to the firing rate; it is also called a static nonlinearity.
- Spike-triggered average: the feature vector that results when many examples of sensory inputs that trigger spikes are aligned relative to the spike time and averaged. This procedure is equivalent to so-called reverse correlation.
- Spike-triggered covariance: the covariance matrix formed from stimulus segments that precede a spike. This matrix quantifies how different stimulus components vary together preceding a spike. The eigenvectors of the matrix form a coordinate frame that captures the stimulus' structure that is relevant to predict a spike.

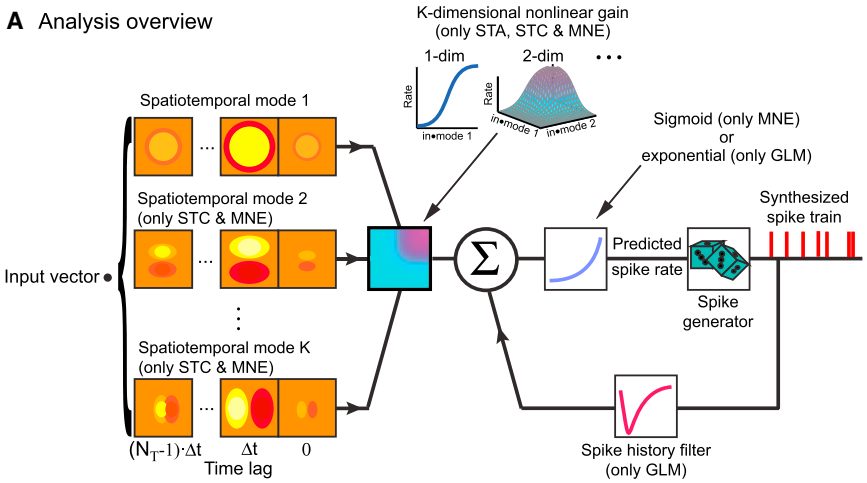
trains recorded from many different areas in the brain (Hagiwara, 1954; Werner and Mountcastle, 1963; Softky and Koch, 1993).

The form of the feature vectors, nonlinearity, and history dependencies can reveal properties of the system that test theoretical concepts, such as how efficiently the stimulus is encoded by a neuron and how robust the encoding is to noise. 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 sensory input (Hosoya et al., 2005; Sharpee et al., 2006). Further, changes in the nonlinearity reveal how the system can modulate its dynamic range as the intensity of the stimulus evolves (Fairhall et al., 2001; Wark et al., 2007; Díaz-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 (Pang et al., 2016). 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. Although we will largely focus on methods to reduce the representation of external variables in order to predict firing, we will further consider a procedure that yields a reduced description of both 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 in that 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., 2015b); 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). Considering data from behaving animals requires 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.

Whenever fitting data, one should avoid fitting structure in the data that is specific to the particular sample chosen as the training set. Further, any identified trends should generalize to other samples from a similar dataset. 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 are used to build the model and 20% for validation. We include methods for model validation and applications to all of our datasets. By permuting the data among the fractions used to fit and to validate, one builds up a so-called jack-knife estimate of the variance for the reliability of the fit.



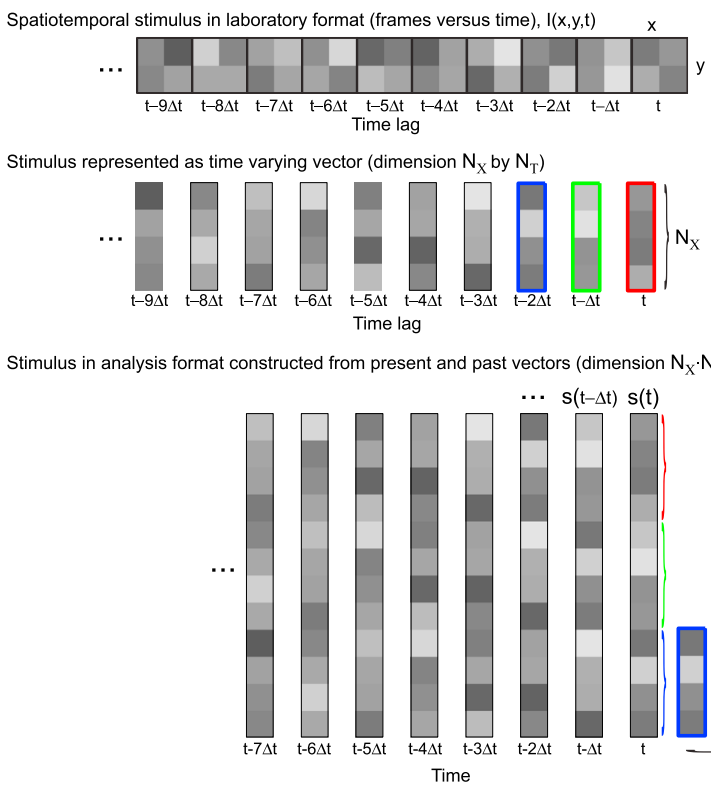


Figure 1. Schematic for the Generation of Spike Trains from Stimuli from Different Classes of Models

(A) An LN model consists of a number of processing steps that transform the input stimulus into a predicted spike train. 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 whose dimensionality is given by the number of features. 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. None of the methods we will consider have all model components; one should choose among the methods depending on the nature of the stimulus, the type of response, and 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.

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. As a brief refresher for the mathematics that we use throughout the Primer, we provide definitions of all essential terms (Box 2) as well as review basic linear algebraic manipulations (Box 3). Lastly, a glossary of all mathematical symbols is provided (Box 4).

The Linear/Nonlinear Modeling Approach

Linear/nonlinear (LN) models 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 partially 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:

$$r(t) = f(r(t' < t), \mathbf{s}(t' < t)). \qquad \text{(Equation 1)}$$

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 spectrotemporal power of a sound, the position of an animal's vibrissae, and so on. The choice of this initial stimulus representation is an important step on its own and could in principle involve a

Box 2. Glossary of Mathematical Terms

- Bayes' rule: in a neuroscience context, it relates the predicted probability of a spike given a stimulus to the measured probability of a stimulus eliciting a spike. More generally, Bayes' rule relates the probability $p(A\,|\,B)$ of event A occurring given that event B has occurred to $p(B\,|\,A)$, the probability of event B occurring given that event A occurred, through $p(B\,|\,A) = p(A\,|\,B)p(B)/p(A)$.
- Coherence: a measure of how two scalar quantities track each other over time, expressed as a function of frequency. Here we use it to assess the relation between predicted and observed spike trains.
- Correlation: a measure of how related two variables are to each other.
- Correlation function: the correlation measured as a function of the lag, e.g., time lag, between two variables. The correlation of a signal with itself is called the autocorrelation; it provides an estimate of how similar a future value is to the current value of a variable.
- Covariance: a measure of how related multiple components of two vector-valued variables are.
- Gradient descent: when searching for the minimum of a function, one can think of the values of this function as a surface. Finding the minimum corresponds to computing the slope, or gradient, of the surface and moving in the direction of the steepest gradient. If the surface is convex, or bowl-like, it is guaranteed that the minimum is global, i.e., there are no points with lower value of the function than the local minimum found by the algorithm.
- Entropy of a probability distribution: this quantity describes the number of states a certain variable can attain and how frequently those states occur. In a neuroscience context, the variable could be the spiking response of a group of neurons; in this example the states represent the patterns of spikes emitted by the population of cells. Entropy increases with the number of states and with the uniformity of the probability of their occurrence. It represents the maximum amount of information a signal can convey about the variable. Specifically, if the probability of event x, drawn from an ensemble of random variables, X, is $p(x)$, then the entropy of that ensemble is $H(X) = \sum p(x)\log_2 p(x)$. When all states, i.e., values of x, are equally likely, the entropy is equal to the log of the number of states.
- Likelihood: the probability of the observed data given the model parameters, understood as a function of the model parameters. It is often convenient to use the log-likelihood because it simplifies the dependence on the model parameters considerably. In maximum likelihood methods, the likelihood is the function being maximized.
- Mutual information: a measure of the co-dependence between two variables that reports how the uncertainty in one variable is reduced by knowing the value of the other. This measure can capture higher-order statistics that correlation and coherence do not.
- Poisson process: a random sequence of events in which the probability of observing an event in a given interval of time is independent of all past and future non-overlapping intervals. For a sufficiently small interval, the probability that a single event occurs is equal to the rate of events times the size of the interval. In a neuroinformatics context, a Poisson spiking process has the property that a spike at any interval in time is independent of previous spiking by the neuron. This implies that the inter-spike interval (ISI) has an exponential distribution when the rate does not change over time.
- Principal component analysis: this is a statistical method to find a set of orthonormal vectors within a space that explains the maximum amount of variance of the data with the fewest vectors.
- Whitening: the procedure by which a signal with arbitrary spectral power is transformed to have a uniform spectral density. For example, a set of variables that have non-zero correlations among themselves will be transformed to a new set that have no correlations. The correlations are measured by a covariance matrix; all cross-correlations among the whitened variables are zero, and the variance of all transformed variables are equal.

nonlinear transformation, e.g., the phase of position 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. The response $r(t)$ is equivalent to the conditional probability of spiking and is of the form

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

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 between the time $t - \Delta t$ and t as $n(t)$.

The LN model is a powerful approach that allows one to approximate the input/output relation (Equation 1) using a plausible amount of experimental data. The key idea is to first find a simplified description of the complex stimulus that captures its relevance to neuronal firing in terms of one or a small number of feature vectors. One then fits the spiking response as a nonlinear function of those few components. Thus the first, "linear" stage of the model acts to reduce the dimensionality of the stimulus. The stimulus $\mathbf{s}(t' < t)$ is in general very high dimensional. For example, a gray-scale image on a 10×10 screen is specified by 100 numbers, which correspond to the light intensity at every location, and these numbers can take any of a range of values. Thus the description of the image is 100-dimensional. If a visual neuron is sensitive, for example, only to the orientation and spatial frequency of the image patch in the center of the screen, this effectively selects a single specific configuration of the 100 values as relevant: all that matters for that neuron's response is how much "oriented bar" there is in the central region of the image. Using linear filtering, one can then take any arbitrary stimulus image, multiply it at every point by the oriented bar configuration, and sum to give a

Box 3. Refresher on Linear Algebra

We use lowercase letters for scalars, i.e., *a* and *b*; we use bold lowercase letters for vectors, i.e., **a** and **b**; and we use bold uppercase letters for matrices, i.e., **A** and **B**. As a refresher in linear algebra, we start with a two-dimensional system, i.e., vectors with two components, to illustrate essential concepts that further apply in high dimension. We define two two-dimensional vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, as

$$\mathbf{v}_1 = \begin{pmatrix} v_1(1) \\ v_1(2) \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \text{ and } \mathbf{v}_2 = \begin{pmatrix} v_2(1) \\ v_2(2) \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

where the numbers in parentheses label the row (Figure B3A).
The inner product of two vectors, denoted by "$\cdot$," is the sum of component-by-component products. Thus

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = 2 \times 0 + -2 \times 2 = -4.$$

The sum of a set of vectors is given by the component-by-component summation. Thus the sum of $\mathbf{v}_1$ and $\mathbf{v}_2$ is (Figure B3A):

$$\mathbf{v}_1 + \mathbf{v}_2 = \begin{pmatrix} 2 \\ -2 \end{pmatrix} + \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

The transpose of a vector, denoted by "$\top$," is found by switching labels of row and columns, so that

$$\mathbf{v}_2^\top = \begin{pmatrix} 0 & 2 \end{pmatrix}.$$

Thus another way to write the inner product of $\mathbf{v}_1$ with $\mathbf{v}_2$ is $\mathbf{v}_1^\top \mathbf{v}_2$, as we multiply a one-by-two vector with a two-by-one vector to get a single number, or scalar. On the other hand, the so-called outer product, $\mathbf{v}_1 \mathbf{v}_2^\top$, multiplies a two-by-one vector with a one-by-two vector to form a two-by-two matrix, i.e.,

$$\mathbf{v}_1 \mathbf{v}_2^\top = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \begin{pmatrix} 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 \times 0 & 2 \times 2 \\ -2 \times 0 & -2 \times 2 \end{pmatrix} = \begin{pmatrix} 0 & 4 \\ 0 & -4 \end{pmatrix}.$$

For any matrix **M**, there exist special vectors **v** that do not change in direction but only in length when they are multiplied by **M**. This is expressed by $\mathbf{M}\mathbf{v} = \lambda \mathbf{v}$, where the matrix **M** is square, i.e., it has the same number of rows as columns. A special but useful class of square matrices have $\mathbf{M} = \mathbf{M}^\top$ and are referred to as symmetric matrices.
For a two-by-two symmetric matrix, the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ and associated eigenvalues λ_1 and λ_2 satisfy $\mathbf{M} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^\top$. The **v** s are orthogonal, i.e., they satisfy $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, and they are normalized to satisfy $\mathbf{v}_1 \cdot \mathbf{v}_1 = \mathbf{v}_2 \cdot \mathbf{v}_2 = 1$. How do we find the **v** s? For concreteness, we consider a matrix **M** defined by

$$\mathbf{M} \equiv \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix},$$

where α is a scalar. This satisfies the eigenvalue equation $\mathbf{M}\mathbf{v} = \lambda \mathbf{v}$, or

$$\begin{pmatrix} 1-\alpha-\lambda & \alpha \\ \alpha & 1-\alpha-\lambda \end{pmatrix} v = 0.$$

There is a so-called trivial solution, $v = 0$, as well as two non-zero eigenvectors. The latter are found by setting the determinant, denoted $|\cdots|$, to zero, i.e.,

$$\begin{vmatrix} 1-\alpha-\lambda & \alpha \\ \alpha & 1-\alpha-\lambda \end{vmatrix} = (1-\alpha-\lambda)^2 - \alpha^2 = (\lambda - 1)(\lambda - 1 + 2\alpha) = 0,$$

(Continued on next page)

Box 3. *Continued*

and the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 1 - 2\alpha$. The corresponding eigenvectors are found from substitution plus normalization and are

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ and } \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

respectively (Figure B3B). The dominant eigenvector, $\mathbf{v}_1$, points in the direction of equal variation of $\mathbf{v}_1(1)$ with $\mathbf{v}_1(2)$.
A final issue is that any symmetric matrix with non-zero eigenvalues has an inverse, denoted $\mathbf{M}^{-1}$, such that the product is equal to the identity matrix, i.e., $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$, where $\mathbf{I}$ has ones along the diagonal and zeros everywhere else. Thus all of the eigenvalues of the identity matrix are one. For the above example,

$$\mathbf{M}^{-1} = \frac{1}{1 - 2\alpha}\begin{pmatrix} 1 - \alpha & -\alpha \\ -\alpha & 1 - \alpha \end{pmatrix},$$

and the eigenvalues of $\mathbf{M}^{-1}$ are λ_1^{-1} and λ_2^{-1}.

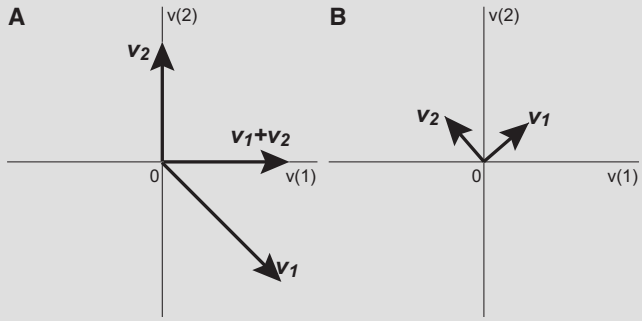


Figure B3. Two-Dimensional Vector Plots
(A) Example vectors.
(B) Eigenvectors of our example matrix **M**. Each has unit length.

single number that quantifies the presence of the relevant feature in the image.

More formally, the relation $f(\cdot)$ in Equation 1 is divided into two parts: a linear and a nonlinear stage. In general, the stimulus 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 (Figure 1), there is some timescale $T = N_T \Delta t$ 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

$$N = N_X \times N_T. \qquad \text{(Equation 3)}$$

This full N-dimensional stimulus is processed by a set of linear filters defined by the feature vectors. These filters act to extract certain components, i.e., linear combinations or dimensions of the stimulus, and possibly also the spike history. Next, a nonlinear stage, which we will denote as $g(\cdot)$, 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, K, that are relevant to the neuron's response is much less than the maximum stimulus dimensionality, N. All methods thus necessarily include a dimensionality reduction step whose goal is to find these relevant K vectors that we will call features. They are denoted by ϕ_i, each of size N with $i = 1, \ldots, K$. The input/output transformation can be written in terms of the ϕ_i as

$$r(t) = g(z_1, z_2, \ldots, z_K, r(t' < t)) \qquad \text{(Equation 4)}$$

where

$$z_i = z_i(t) \equiv \phi_i \cdot \mathbf{s}(t' < t) \qquad \text{(Equation 5)}$$

is the projection of the stimulus on the i^{th} feature, i.e., a component-wise multiplication of the stimulus and the feature, followed by summation. The time-invariant features ϕ_i are vectors that 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. More complicated projections could in principle be used, but the determination of such projections typically will require more data when the z_i are nonlinear functions of the stimulus.

The second assumption is that the nonlinear stage is taken to be stationary in time, i.e., $g(\cdot)$ has time dependence only through

Box 4. Glossary of Symbols

- N, number of stimulus dimensions
- N_X, number of spatial (or spectral) points in stimulus
- N_T, number of time points in stimulus
- $r(t)$, predicted firing rate
- $\mathbf{s}(t)$, stimulus presented at time t ($N\times1$ vector)
- ϕ_i, the i-th feature vector, $i=1,\ldots,K$ ($N\times1$ vector)
- K, number of feature vectors in LN model
- z_i, projection of stimulus onto i-th feature vector
- $g(z_1, \cdots z_K)$, nonlinearity specified as a function of the stimulus projections z_i
- t_b, set of all times that the reduced stimulus belongs to bin b
- ϕ_{sta}, spike-triggered average ($N\times1$ vector)
- $\bar{\mathbf{s}}$, average stimulus ($N\times1$ vector)
- M, number of stimuli
- $n(t)$, number of spikes at time t
- n_T, total number of spikes in time series
- $\mathbf{C}_{\text{p}}$, underlying stimulus covariance ($N\times N$ matrix)
- $\mathbf{C}_{\text{s}}$, spike-triggered stimulus covariance (STC) ($N\times N$ matrix)
- $\mathbf{C}_{\text{r}}$, spike-triggered stimulus covariance conditioned on randomly shifted spike trains ($N\times N$ matrix)
- $\Delta\mathbf{C}$, matrix of covariance differences ($N\times N$ matrix)
- $\Delta\mathbf{C}_{\text{r}}$, matrix of covariance differences with respect to randomly shifted spike trains ($N\times N$ matrix)
- λ_i, the i-th eigenvalue of a matrix
- $\mathbf{u}_i, \mathbf{v}_i$, the i-th eigenvector of a matrix
- $\mathbf{I}$, identity matrix
- $\phi_{\text{stc},i}$, the i-th STC eigenvector ($N\times1$ vector)
- $\widehat{\phi}_{\text{sta}}$ and $\widehat{\phi}_{\text{stc},i}$, decorrelated STA and STC feature vectors ($N\times1$ vectors)
- $\mathbf{C}_{\text{p},L}^{-1}$, pseudo-inverse of rank L of underlying stimulus covariance
- $a, \mathbf{h}, \mathbf{J}$, parameters of maximum noise entropy models that satisfy zeroth, first, and second order constraints, respectively
- r_0, mean spike rate
- c, parameter of Generalized Linear Model (GLM)
- ϕ_{glm}, feature vector of GLM ($N\times1$ vector)
- $\boldsymbol{\psi}$, spike history filter of single-neuron GLM
- κ, regularization penalty in GLM
- w_{i}, raised cosine basis function for spike history filter
- B, number of functions included in raised cosine basis
- t_0, t_1, t_2, parameters of raised cosine basis
- $\boldsymbol{\psi}_{ij}$, spike history filter, from neuron j to neuron i, of multi-neuron GLM
- $\mathcal{L}$, likelihood function
- $\tilde{C}(f)$, spectral coherence at frequency f
- Φ, phase in whisk cycle

the stimulus and the history of the neuron's spiking response that, like the stimulus, may be high dimensional. The nonlinearity $g(\cdot)$ can be determined nonparametrically using the probabilistic interpretation of Equation 1 given in Equation 2. We consider for now only dependence on the stimulus and not on the history of spiking, i.e., $r(t)=g(z_1,z_2,\ldots,z_K)$.

A general means to estimate the nonlinearity is to determine the expectation of the response within each stimulus bin (Chichilnisky, 2001). Given $z_i(t)$, i.e., the projections of the stimulus on the K relevant feature vectors (Equation 5), the function $g(z_1,\ldots,z_K)$ can be computed by first discretizing each of the K stimulus components into N_B bins. The resolution at which the nonlinearity is estimated is limited by the need to ensure that each of the $(N_B)^K$ bins contains multiple data points for a statistically robust result. Then the measured response rate, $r(t)$, is averaged over all the time points where the stimulus belongs to each bin. For concreteness, suppose that $z_{1,b},\ldots,z_{K,b}$ is the point in the middle of the bin b, and the lower and upper boundaries defining that bin are $z^l_{1,b},\ldots,z^l_{K,b}$ and $z^u_{1,b},\ldots,z^u_{K,b}$, respectively. The value of $g(z_{1,b},\ldots,z_{K,b})$ will be set to

$$g(z_1,z_2,\ldots,z_K)=\frac{1}{T_b}\sum_{t_b} r(t_b), \qquad \text{(Equation 6)}$$

where the t_b comprise the set of all times for which the stimulus belongs to the bin b, i.e.,

$$z^l_{1,b}\le z_1(t_b)<z^u_{1,b},\ldots,z^l_{K,b}\le z_K(t_b)<z^u_{K,b}, \qquad \text{(Equation 7)}$$

and T_b is the number of such samples in the data. Finally, the value of $g(z_1,\ldots,z_K)$ for all points is found by interpolating between the values at the center points for each of the bins.

Box 5. Nonlinearity for Neuronal Responses

We discussed two alternatives to compute the nonlinearity $g(\cdot)$ in the main text. Here we show that these are equivalent if the response is binary. In the more general method (Equations 6 and 7), $g(\cdot)$ is set to the average response that corresponds to stimuli in each specific bin, i.e.,

$$g(z_{1,b}, \ldots, z_{K,b}) = \frac{1}{T_b} \sum_{t_b} r(t_b).$$

Assume that the response is binary and that each bin is sufficiently well sampled. Then average response is equal to the probability of a spike given that the stimulus belongs to a certain bin b, i.e.,

$$g(z_{1,b}, \ldots, z_{K,b}) = p(\text{spike} \mid \mathbf{s} \text{ in bin } b) = p(\text{spike} \mid z_{1,b}, \ldots, z_{K,b}).$$

With the use of Bayes' rule, this becomes

$$g(z_{1,b}, \ldots, z_{K,b}) = \frac{p(z_{1,b}, \ldots, z_{K,b} \mid \text{spike}) p(\text{spike})}{p(z_{1,b}, \ldots, z_{K,b})},$$

and we recover Equation 9.

A more intuitive approach to find the nonlinear transformation, $g(\cdot)$, applies for the particular case of only binary responses, i.e., spike or no-spike per sample interval Δt. This is always possible for spike trains sampled at a sufficiently high rate. Here we can use Bayes' rule, i.e.,

$$p(\text{spike} \mid \mathbf{s}(t)) = \frac{p(\mathbf{s}(t) \mid \text{spike})\, p(\text{spike})}{p(\mathbf{s}(t))}, \qquad \text{(Equation 8)}$$

to determine an input/output relation in terms of the reduced variables defined above (Equation 5):

$$g(z_1, z_2, \ldots, z_K) = p(\text{spike} \mid z_1, z_2, \ldots, z_K)$$

$$= \frac{p(z_1, z_2, \ldots, z_K \mid \text{spike})\, p(\text{spike})}{p(z_1, z_2, \ldots, z_K)}. \qquad \text{(Equation 9)}$$

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

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

The underlying stimulus distribution and the spike-conditional distribution are estimated by binning the K-dimensional stimulus subspace along each of the feature vectors that span this space. The Bayesian procedure (Equations 8 and 9) can be deduced as a special case of the expectation rule (Equations 6 and 7) (Box 5).

It is often useful to examine the dependence of the input/output relation $g(z_1, z_2, \ldots, z_K)$ as a function of only one variable. This is referred to as the marginal gain and is found by integrating over all other variables. For example, the marginal of z_1 is given by

$$g(z_1) = \int dz_2 \cdots dz_K\, g(z_1, z_2, \ldots, z_K). \qquad \text{(Equation 10)}$$

Relation to Expansion Modeling

It is worthwhile to briefly contrast the LN approach with traditional nonlinear methods. 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 Naka, 1972; Marmarelis and Marmarelis, 1978), i.e.,

$$r(t) = f(\mathbf{s}(t' < t)) = \int_{t' < t} dt'\, f_1(t')\, \mathbf{s}(t - t') + \int_{t' < t} dt' \int_{t'' < t} dt''\, f_2(t', t'')\, \mathbf{s}(t - t')\, \mathbf{s}(t - t'') + \ldots, \qquad \text{(Equation 11)}$$

where the functions $f_1(\cdot)$, $f_2(\cdot)$, etc. are weighting functions called 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). However, 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 approach differs in that no attempt is made to approximate the nonlinearity in

successive orders. Rather, the nonlinearity is explicitly introduced as a component of the model.

Nonparametric Models

An important goal of this type of approach is to drive the system with a wide variety of inputs so that one explores, and the model captures, as much richness in the response as possible. One approach to this is to stimulate with white noise, an input that samples a wide space of possibilities. However, one should bear in mind that that this is not a natural input for most sensory systems and might drive the system in ways that rarely occur in nature or put it into an unusual state of adaptation. While these possibilities raise interesting issues for future study, the results of noise stimuli often give strong clues as to the realm of structured stimuli that are relevant.

For white noise, the value of the stimulus at one location or time is unrelated to its value at any other location or time—that is, there are no correlations in the input. This means that all frequencies are represented in the stimulus up to a smoothing cutoff, which might be determined by limitations on how the stimulus is produced, or chosen using a reasonable guess at the fastest possible response timescale of the system. 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 (Figure 3B). 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.,

$$\overset{N_T \text{ time points}}{\longleftrightarrow} \quad \left.\begin{pmatrix} I(1,t-(N_T-1)) & \cdots & I(1,t) \\ \vdots & \ddots & \vdots \\ I(N_X,t-(N_T-1)) & \cdots & I(N_X,t) \end{pmatrix}\right\updownarrow N_x \text{ spatial positions} \qquad \text{(Equation 12)}$$

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 $N = N_X \times N_T$ (Equation 3) vector that indexes the N_T frames that go back in time by $N_T \Delta t$ (Figure 3B):

$$\mathbf{s}(t) = \begin{pmatrix} I(1,t-(N_T-1)) \\ \vdots \\ I(N_X,t-(N_T-1)) \\ \vdots \\ I(1,t) \\ \vdots \\ I(N_X,t) \end{pmatrix}. \qquad \text{(Equation 13)}$$

Spike-Triggered Average

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 that is set by the stimulus design (all dots in Figure 2A). The spike-triggering stimuli are a smaller cloud, or subset, of these stimuli (red dots in Figure 2A). Dimensionality reduction seeks to find the stimulus subspace that 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 equals 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 (STA), denoted ϕ_{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.,

$$\phi_{\text{sta}} = \frac{1}{n_{\text{T}}} \sum_t n(t)(\mathbf{s}(t) - \bar{\mathbf{s}}), \qquad \text{(Equation 14)}$$

where $n(t)$ is the number of spikes at time t, n_{T} is the total number of spikes, $\bar{\mathbf{s}}$ is the average stimulus, i.e.,

$$\bar{\mathbf{s}} = \frac{1}{M} \sum_t \mathbf{s}(t), \qquad \text{(Equation 15)}$$

and M is the total number of stimuli presented in the experiment. As for the case of the stimuli (Figure 1B), the STA is organized as a vector of length N that indexes the N_T frames back in time from $t = 0$ to $t = (N_T - 1)\Delta t$ (Equation 3), i.e.,

$$\phi_{\text{sta}} = \begin{pmatrix} \phi_{\text{sta}}[1] \\ \vdots \\ \phi_{\text{sta}}[N_X] \\ \vdots \\ \phi_{\text{sta}}[2N_X] \\ \vdots \\ \phi_{\text{sta}}[N_T \times N_X] \end{pmatrix}. \qquad \text{(Equation 16)}$$

For a Gaussian stimulus, the underlying distribution of stimulus values projected onto the STA, $p(z) = p(\phi_{\text{sta}} \cdot \mathbf{s}(t))$, is also Gaussian. Often in experiments, however, the stimulus is binary, so that the stimulus in each pixel or time point takes one of two values. If the stimulus has a large number of components, the central limit theorem ensures that these projections, as a sum of many random values weighted by the feature vector components, will have a Gaussian distribution. This distribution can be either computed analytically from the statistics used to construct the stimuli or accurately fit from data.

The nonlinearity can be estimated as the expectation of the response (Equations 6 and 7) or, when the sampling interval is sufficiently fine, with the use of Bayes' rule (Equations 8 and 9). The conditional histogram defining $p(z|\text{spike}) = p(\phi_{\text{sta}} \cdot \mathbf{s}(t)|\text{spike})$ 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 apply a smoothing spline to the conditional

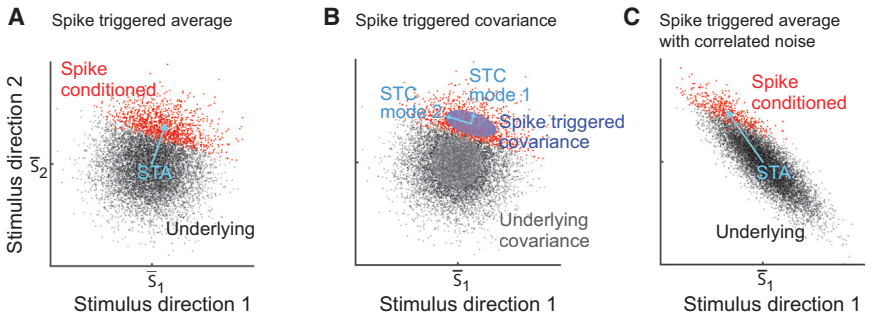


Figure 2. Schematic of Stimulus Samples Plotted in Two Arbitrary Directions in Stimulus Space

(A–C) Each stimulus is plotted as a single dot, its projection into two dimensions. The red-dotted stimuli precede spikes.

(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 underlying stimulus covariance $\mathbf{C}_p$ (solid gray circle), forms one set of vectors, and the covariance of the spike-triggered stimuli, $\mathbf{C}_s$, forms a second set. The two dominant vectors comprising their difference, i.e., $\Delta\mathbf{C} = \mathbf{C}_s - \mathbf{C}_p$, yield the dominant STC two modes. The mode is significant only if its length is larger than the radius of the underlying stimulus covariance.

(C) The naively computed STA for the case of correlated or colored noise, where the variance of the underlying stimulus distribution is heterogeneous.

distribution or reduce the number of bins used to estimate the distributions from the data.

Calculating the STA for Retinal Ganglion Cells. We consider the case of a binary checkerboard stimulus used to drive spiking in retinal ganglion cells (Figure 3), the neurons that output visual information from the retina. In this experiment, the pixel values were chosen from a binary distribution (Figure 3A). We applied the above formalism to the stimulus set reorganized as three consecutive frames for a stimulus dimension of N = 14^2 × 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 Equations 14 and 15 for each of 53 retinal ganglion units; an example is shown in Figure 4. The aim is to choose the stimulus dimensionality, i.e., $N_T \times N_X$, so that any structure in the STA returns to zero at the boundaries. Further, the dimension of the stimulus should be sufficiently large to allow resolution of structure within the STA yet small enough to allow sufficient averaging over the effects of noise.

Here we tried both $N_T = 3$ time points with a larger patch size of 14×14 pixels (Figures 4A and 4B) and $N_T = 6$ frames with a smaller patch of 10×10 pixels (Figures 4C and 4D). The key feature is a central spot of excitation that rises and falls over three frames (Figures 4A and 4B). 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 underlying stimulus distribution to be well sampled (Figure 4C). This distribution is consistent with a Gaussian, as can be expected for a projection on any direction for a white noise stimulus (Figures 4B and 4D). The coarse time bins contained up to three spikes per bin, and thus we used the expectation rule to calculate the nonlinearity (Equations 6 and 7). The observed nonlinearity is found to be monotonic (Figure 4D).

Interpreting the STA. The STA procedure (Equations 14 and 15) has a strong theoretical basis. It has been shown (Chichilnisky, 2001; Paninski, 2003) that ϕ_{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 amounts of data and an elliptical noise distribution (Paninski, 2003), the STA feature is guaranteed to correctly recover the dependence of the neuron's response on this single feature. Geometrically, the vector ϕ_{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 ϕ_{sta}. For example, the spike-triggering cloud of stimuli points might be symmetric, such that the average lies at the origin, but the shape is nonetheless very different from the cloud consisting of all stimuli, i.e., the underlying stimulus distribution $p(\phi_{sta} \cdot \mathbf{s}(t))$. The spike-triggered covariance (STC), discussed next, makes use of this additional information.

Spike-Triggered Covariance

While generally the STA 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 ϕ_{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:

$$p(\text{spike} \mid \mathbf{s}(t)) = p(z_1, z_2, \ldots, z_K), \qquad \text{(Equation 17)}$$

where, as a reminder, $z_i = z_i(t) = \phi_i \cdot \mathbf{s}(t)$ is the projection of the stimulus at time *t* on the *i*-th identified feature vector ϕ_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 second-order statistics of the stimulus itself. These are captured by its covariance matrix, also referred to as the underlying stimulus covariance, i.e.,

$$\mathbf{C}_p = \frac{1}{M-1} \sum_t (\mathbf{s}(t) - \bar{\mathbf{s}})(\mathbf{s}(t) - \bar{\mathbf{s}})^\top, \qquad \text{(Equation 18)}$$

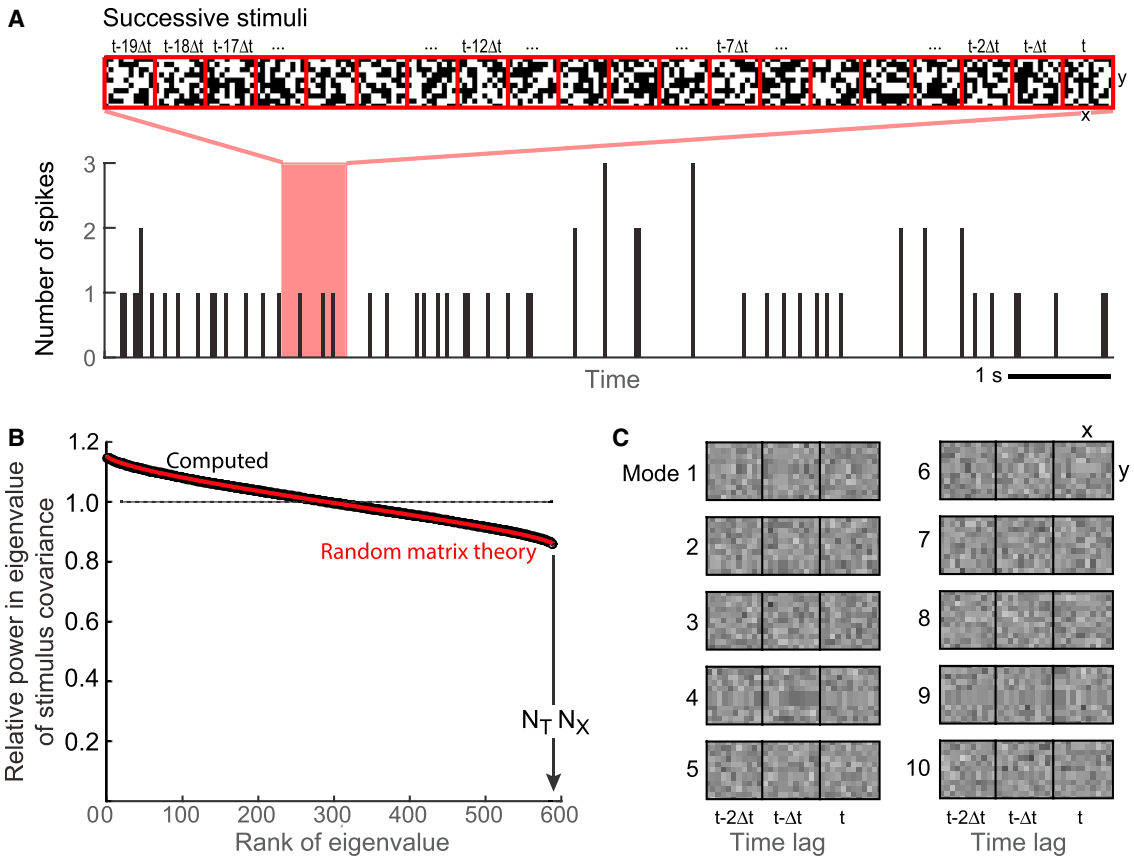


Figure 3. Spike Responses from Salamander Retinal Ganglion Cell 3 for a Visual Checkerboard Stimulus, Used to Illustrate the Methods with a "White Noise" Stimulus

(A) Each pixel in the checkerboard was refreshed each Δt = 33.33 ms with a random value, and the spikes were recorded within the same interval.

(B) We constructed the covariance matrix of the stimulus (Equation 18) and plotted its spectrum (black).The eigenvalues are all close to the variance of a single pixel, $\sigma^2 = 1$, for the checkerboard stimulus. We compared this spectrum to that expected theoretically for the same-sized random matrix (Marčenko and Pastur, 1967) with signal-to-noise parameter $\gamma = n/N$ (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 eigenvectors of the stimulus covariance matrix (Equation 18) that correspond to the largest eigenvalues are seen to contain no spatial or temporal structure.

Methods: The dataset consists of 53 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 $40^2 = 1600$ 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 spatial receptive field of each ganglion cell, with a cut-off of 200 cm^{-1} in spatial frequency. For each cell, we extracted either the $14^2 = 196$ or the $10^2 = 100$ pixel region with modulated activity; these give rise to 2^{196} or 2^{100} potential patterns, respectively. Each time series was 60–120 min long and contained between 1,000 and 10,000 spikes, but sampled a tiny fraction of the potential patterns.

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.,

$$\mathbf{C}_p = \sum_{i=1}^{K} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\top}, \qquad \text{(Equation 19)}$$

where the eigenvectors of $\mathbf{C}_p$ 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 underlying stimulus distribution are equal and $\mathbf{C}_p$ is a diagonal matrix with $\mathbf{C}_p = \sigma^2\mathbf{I}$, 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 underlying stimulus covariance (Equation 19) affects the spectrum slightly, but in a predictable way; the spectrum of eigenvalues of the stimulus covariance matrix is close

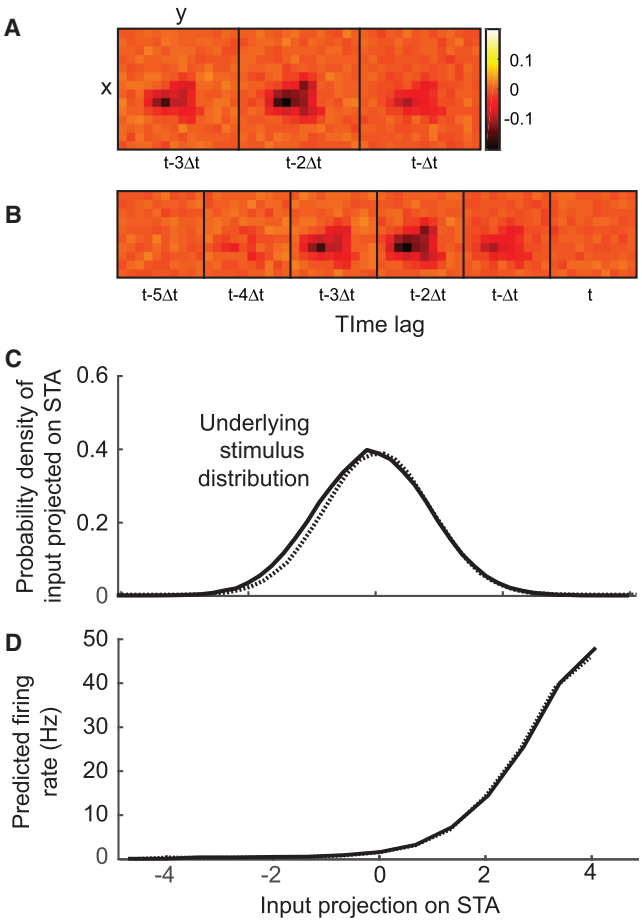


Figure 4. The Spike-Triggered Average, ϕ_{sta}, for the Responses of Retinal Ganglion Cell 3

(A and B) We considered two stimulus representations. In (A), we show a short sequence where we retain three stimulus frames in the past ($N_T = 3$), and the frame was $N_X = 14^2 = 196$ pixels. We chose the optimal lag for which the cell's response is maximally modulated by the stimulus. In (B), we show a long sequence where $N_T = 6$, but the frame was cropped such that $N_X = 10 \times 10 = 100$. We chose the first six frames into the past.

(C) The underlying stimulus distribution computed for both representations, (A) and (B), in solid and dashed curves, respectively.

(D) The expectation procedure (Equations 6 and 7) was used to obtain the input/output nonlinearity for both representations, (A) and (B), in solid and dashed curves, respectively.

to constant (black dots in Figure 3B), in agreement with the analytical spectrum calculated for the same stimulus dimension and same number of samples (red dots in Figure 3B). Although we could have computed the underlying stimulus distribution without finite size limitations, it is instructive to see this effect. The dominant eigenvectors, shown in a space-time format, appear featureless, as they should (Figure 3C).

Our goal is to find the directions in stimulus space in which the variances of the spike-triggering stimuli differ relative to the underlying stimulus distribution of stimuli. These can be found through the covariance difference matrix (de Ruyter 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

$$\mathbf{\Delta C} = \mathbf{C}_s - \mathbf{C}_p, \quad \text{(Equation 20)}$$

and the STC matrix $\mathbf{C}_s$ is computed relative to the spike-triggered average (Equations 14 and 15) and given by

$$\mathbf{C}_s = \frac{1}{n_T - 1} \sum_t n(t)(\mathbf{s}(t) - \phi_{sta})(\mathbf{s}(t) - \phi_{sta})^\top. \quad \text{(Equation 21)}$$

The underlying stimulus covariance matrix $\mathbf{C}_p$ is given by Equation 18, and we recall that n_T is the total number of spikes.

The matrix $\mathbf{\Delta C}$ (Equation 20) may be expanded in terms of its eigenvalues, λ_i, and eigenvectors, $\phi_{stc,i}$, i.e.,

$$\mathbf{\Delta C} = \sum_{i=1}^{N} \lambda_i \phi_{stc,i} \phi_{stc,i}^\top. \quad \text{(Equation 22)}$$

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

Practical Considerations in Computing the STC. Some eigenvalues will emerge from the background simply because of $\sqrt{N}$ noise from the finite number of samples. 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 at random from the same stimulus. We compute, for a large number of repetitions, a STC matrix using randomly chosen spike times, t_r, to select the same number of stimulus samples at random, i.e.,

$$\mathbf{C}_r = \frac{1}{n_T - 1} \sum_{t_r} n(t_r)(\mathbf{s}(t_r) - \phi_{sta})(\mathbf{s}(t_r) - \phi_{sta})^\top. \quad \text{(Equation 23)}$$

The corresponding matrix of covariance differences $\mathbf{\Delta C}_r = \mathbf{C}_r - \mathbf{C}_p$ and its eigenvalues are computed for each random choice. The eigenvalues of all matrices $\mathbf{\Delta C}_r$ form a so-called "null distribution." Eigenvalues of $\mathbf{\Delta C}$ (Equation 20) 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 that results from temporal correlations in the spike train, e.g., a tendency to spike in bursts. If such structure exists, one can compute the matrix $\mathbf{C}_r$ (Equation 23) 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, $\phi_{stc,i}$, are the corresponding significant eigenvectors of the covariance difference matrix. If there is a non-zero STA, ϕ_{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, it is convenient to 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

$$\phi^{\perp}_{\mathrm{stc},i} = \phi_{\mathrm{stc},i} - \frac{\phi_{\mathrm{stc},i} \cdot \phi_{\mathrm{sta}}}{\|\phi_{\mathrm{sta}}\|^2}\phi_{\mathrm{sta}},\ i = 1, \ldots K-1. \qquad \text{(Equation 24)}$$

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

For the case of white noise, where the variance of the stimulus is equal along every direction, the eigenvalues of the underlying stimulus 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 is the case for correlated noise, the significance threshold for each eigenvalue of $\mathbf{C}_{\mathrm{s}}$ is different. In this case, subtracting the underlying stimulus covariance allows one to test whether the variance of the spike-triggered distribution is different from that of the underlying stimulus distribution along each direction.

The STA and the set of orthogonalized STC vectors are then used to calculate 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 either the expectation (Equations 6 and 7) or Bayesian (Equations 8 and 9) procedure. 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.

Calculating the STC for Retinal Ganglion Cells. The STC features were computed according to Equation 20 for a set of retinal ganglion units; results for the same representative unit used for the STA features (Figures 4 and 5A) are shown in Figure 5. There are four STC features (Figure 5A) that are statistically significant (Figure 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 spike.

We complete the model by calculating the nonlinearity (Equations 6 and 7). We first project out the component along the STA feature from the STC features (Equation 24) 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., $p(\mathrm{spike}\,|\,\phi_{\mathrm{sta}} \cdot \mathbf{s}, \phi^{\perp}_{\mathrm{stc},2} \cdot \mathbf{s})$ (Figure 5C). The one-dimensional nonlinearities for the STA and orthogonal STC mode can be recovered from this function by projecting along the respective axes (Equation 10) (Figure 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, for example, if the neuron is sensitive to a feature independent of its sign.

Interpreting the STC. 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 underlying 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 retinal ganglion cell (Fairhall et al., 2006; Gollisch and Meister, 2008) 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).

The spectral decomposition of the symmetric matrix $\mathbf{\Delta C}$ always returns orthogonal components. Thus 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. Finding the appropriate rotations from the orthogonal feature vectors to stimulus components can strengthen the potential link between the functional model and underlying properties of the neural circuit, but requires additional assumptions. This is, in general, a difficult problem (Hong et al., 2008; Kaardal et al., 2013; Ramirez et al., 2014).

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 (Nelson and MacIver, 2006; Kleinfeld et al., 2006; Schroeder et al., 2010; Prescott et al., 2011). While one could, for example, sample the natural scene statistics of a forest

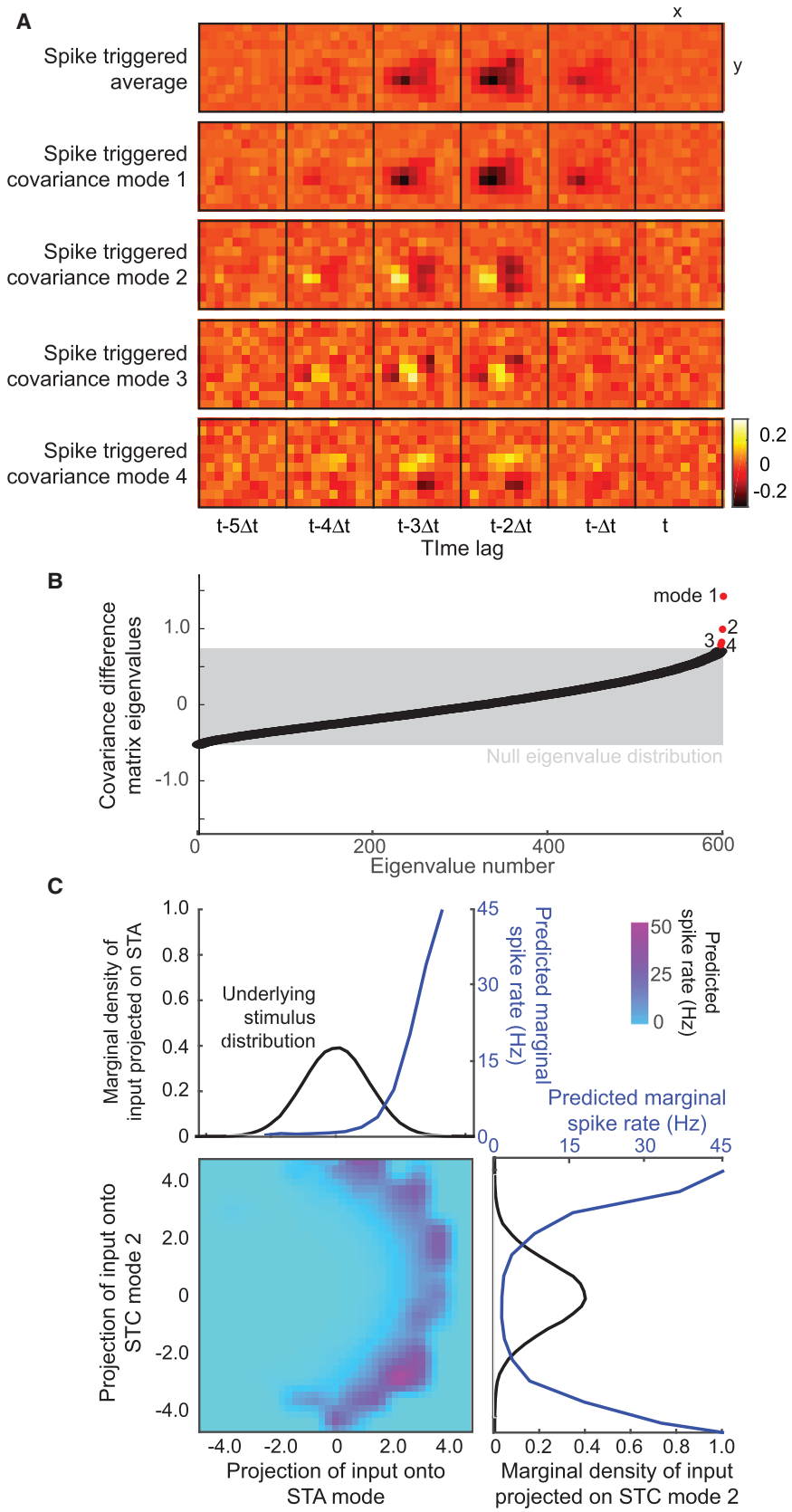


Figure 5. Spike-Thriggered Covariance Features for the Response of Retinal Ganglion Cell 3

(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 $\phi_{stc,1}$ has 0.93 overlap with ϕ_{sta}, while $\phi_{stc,2}$ through $\phi_{stc,4}$ have only a 0.20, 0.11, and a 0.05 overlap, respectively.

(B) The significance of each candidate STC feature, i.e., eigenvectors of $\mathbf{\Delta C}$ (Equation 20), were determined by comparing the corresponding eigenvalue (red and black) to the null distribution (gray shaded area). We used 1,000 repetitions of the calculation for randomized spike trains, corresponding to a confidence interval of 0.001.

(C) The nonlinearity in the space spanned by the STA and the second orthogonalized STC feature, after the STA feature was projected out (Equation 24), $\phi^{\perp}_{stc,2}$, completes the construction of the spiking model. The nonlinearity is found by the expectation procedure (Equations 6 and 7). The marginals of this distribution give nonlinearities with respect to the STA (top) and second STC features (right) alone.

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.

Calculating the STC with 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, $\mathbf{C}_\mathrm{p}$, whose eigenvalue spectrum was nearly flat (Figure 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 an underlying stimulus covariance matrix, $\mathbf{C}_\mathrm{p}$, 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 affected by the correlations within the stimulus (Bialek and van Steveninck, 2005) (compare Figure 2A and Figure 2C). The removal of these correlations is a process referred to as decorrelation or whitening. This may be applied to the case of so-called colored noise, where the power in different frequency bands is not equal, as it is in white noise. In this case, whitening corresponds to equalizing the power at each frequency. In the time domain, this corresponds to dividing by the underlying stimulus covariance matrix. Thus one can whiten the stimulus itself by dividing by the underlying stimulus covariance matrix (Theunissen et al., 2001; Schwartz et al., 2006), i.e.,

$$\widehat{\mathbf{s}}(t) = \mathbf{C}_\mathrm{p}^{-\frac{1}{2}}\mathbf{s}(t), \qquad \text{(Equation 25)}$$

and then proceed with the STA and STC analysis as defined by Equations 14, 15, 18, and 20 to 22 but with $\mathbf{s}(t)$ replaced by $\widehat{\mathbf{s}}(t)$. The matrix $\mathbf{C}_\mathrm{p}^{-1/2}$ is defined by

$$\mathbf{C}_\mathrm{p}^{-\frac{1}{2}} = \sum_{i=1}^{K} \lambda_i^{-\frac{1}{2}} \mathbf{v}_i \mathbf{v}_i^\top . \qquad \text{(Equation 26)}$$

where $\mathbf{C}_\mathrm{p}^{-1/2}$ has the same eigenvectors as $\mathbf{C}_\mathrm{p}$ (Equation 19) but the eigenvalues are the square root of the inverted eigenvalues.

Alternatively, the feature vectors may be calculated and the effect of correlations removed by dividing the feature vectors by the underlying stimulus covariance matrix (Equation 18). We denote the whitened features as $\widehat{\phi}_\mathrm{sta}$ and $\widehat{\phi}_{\mathrm{stc},i}$, where

$$\widehat{\phi}_\mathrm{sta} = \mathbf{C}_\mathrm{p}^{-1}\phi_\mathrm{sta}, \qquad \text{(Equation 27)}$$

$$\widehat{\phi}_{\mathrm{stc},i} = \mathbf{C}_\mathrm{p}^{-1}\phi_{\mathrm{stc},i}, \qquad i = 1, \ldots, K-1, \qquad \text{(Equation 28)}$$

and ϕ_sta and $\phi_{\mathrm{stc},i}$ are the estimates defined by Equations 14 and 22, respectively. As for the case of the matrix $\mathbf{C}_\mathrm{p}^{-1/2}$, the matrix $\mathbf{C}_\mathrm{p}^{-1}$ has the same eigenvectors as $\mathbf{C}_\mathrm{p}$ (Equation 19), but now the eigenvalues are simply inverted, i.e.,

$$\mathbf{C}_\mathrm{p}^{-1} = \sum_{i=1}^{K} \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^\top . \qquad \text{(Equation 29)}$$

Recall that $\mathbf{C}_\mathrm{p}$ and thus $\mathbf{C}_\mathrm{p}^{-1}$ are close to the identity matrix for white noise. The decorrelation procedure is also applied when producing the null eigenvalue distribution used to determine the significance of the STC features (Equation 23).

Regularization of the Inverse Covariance Matrix. The whitening procedure is usually numerically unstable, as it tends to amplify noise (David et al., 2004; Sharpee et al., 2008). This is because decorrelation attempts to equalize the variance in all directions. Yet the eigenvector decomposition of the underlying stimulus 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 $\mathbf{C}_\mathrm{p}$. In this case, it is best to simply remove stimulus components with small variance. This is done by replacing $\mathbf{C}_\mathrm{p}^{-1}$ with the pseudoinverse, a matrix in which the λ_i^{-1} is set to 0 for λ_i below a certain threshold, i.e., stimulus components along small eigenvalue modes have simply been discarded. The number of remaining non-zero λ_i^{-1} is called the order of the pseudoinverse (Penrose, 1955).

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:

$$\mathbf{C}_{\mathrm{p};L}^{-\frac{1}{2}} = \sum_{i=1}^{L} \lambda_i^{-\frac{1}{2}} \mathbf{v}_i \mathbf{v}_i^\top . \qquad \text{(Equation 30)}$$

$$\mathbf{C}_{\mathrm{p};L}^{-1} = \sum_{i=1}^{L} \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^\top \qquad \text{(Equation 31)}$$

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 choose a cutoff for directions in stimulus space for which the variance is considered to be too small to 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.

STA and Covariance from Thalamic Spiking during Whisking in Rat. Rat whisking provides an excellent example of active sensing in that 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 (Figure 6A). Whisking consists of an underlying 6–10 Hz rhythm whose overall maximum amplitude, or envelope, and local mean, or set-point, change slowly in

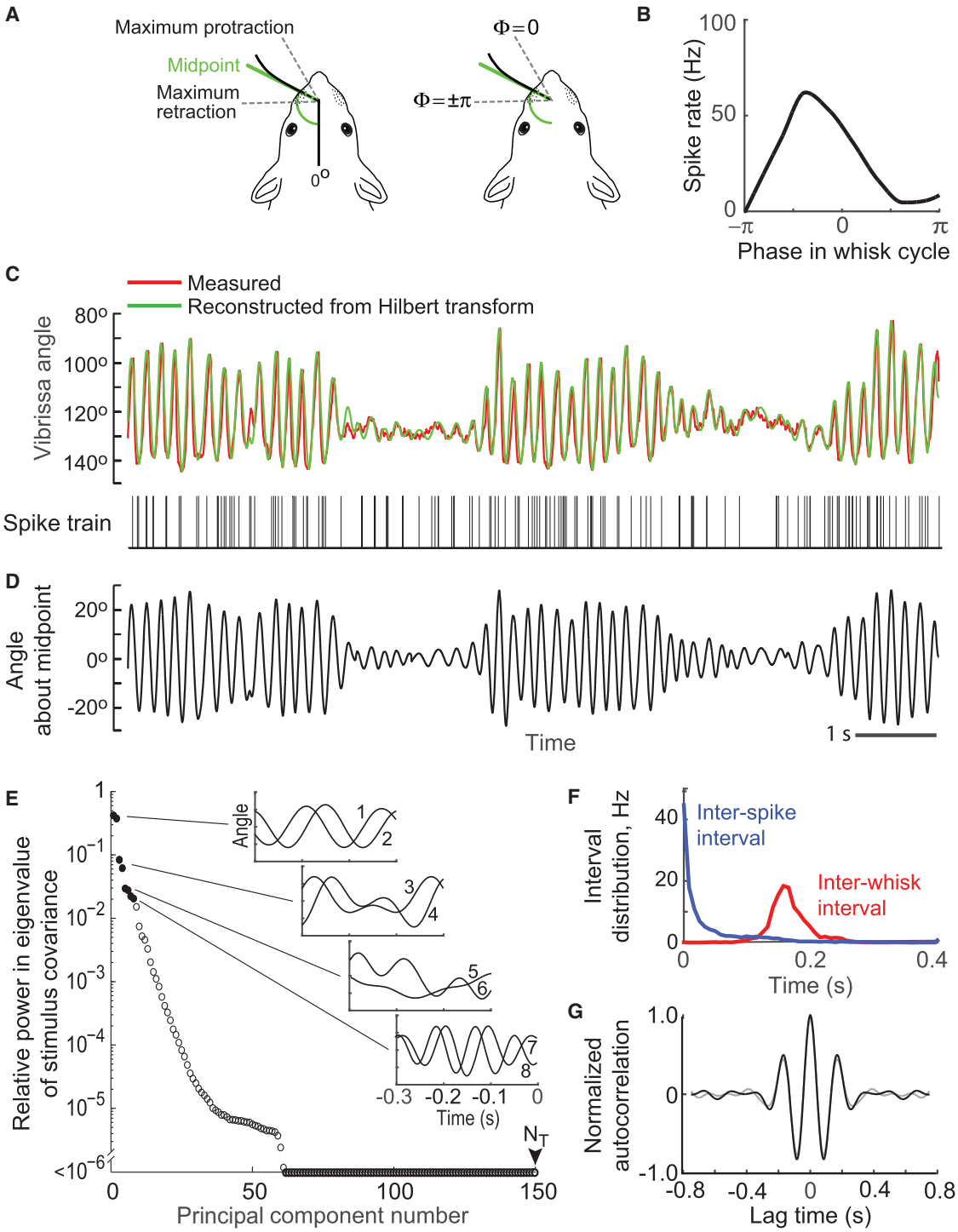


Figure 6. Spike Responses from Thalamic Cell 57 in Response to Whisking in Air without Contact

(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 $\theta_{\text{whisk}}(t) = \theta_{\text{protract}} - \theta_{\text{amp}}(1 - \cos(\Phi(t)))$ (Deschênes et al., 2016).

(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 slowly varying components θ_{protract} and θ_{amp} and the rapidly varying component $\Phi(t)$ were found by the Hilbert transform and the whisk reconstructed.

(D) Reconstructed whisk, leaving out slowly varying mid-point $\theta_{\text{protract}} - \theta_{\text{amp}}$. The self-motion stimulus is taken as the vibrissa position up to 300 ms in the past with $\Delta t = 2$ 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 (Equation 18). Note the highly structured dominant modes.

(legend continued on next page)

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) (Figure 6A), as many neurons have a preferred phase for spiking (Figure 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., 2015b) (Figure 6C); free whisking in air is a means to study the reafferent signal alone, as a touch-based sensory input must be decoded relative to the reafferent signal of vibrissa position (Kleinfeld and Deschênes, 2011).

To ensure that the mean firing rate is stationary over the time course of each behavioral epoch, we decomposed the whisking stimulus by computing the local phase and envelope using a Hilbert transform (Hill et al., 2011a) and removing shifts in the set-point of the motion (cf. green and blue traces of the reconstruction in Figure 6C). We then reconstructed the stimulus as changes in angle with respect to the set-point (Figure 6D). To analyze the spiking data relative to the reconstructed stimulus, 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 vibrissa, the dimensions of the stimulus are $N_X = 1$ and $N = N_T$. The underlying stimulus covariance (Equation 19) has eigenvalues that fall off dramatically by a few orders of magnitude (Figure 6E), in contrast with the nearly flat spectrum of white noise (Figure 3B). The dominant eigenvectors appear as sines and cosines at the whisking frequency (modes 1 and 2 in Figure 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 the inter-whisk interval shows a peak at the whisking frequency (Figure 6F), consistent with the form of the autocorrelation (Figure 6G). However, despite the presence of a strong rhythmic component in the stimulus, the inter-spike interval for a representative neuron appears largely exponential (Figure 6F).

We first consider the case of the feature vectors without whitening. We computed the STA feature (Equation 14) (Figure 7A) and the three significant STC features (Equation 20) (Figures 7A and 7B) for neurons in vibrissa thalamus. The STA feature appears as a decaying sine wave (Figure 7A), and the dominant STC feature appears as a phase-shifted version of the ϕ_{sta} (gray, Figure 7A). The overlap of $\phi_{\text{stc},1}$ with ϕ_{sta} is small, i.e., -0.06. Thus the dominant unwhitened STC feature could be safely orthogonalized relative to the unwhitened STA feature (Equation 24) 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 (Equation 29), where L was varied between 2 and 40. For each value of L we computed a predictive model, as described later, and chose the value of L for the whitening procedure that provided the best predictability. We show the decorrelated (Equation 27) and regularized (Equation 31) STA feature (Figure 7A) and the two significant STC features (Figures 7A and 7B). Here, the whitened STA and both whitened STC features are fairly similar to those for the unwhitened case, even though the analysis was restricted to a L-dimensional subspace spanned by the leading eigenvalues of $\mathbf{\Delta C}$ (Equation 20) after whitening. We next constructed the nonlinear input/output surface for the cell (Equation 9) using the whitened STA feature vector and the orthogonalized (Equation 24) first whitened STC feature vector (Figure 7C). The nonlinearity with respect to the STA feature alone appears as a saturating curve.

Before we leave the approach of nonparametric models, for which the features and nonlinearity are determined only by data, we note the method of maximally informative dimensions (Sharpee et al., 2003, 2004; Rowekamp and Sharpee, 2011) as an alternative means to find spike-triggering features and an arbitrary nonlinearity (Box 6). Rather than using a geometrical approach, this method uses the mutual information between the stimulus and the spike as a measure of the quality of the feature.

Models with Constrained Nonlinearities

The ability to find nonparametric stimulus features and nonlinearity can be severely constrained by the size of the dataset. As we have seen, with realistic amounts of data, such models are often undersampled, 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 certain aspects of the fitting problem.

Fixing the form of the nonlinearity allows one to pose a so-called "noise model" for the responses given the stimulus and the choice of model parameters. One then estimates the parameters of the model that best account for the data through an approach known as maximum likelihood. The likelihood is the probability of the observed data given a choice of model parameters, understood as a function of those parameters. Maximization of the likelihood function provides an estimate of the model parameters 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, has no local minima or maxima, and

(F) The inter-whisk and inter-spike intervals.
(G) The autocorrelation of whisking. Black is over all trials and gray is only over large amplitude whisks. Note the narrow band nature of this dataset.
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., 2015b). 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., 2015a); 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 4–14 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 dynamics of vibrissa motion only when the amplitude θ_{amp} was relatively high; therefore we constructed the models and tested their predictions only for periods when $\theta_{\text{amp}} \geq 10^\circ$.

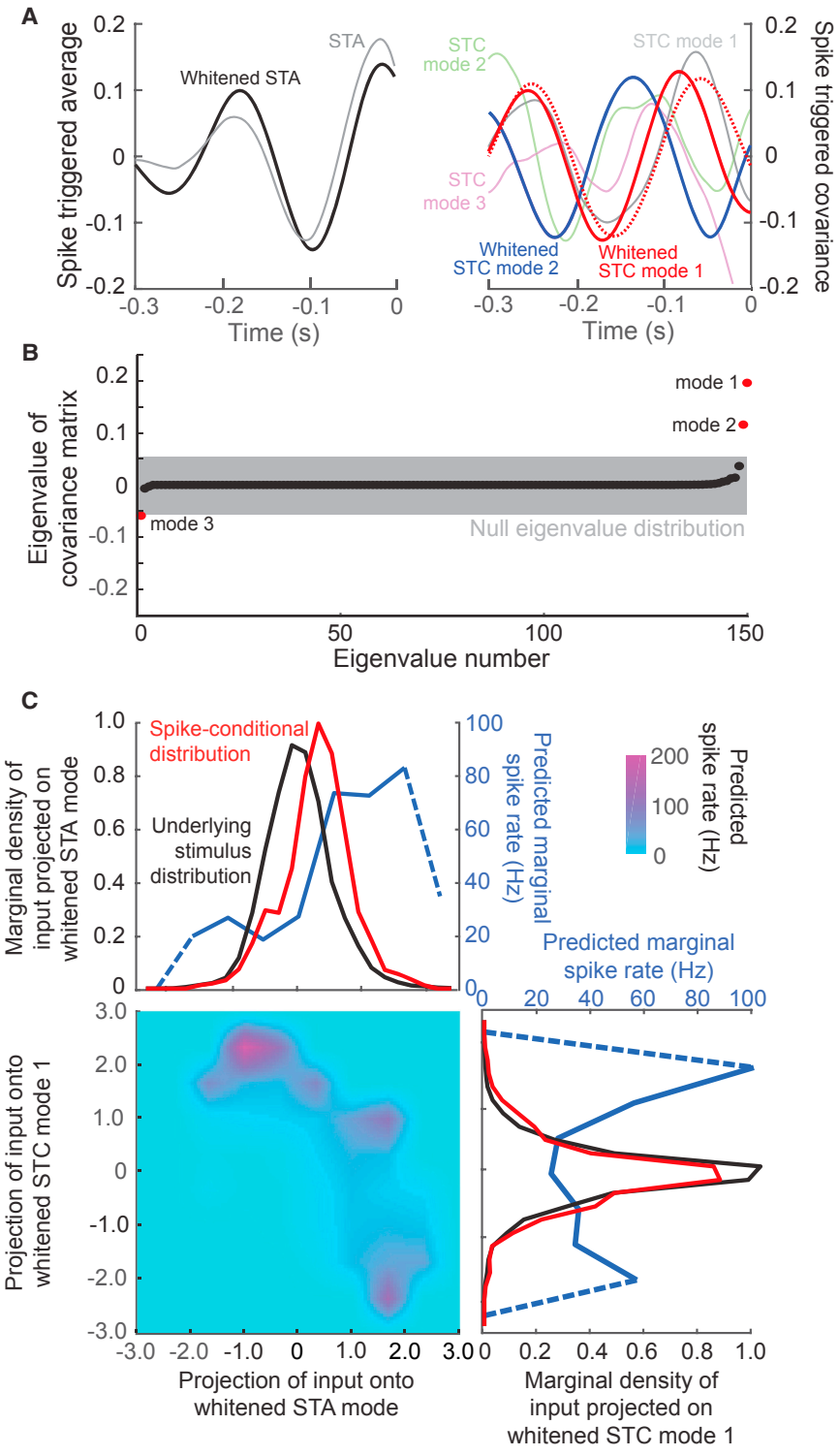


Figure 7. The Spike-Triggered Average and Spike-Triggered Covariance Feature Vectors for the Response of Thalamic Cell 57 in the Rat Vibrissa System

(A) The STA feature and the same feature computed for the whitened stimulus (Equation 30 with L = 18), along with the leading STC features calculated with and without whitening (Equation 30 with L = 18). The dashed curve is after projecting out the STA vector from STC mode 1.

(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 L = 18 eigenvectors, of which two were significant.

(C) A two-dimensional model of the nonlinearity for $\hat{\phi}_{sta}$ and the leading STC feature, $\hat{\phi}_{stc,1}$, both computed after whitening. We further plot the two marginals; dashed lines correspond to rare events.

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 for which the solution is unique as a result of convexity, the model may be accounting for variation that is specific to the portion of the data used for the fit. This is a phenomenon known as overfitting, and it manifests as a decrease in predictability of the model on novel datasets relative to the quality of the fit obtained with 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 jack-knife resampling, 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.

Maximum Noise Entropy Method

A principled way to specify a probabilistic model of the input-output transformation, $f(\cdot)$, is by searching for a conditional probability distribution of stimuli and responses, $p(\text{spike}\,|\,\mathbf{s})$. This can be done under the assumption that the variability in the response is described by a maximum entropy distribution, i.e., a distribution that is the least structured given

Box 6. Maximally Informative Dimensions

This method uses the mutual information between the stimulus and the spike as a measure of the quality of the feature (Sharpee et al., 2003, 2004; Rowekamp and Sharpee, 2011). An "informative" dimension is one in which, when a spike occurs, the spread of possible stimulus values along that dimension, as quantified by the entropy, is as small as possible. Thus, the MID approach implements a search to locate a feature that minimizes this entropy or, equivalently, maximizes the mutual information between stimulus and spikes. To understand this approach, we return to the definition of the nonlinearity based on the Bayesian procedure (Equations 8 and 9), 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

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

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, $p(z_1 \mid \text{spike})$ and $p(z_1)$, are as different as possible. One method for evaluating the difference between two probability distributions is the Kullback-Leibler divergence, defined as $D_{\text{KL}}(p,q) = \int dz\, p(z)\log[p(z)/q(z)]$, where $p(z)$ and $q(z)$ are probability distributions. Here, maximizing mutual information is equivalent to searching 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/.

stimulus and constraints set by measures on the data. In this approach, called the maximum noise entropy (MNE) 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 MNE joint distribution, with constraints to second order (Globerson et al., 2009; Fitzgerald et al., 2011a, 2011b) is given by

$$p(\text{spike} \mid \mathbf{s}) = \frac{1}{1+\exp\{a + \mathbf{h}\cdot\mathbf{s} + \mathbf{s}^{\top}\mathbf{J}\mathbf{s}\}}. \qquad \text{(Equation 32)}$$

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 parameters of the distribution (Equation 32) that best fit the data, i.e., have highest likelihood, are found via a gradient ascent algorithm. For each set of parameter values, the likelihood function is computed and the parameters are modified such that the likelihood function will increase in the next step, until a maximum is reached. Note that changes made to the parameters are not arbitrary: parameters must be changed such that a set of constraints is satisfied. Equation 32 represents a probability distribution that is normalized, so $\sum_{t_s} p(\text{spike} \mid \mathbf{s}(t_s)) = 1$. Additionally, the moments of the distribution (Table 1) must match those computed from the data. There is no need to use a spectrally white stimulus with MNE. Lastly, by convention, one seeks the minimum value of the negative of the logarithm of the likelihood rather than the maximum of the likelihood.

Interpreting the MNE Model. How does the MNE model (Equation 32) ensure the maximal variability in the spike rate? Consider the maximum entropy distribution (Equation 32) without any constraints, i.e., $a = \mathbf{h} = \mathbf{J} = 0$. The probability of a spike given a stimulus then is $p(\text{spike} \mid \mathbf{s}) = 1/2$ and can be thought of as the least structured spiking model. At every time

Table 1. Moments for the MNE Models

Moment	0	1	2
Element	scalar	$[i]$-th component of a vector	$[i,j]$-th component of a matrix
Symbol	$\langle r(t)\rangle$	$\langle r(t)\mathrm{s}[i](t)\rangle$	$\langle r(t)\mathrm{s}[i](t)\mathrm{s}[j](t)\rangle$
Data	n_T/M	$\frac{1}{M}\sum_t n(t)\mathrm{s}[i](t)$	$\frac{1}{M}\sum_t n(t)\mathrm{s}[i](t)\mathrm{s}[j](t)$
Model	$p(\text{spike})$	$p(\text{spike})\sum_{t_s}\mathrm{s}[i](t_s)p(\text{spike} \mid \mathbf{s}(t_s))$	$p(\text{spike})\sum_{t_s}\mathrm{s}[i](t_s)\mathrm{s}[j](t_s)p(\text{spike} \mid \mathbf{s}(t_s))$

n_T is the total number of spikes and t_s are the spike times.

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 $p(\text{spike}\,|\,\mathbf{s}) = p(\text{spike})$, but the overall firing rate is constrained to be the experimentally measured rate, denoted r_0. Now the goal of the fitting procedure is to find a such that $r_0 = p(\text{spike}\,|\,\mathbf{s}) = 1/(1+e^a)$, 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 (Equation 32). 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 equals that of the neuron. The parameters **h** and **J** act as linear feature vectors analogous to ϕ_{sta} and the $\phi_{\text{stc},i}$:

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

$$p(\text{spike}\,|\,\mathbf{s}) = \frac{1}{1+\exp\{a+\mathbf{h}\cdot\mathbf{s}\}}. \qquad \text{(Equation 33)}$$

This is equivalent to a STA model with a feature vector feature $\phi_{\text{sta}} = \mathbf{h}$ and a sigmoidal nonlinearity.

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

$$\mathbf{J} = \sum_{i=1}^{K} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\top}. \qquad \text{(Equation 34)}$$

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 Equation 32 as:

$$\mathbf{s}^{\top}\mathbf{J}\mathbf{s} = \sum_{i=1}^{K} \lambda_i (\mathbf{s}\cdot\mathbf{u}_i)(\mathbf{u}_i\cdot\mathbf{s}) = \sum_{i=1}^{K} \lambda_i z_i^2. \qquad \text{(Equation 35)}$$

Therefore, the eigenvectors of **J** with large eigenvalues, in absolute value, can be viewed as analogs of the STC features $\phi_{\text{stc},i}$ with a quadratic-sigmoidal nonlinearity. The match is not exact, as **J** is fitted in together with the linear component **h**, whereas $\phi_{\text{stc},i}$ were calculated from the covariance difference matrix (Equation 20), independently of the STA. Similarly to the STC method, the eigenvectors of **J** are orthogonal to each other by construction. Thus we may not interpret these spatiotemporal vectors as independent aspects of the input that drive the cell's response.

In the STC approach, the significance of a given feature was determined by comparing the corresponding eigenvalue of $\Delta\mathbf{C}$ (Equation 20) 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 **J** correspond to features, denoted $\mathbf{u}_i$, that significantly modulate the spiking output. We accomplish this by shuffling the entries of **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 **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.

Relation to Minimal Mutual Information. From an information theoretic point of view, the problem could be posed differently (Globerson et al., 2009). In the MNE approach, one seeks a minimally structured conditional distribution of responses $p(\text{spike}\,|\,\mathbf{s})$. In contrast, in the Minimal Mutual Information (MinMI) approach, one searches for the model, i.e., the distribution $p(\text{spike}\,|\,\mathbf{s})$, where the responses provide the least information about the stimulus. Thus the goal with MinMI is to find a lower bound on the information content sent to a hypothetical downstream population. The problems with the MinMI approach are that the resulting model is typically highly structured and biologically unrealistic, and that the lower bound represents a worst case scenario that is unlikely to be attained by a biologically plausible model.

MNE Models for Thalamic Spiking during Whisking in Rat. We applied the MNE procedure to the datasets obtained from thalamic recordings while rats whisked in air (Figure 8). As expected, the calculated first-order feature, **h**, closely approximated the STA feature (Figure 8A). We found that the top nine of $N = 150$ eigenvectors of **J** were statistically significant (Figure 8B). The dominant feature, $\mathbf{u}_1$, makes a substantial contribution at short times, like the dominant STC feature $\phi_{\text{stc},1}$ (Figure 8A), but decays much more rapidly than the STC feature. The higher-order features calculated from **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 in applying the MNE model. First, in its raw form, the fitting procedure can generate high-frequency components that are not represented in the stimulus and thus not constrained. In the present case, we address this by removing the components orthogonal to the first 15 principal components of the stimulus from the feature vectors of the model. Second, we use the full matrix **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 **J** do not sum exactly to zero. Third, while overfitting is always a potential problem, this did

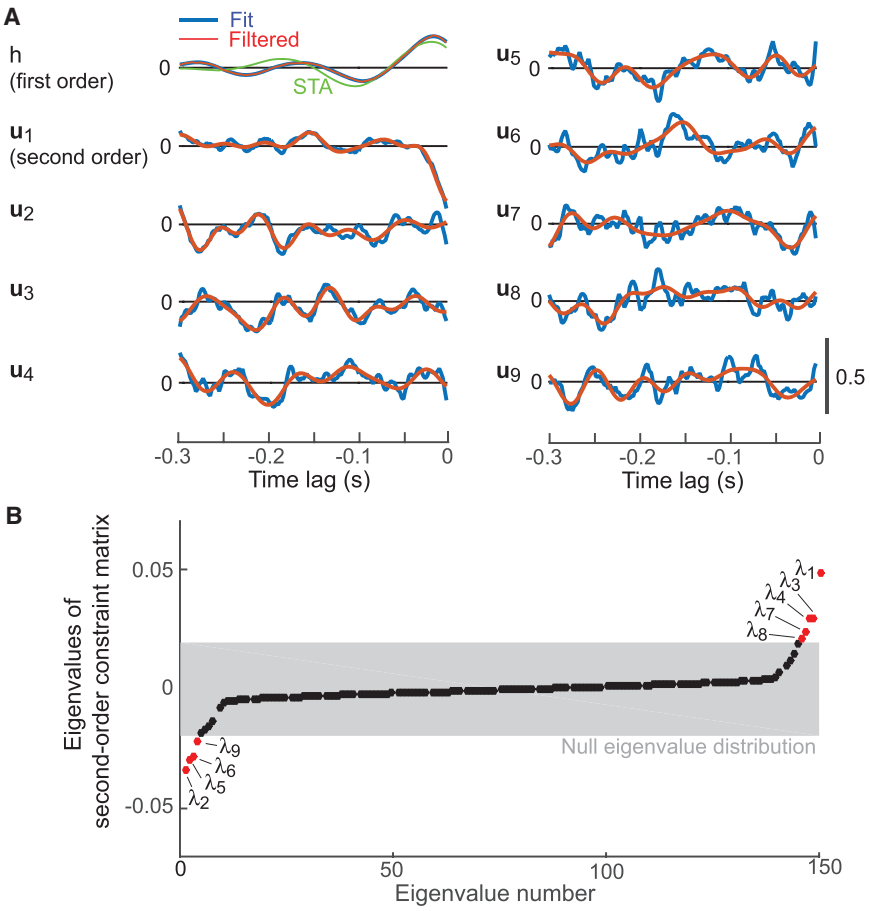


Figure 8. The Dominant Features Calculated by the Maximum Noise Entropy Method for Example Thalamic Cell 57

(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., **h**, and statistically significant second-order feature vectors, i.e., eigenvectors of **J** (Equation 34). 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 **J** to a null distribution.

not arise with this dataset, possibly because of the rapid fall-off of the eigenvalues for the covariance of the stimulus matrix (Figure 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.

Separability

The feature vectors in the first two models we discussed, namely STA and STC, are computed directly from the spike-triggered and underlying stimulus distributions and do not require a fitting procedure to be applied. As such, they do not degrade significantly if 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. For algorithmically fit models, the number of parameters scales with the dimensionality of the stimulus. In MNE, for example, the scaling is linear for a first-order model (Equation 33) and quadratic for a second-order model (Equation 34). Therefore, these models may suffer from overfitting, as a presentation of a large number of stimulus samples 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 neuron, they may greatly reduce the number of spikes needed to fit the model or help prevent overfitting.

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 (Equation 12) 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 $\phi_i^{X,d}$ and $\phi_i^{T,d}$, respectively, where d labels the mode.

We express ϕ_i in the same form of a matrix for the space time stimulus (Equation 12), i.e.,

$$\phi_i(x,t) = \overbrace{\begin{pmatrix} \phi_i(1,1) & \cdots & \phi_i(1,N_T) \\ \vdots & \ddots & \vdots \\ \phi_i(N_X,1) & \cdots & \phi_i(N_X,N_T) \end{pmatrix}}^{N_T \text{ time points}} \Big\} N_x \text{ spatial positions}$$

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

(Equation 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 spatial and temporal contribution, i.e., only the $d = 1$ term in Equation 36 is used. 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 $N_X + N_T + 1$.

Separability of the Nonlinearity. Another important form of separability relates to the nonlinear function $g(\cdot)$ (Equations 7 and 9). 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:

$$g(z_1, \ldots, z_K) = g_1(z_1) \times \ldots \times g_K(z_K). \quad \text{(Equation 37)}$$

This approximation is equivalent to assuming that the joint conditional probability distribution over the projections of the stimulus on the feature vectors, $p(z_1, z_2, \ldots, z_K|\text{spike})$, 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 feedforward 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 (GLM).

Generalized Linear Models

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, along with other more general dependencies, can be incorporated in the framework of a GLM (Nelder and Wedderburn, 1972; Brown et al., 1998).

GLMs are a flexible extension of standard linear models that allow one to incorporate nonlinear 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 spike counts, 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.,

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

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 (Figure 1A). In its simplest form, the spike rate is given by

$$r(t) = g\left(c + \sum_{t'<t} \phi_{\text{glm}}(t') \cdot \mathbf{s}(t') + \sum_{t'<t} \boldsymbol{\psi}(t') n(t')\right), \quad \text{(Equation 39)}$$

where the parameter c sets the overall level of the firing rate, the sum $\sum_{t'<t} \phi_{\text{glm}}(t') \cdot \mathbf{s}(t')$ is the familiar projection of the stimulus onto the spatiotemporal feature $\phi_{\text{glm}}(t)$, 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. 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, $\mathcal{L}(\Theta) = P(n(t) \mid \Theta)$, 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 be 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). We adopt an exponential non-linearity for all subsequent analyses.

Rather than maximize the likelihood function, we maximize the logarithm of the likelihood function, referred to as the log-likelihood, which for Poisson spiking is

$$\log\mathcal{L}(\Theta) = \sum_t n(t) \log(r(\mathbf{s}(t) \mid \Theta)\Delta t) - \sum_t r(\mathbf{s}(t) \mid \Theta)\Delta t, \quad \text{(Equation 40)}$$

where $r(\mathbf{s}(t) \mid \Theta)$ is the predicted firing rate, and we drop the $n(t)!$ term, as it is independent of the model. With this, the computational fitting problem we solve is simply

$$\underset{\Theta}{\text{argmax}}\, (\log\mathcal{L}(\Theta)), \quad \text{(Equation 41)}$$

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

Overfitting and Regularization. As for other methods, the model that best fits the training data may 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 number of time points in the experiment M, 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 a model that contains 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 and other algorithmically fit models is the appearance of high-frequency components in the feature vector when the stimulus is slowly varying. This occurs because the fast variations minimally affect the predicted spike-trains and the likelihood when the slowly varying stimulus is projected onto them (Equations 40 and 41). While their effect on the log-likelihood may be minimal, they obstruct interpretation of the feature vectors ϕ_{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, i.e.,

$$\underset{\Theta}{\text{argmax}}(\log\mathcal{L}(\Theta) - \text{Q}(\Theta)). \qquad \text{(Equation 42)}$$

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,

$$\underset{\Theta}{\text{argmax}}(\log\mathcal{L}(\Theta) - N_\Theta), \qquad \text{(Equation 43)}$$

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 $\boldsymbol{\Theta}_{MAP}$, where

$$\boldsymbol{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}}\log p(\Theta \mid \mathbf{s};\mathbf{r}) = \underset{\Theta}{\text{argmax}}(\mathcal{L}(\Theta;\mathbf{s},\mathbf{r}) + \log p_\Theta(\Theta)). \qquad \text{(Equation 44)}$$

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 of the form

$$Q(\Theta) = \kappa\Theta^\top \mathbf{D}\Theta. \qquad \text{(Equation 45)}$$

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 κ is a regularization parameter that determines the weight given to the prior probability compared to the likelihood. It 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 iterations that recovers a reasonable solution.

Choice of Basis. Overfitting can also 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 (Equation 18) 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 to represent spike history filters is a 'raised cosine' basis, denoted by $\text{w}_i(t)$. This specific basis, despite a complicated functional form, describes a set of bumps whose peaks are tightly spaced near the time of the spike and become increasingly sparse for earlier times (Pillow et al., 2008). The first function, w_0, is given by

$$\text{w}_0(t) = \begin{cases} 1, & 0 < t < t_0 \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(Equation 46)}$$

and the additional w_i are given by

$$w_i(t) = \frac{1}{2}\begin{cases} 1 + \cos\left(\frac{\pi}{2}\left(\eta \log\left(\frac{t+t_1}{t_0+t_1}\right) - i + 1\right)\right), & t \geq t_0 \text{ and } i - 3 < \eta \log\left(\frac{t+t_1}{t_0+t_1}\right) < i + 1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{(Equation 47)}$$

where t_0 is the refractory time, t_1 sets the density of the basis functions, t_2 sets the onset of decay, B is the number of functions with $i = 0, \ldots, B-1$, and $\eta \equiv (B-2)/\log((t_2+t_1)/(t_0+t_1))$. This basis is well resolved where the spike history filter is expected to change most rapidly.

Stability. Despite much theory surrounding their application (Paninski, 2004; Brown et al., 1998), correctly specifying a

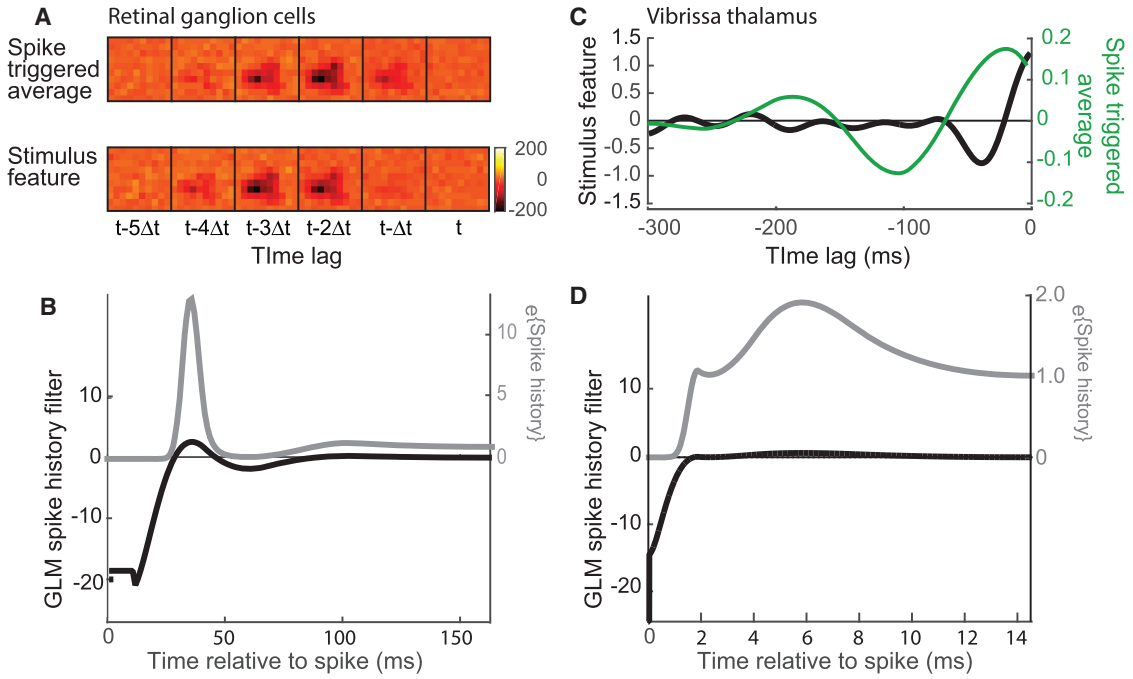


Figure 9. The Fit of the Generalized Linear Model for the Responses of Retina Ganglion Cell 3 and Thalamic Vibrissa Cell 57
(A and C) The stimulus feature ϕ_{glm} compared with the previously calculated STA feature.
(B and D) The spike history filters ψ (black curves) along with the exponent of the filter (gray). The history filters were expanded in the basis of equations 46 and 47 using $B = 5$, $t_0 = 0.333$ s (retinal ganglion cells) or $t_0 = 0.020$ s (thalamic cells), $t_1 = 0.4$ s, and $t_2 = 2.0$ s.

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, i.e., yield a sensible prediction that can be compared to data. Unstable models usually have predicted spike rates that diverge to infinity when stimulated by novel stimuli. While such models may still provide insight from the form of their feature vectors, they are not able to predict spike trains for 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.

GLMs for Retinal Ganglion and Thalamic Cells. We fit GLMs for the set of retinal ganglion cells stimulated with white noise (Figure 3) and thalamic neurons stimulated by self-motion of the vibrissae (Figure 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 (Equations 46 and 47) were used as basis of the spike history filter. For the same representative neuron used previously, the feature vector ϕ_{glm} corresponds to a transient spot of illumination that is similar to the STA feature yet slightly delayed in time (Figure 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 timescale similar to the stimulus refresh, Δt. Since the effect of the spike history filter is exponentiated (Equation 39), we plot both the result of the fit (black line, Figure 9B) and the exponent of the filter (gray line, Figure 9B) to more clearly illustrate the effect that this component of the model has on the predicted firing rate.

The GLM fit in the case of correlated noise gives a less intuitive but perhaps more interesting result. Here the 12 leading terms of a PCA of the stimulus were used as the basis functions for the stimulus feature vector, and, again, raised cosines were used as the basis of the spike history filter. We observe that the feature vector ϕ_{glm} oscillates for less than one whisk cycle and returns to baseline very quickly (Figure 9C); it is quite different from the ϕ_{sta}. Further, the spike history shows a significant excitatory component, delayed by ~6 ms, that is likely to generate a burst of spikes at an approximately fixed position in the whisk cycle (Figure 9D). The GLM analysis therefore suggests that the thalamic cell is very responsive to changes in position of the vibrissae but has little dependence on the history of the stimulus or spiking at times earlier

Box 7. Spectral Coherence as Regression

Coherence may be viewed in analogy to the more familiar Pearson correlation coefficient in linear regression. The expected value of the predicted rate given the observed rate is

$$\mathcal{E}(\tilde{r}(f)\,|\,\tilde{r}_s(f)) = \tilde{b}(f)\tilde{r}_s(f),$$

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

$$\tilde{b}(f) = \tilde{C}(f)\sqrt{\frac{\left\langle |\tilde{r}(f)|^2 \right\rangle}{\left\langle |\tilde{r}_s(f)|^2 \right\rangle}}.$$

Thus, to the extent that real and imaginary parts of both $\tilde{r}(f)$ and $\tilde{r}_s(f)$ may be considered as Gaussian variables, $\tilde{C}(f)$ forms part of the regression coefficient. The variance of the expectation, denoted $\mathcal{V}(\tilde{r}(f)\,|\,\tilde{r}_s(f))$, is given by

$$\mathcal{V}(\tilde{r}(f)\,|\,\tilde{r}_s(f)) = \left(1 - \left|\tilde{C}(f)\right|^2\right)\left\langle |\tilde{r}(f)|^2 \right\rangle$$

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

than approximately 80 ms in the past, which corresponds to about half of a whisk cycle.

Model Evaluation

How well does each of the models perform in predicting the spike rate for data that have the same statistical properties as the training set but are otherwise novel? For each model and dataset, 80% of the data were used as the training set for fitting the model, and the remaining 20% were 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 $r_s(t)$ 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 in general for natural stimuli, one only has a single presentation of the stimulus as 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.

Log-Likelihood

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

$$\log\mathcal{L}(\phi_i) = \sum_t (n(t)\log(r(t)\Delta t) - r(t)\Delta t). \qquad \text{(Equation 48)}$$

where $r(t)$ is the predicted response of the model under evaluation. 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 $\log(r(t)\Delta t)$ 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.,

$$\log\mathcal{L}_{\text{null}} = \sum_t (n(t)\log(\langle n(t)\rangle) - \langle n(t)\rangle), \qquad \text{(Equation 49)}$$

where $\langle n(t)\rangle$ is the average spike count in each bin. The confidence level is determined by a jack-knife procedure (Sokal and Rohlf, 1995) in which the 20% testing part of the data is permuted with the training part.

Spectral Coherence

A complementary metric for the fidelity of the predicted spike trains is the spectral coherence between the predicted and measured responses. Coherence provides a measure of correlation between signals in the frequency domain and thus 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 $\tilde{r}_s(f)$ as the Fourier transform of the predicted and measured rates, respectively. The spectral coherence, denoted $\tilde{C}(f)$, is:

$$\tilde{C}(f) = \frac{\left\langle \tilde{r}(f)\tilde{r}_s^{\star}(f) \right\rangle}{\sqrt{\left\langle |\tilde{r}(f)|^2 \right\rangle \left\langle |\tilde{r}_s(f)|^2 \right\rangle}}. \qquad \text{(Equation 50)}$$

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 about 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 (Box 7).

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 for the retinal ganglion cells (Figures 10A and 10B). The GLM yields representative spike trains, as opposed to rates, so that we calculated 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 (Equation 48), while the GLM has the lowest, lower even than the STA (Figure 10C). The relatively poor performance of the GLM may imply overfitting of the training data, as models that involve more parameters have a larger log-likelihood. All models perform better than the null expectation (Equation 49) (Figure 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 (Figure 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, i.e., STA, STC plus STA, and GLM, show a rather strong frequency dependence. The coherence is substantially below $\left|\tilde{C}(f)\right| = 1$ at all frequencies, yet it is highly statistically significant (Figure 10E). Consistent with expectations from the log-likelihood (Figure 10C), the STC plus STA model has an approximately 5% improved coherence at all frequencies (Figures 10E and 10F). The GLM yielded inferior predictions. 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 (Figure 10E). This implies that the predicted spikes arrive with a brief time delay that is estimated to be $(1/2\pi)(\Delta\text{phase}/\Delta f) = -25$ ms or less than Δt (inset, Figure 10A).

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 (Figures 4C and 4D). The coherence shows an improvement with the STC plus STA model (Figures 10E and 10F). The GLM gives the poorest 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 overfitting (results not shown). The parameters from fitting the stimulus set with $N = 600$ (Figures 4C, 5A, and 10A) led to a stable calculation of the linear feature, **h**, and three statistically significant second-order features (Equation 32). Yet the model gave poor predictions, with a log-likelihood metric that was lower for the MNE model than for the null hypothesis (Equation 49) and a spectral coherence that was relatively small. To reduce overfitting, we truncated the stimulus. The log-likelihood for this reduced model increased, and there was a concomitant increase in the spectral coherence at all frequencies, although the coherence was still lower than that achieved with the other models.

Validation of Models with Correlated Noise from Self-Motion

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

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., ϕ_{sta}, and the feature vector after whitening $\widehat{\phi}_{\text{sta}}$ (Figures 7A and 11B). The relative values of the log-likelihood function were too uncertain to offer insight. Interestingly, the spectral power for the whitened and nonwhitened feature vectors are essentially the same at all frequencies (Figure 11D). Further, we observe that whitening slightly albeit insiginicantly increases the coherence between the predicted and the measured rates at the whisking frequency as well as at other frequencies (Figures 11E and 11F).

We computed the predicted rate from the feature vectors for the STC plus STA model, i.e., $\phi_{\text{stc},1}$ and ϕ_{sta}, and the feature vectors after whitening, i.e., $\widehat{\phi}_{\text{stc},1}$ and $\widehat{\phi}_{\text{sta}}$ (Figures 7A and 11B). Unlike the case for the STA feature alone, whitening increases the coherence between the predicted and the measured rates at the whisking frequency (Figure 11E), with $\left|\tilde{C}(f)\right|$ increasing from 0.65 to 0.70, as well as at other frequencies. The exception is that the coherence below about 1 Hz is better for the nonwhitened STC plus STA feature vectors (Figure 11E).

Across all models, the best predictability at the whisking frequency occurred with the whitened STC plus STA model, albeit by an increase of only approximately 10% compared with the other models. Some of the models exhibited a slight phase advance at the whisking frequency. The largest advance occurred for the STA model and 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" or "worse" at all frequencies. The MNE model appeared to be the least coherent with the measured train at the lowest frequencies, yet the phase lag with the NME model was minimal near the whisking frequency (Figure 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 (Figure 6B):

$$p(\text{spike}\,|\,\Phi) = \frac{p(\Phi\,|\,\text{spike})\,p(\text{spike})}{p(\Phi)}. \qquad \text{(Equation 51)}$$

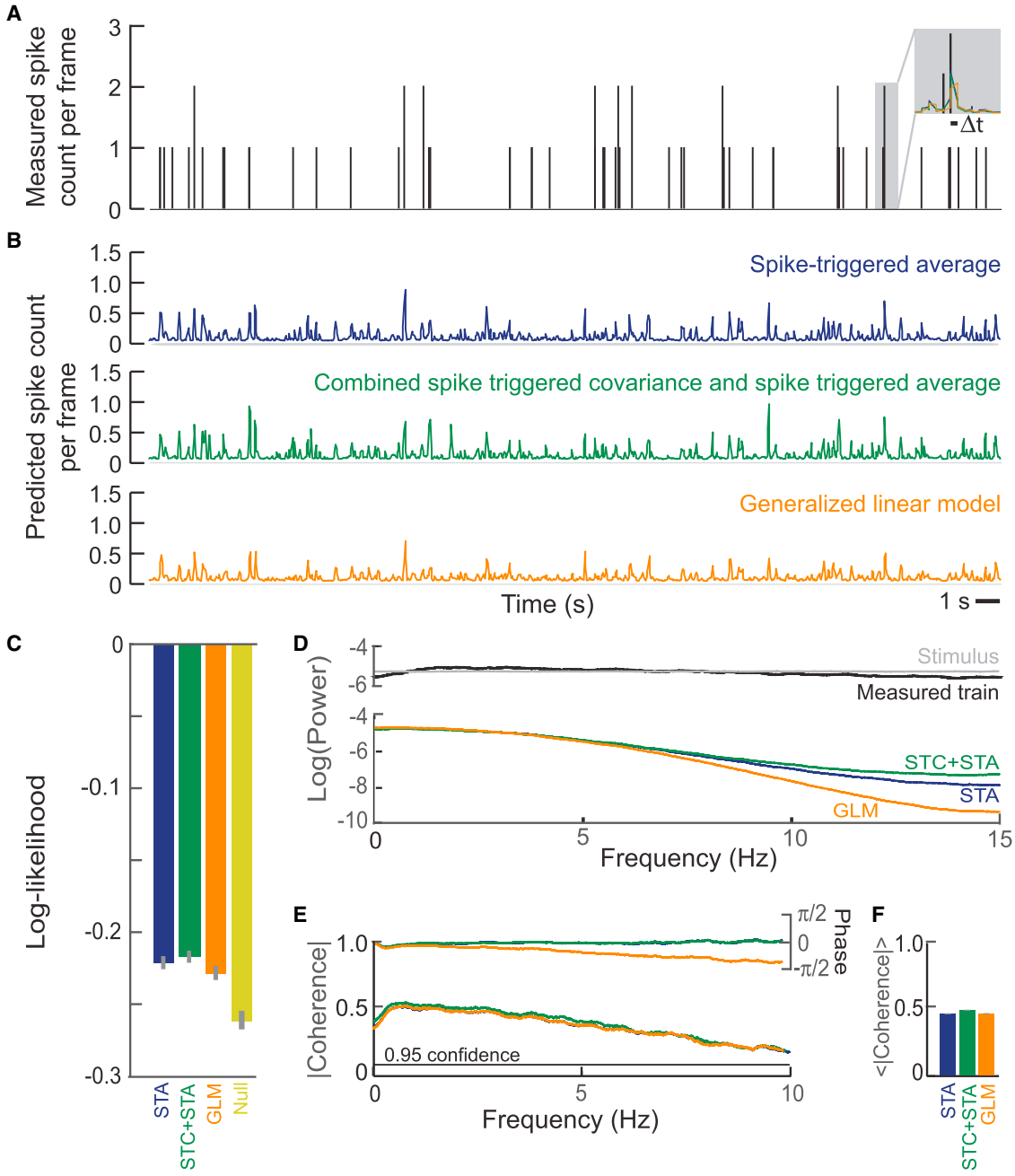


Figure 10. Summary of the Performance of Model Predictions for the Retinal Ganglion Cell 3

(A–F) 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.

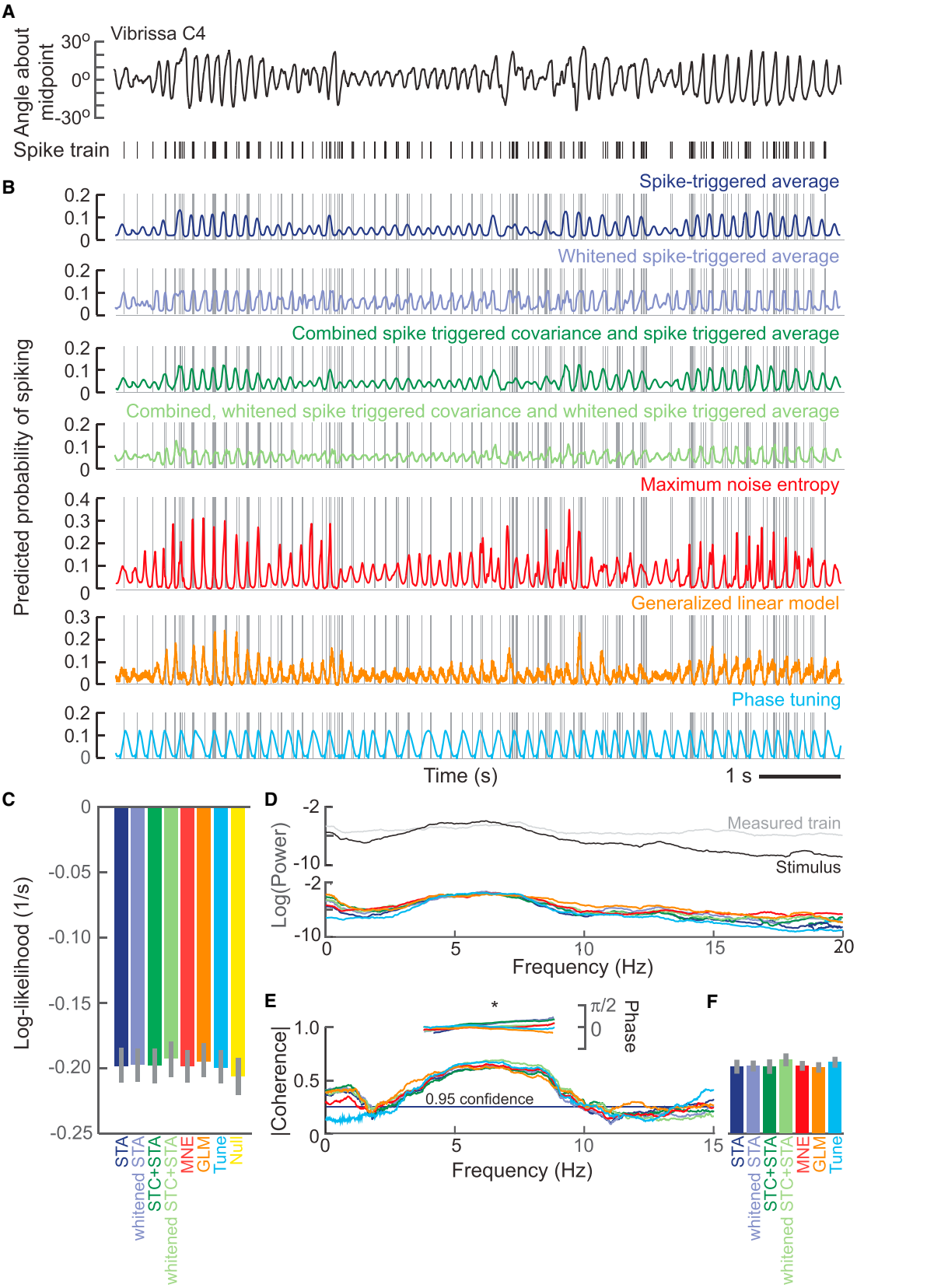
(B) The predicted spike count per frame obtained by computing the probability of a spike corresponding to each stimulus frame (top, STA; middle, STC; bottom, GLM). Note that to generate a prediction from the GLM at time t we need the history of the spike train up to that point $t' < t$, which is not 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 (Equation 48) of each model given the test set, which quantifies the quality of the prediction. We further include the log-likelihood for the null condition (Equation 49). The bars are one standard error jack-knife estimates.

(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 across four epochs of 920 s of data or reconstructions.

(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 and four epochs of data and reconstructions.

(F) The coherence averaged from 0.5 to 5.0 Hz, together with one standard error jack-knife estimates for the average.



(legend on next page)

The phase tuning model achieves the same high level of coherence at the whisking frequency as the whitened STC plus STA model (Figures 11D–11F). This suggests that the feature vectors are largely acting as broadband filters. Of course, the tuning model performs badly for frequencies away from the ~6 Hz whisking peak (Figure 11E).

Finally, we consider two additional thalamic neurons that had extreme response properties (Figure 12). The first is a neuron that tended to spike with respect to changes in the amplitude of whisking (Figures 12A–12D). Here the whitened STA and STC plus STA models did well, the MNE model exhibited significant coherence over the broadest frequency range, and the GLM did poorly (Figure 12E). On the other hand, we consider a neuron that responds almost solely to the phase of whisking (Figures 12F–12H). The whitened STA and STC plus STA models performed best at the whisking frequency (Figures 12I and 12J); the phase tuning model performs particularly well.

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, the whitened STC plus STA model had the highest coherence at the whisking frequency (Figure 11F). Yet we were also 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.

A noteworthy issue concerns the jack-knife estimate of the standard error. Unlike the case of a continuous record for retinal ganglion cells, the thalamic neurons were recorded for one to two dozen whisking bouts, each a few seconds long. Thus the test set for each jack-knife consisted of a recording of variable length, as opposed to exactly 20% of the data. Additionally, whisking variables such as amplitude and frequency are not stationary but change from bout to bout during active sensing behavior. These experimental issues, taken together with the log-likelihood being a shallow function that depends on the average firing rate, led to systematic differences between the different jack-knives that are larger than the differences between models; note relatively large error bars in the figure (Figure 11C). These issues were partly resolved by looking at the spectral coherence (Figures 11E and 11F), which is less susceptible to systematic differences.

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

$$r_i(t)=\exp\left\{c_i+\sum_{t'<t}\phi_i(t')\cdot\mathbf{s}(t')+\sum_{j=1}^{M}\sum_{t'<t}\psi_{ij}(t')n_j(t')\right\}. \quad \text{(Equation 52)}$$

The incorporation of such network filters has 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 inter-neuronal 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 feature vector (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 (Equation 52) 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). However, expanding the stimulus to encompass the spiking history of the rest of the network stands in contrast to validation of a GLM that represents a network with feedback between neurons, 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.

Figure 11. Summary of the Performance of Predicted Spike Trains for Thalamic Neuron 57

(A–F) Seven means of analysis are compared, i.e., STA, STA after whitening of the stimulus, STC plus STA, STA and STC 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.

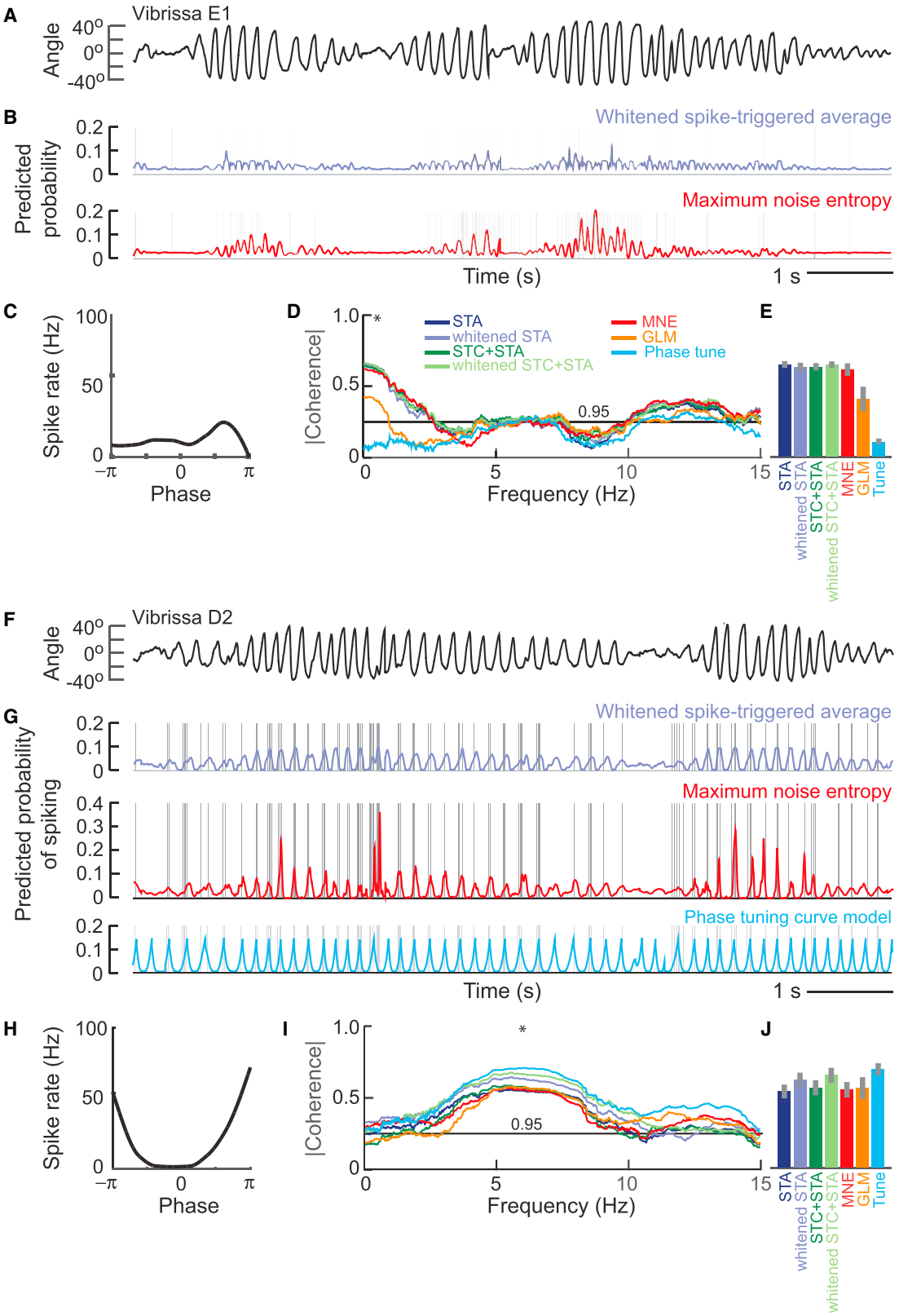
(B) The predicted probability of spiking per 2 ms time bin obtained by computing by each model and the corresponding stimulus. Note that to generate a prediction from the GLM at time t, we need the history of the spike train up to that point, $t' < t$, which is not deterministic due to the Poisson variability. 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 (Equation 48) of each model given the test set, which quantifies the quality of the prediction. The bars are one standard error computed as a jack-knife estimate.

(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 (Equation 50). Coherence was computed with a half-bandwidth of 1.2 Hz as an average over 49 spectral estimators.

(F) The magnitude of the coherence at the peak of the spectrum for whisking (*), 6.7 Hz, with one standard error jack-knife estimates.



(legend on next page)

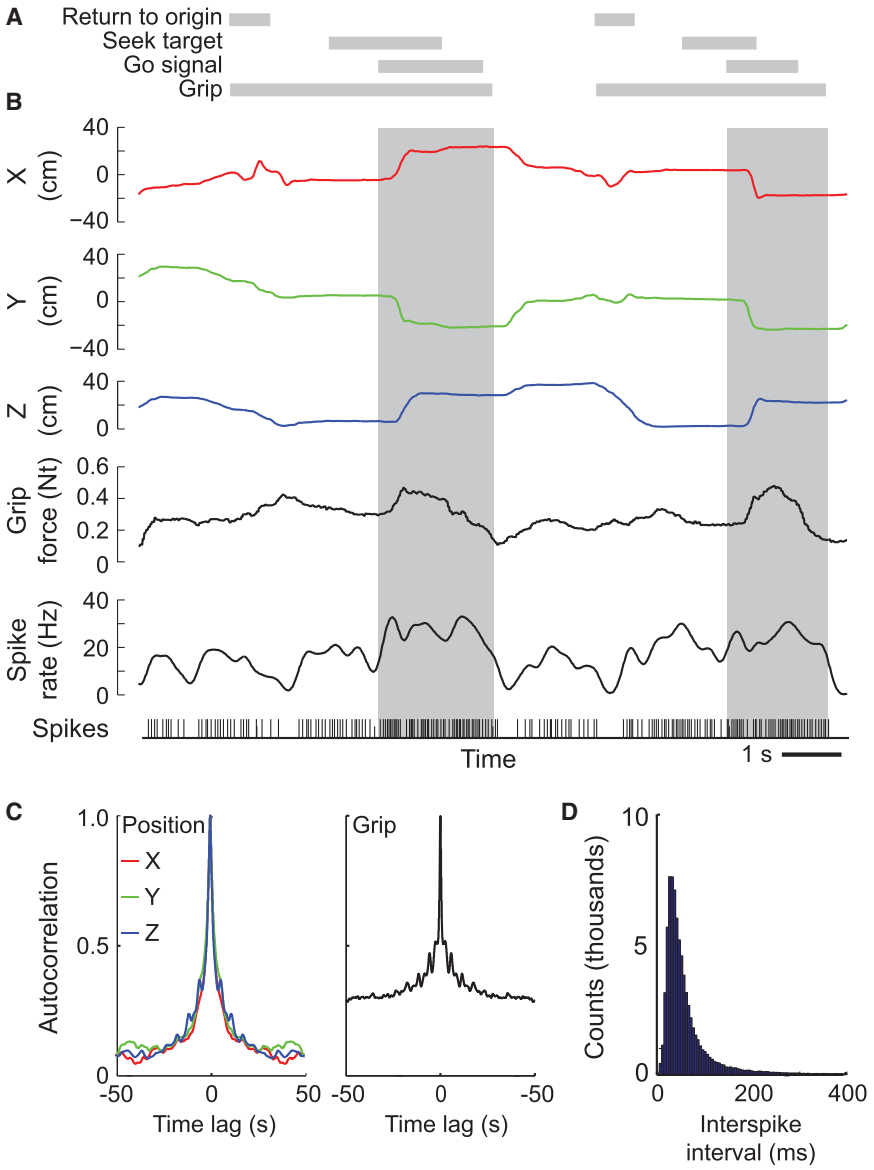


Figure 13. Summary of the Three-Dimensional Monkey-Based Reach Task with Spike Data from Unit 36

(A–D) Analysis is based on a single ~90 min recording session in which the monkey performed 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 eight 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 autocorrelation.

(D) Distribution of inter-spike 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 100 Hz. Of the isolated units, we selected those that showed no evidence of contamination based on inspection of the inter-spike interval distribution. Analysis was performed from the time of the Go signal until the grip was released; see gray band in (B).

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 performed a grip-and-reach motor task (Engelhard et al., 2013). The GLM consists of feature vectors that relate to hand motion, as measured by a cursor trajectory and grip force (Figure 13A), that were modeled with Gaussian-bump basis functions. Since motor neurons can encode future motor outputs, the stimulus feature encodes both past and future relationships relative to the current time bin (Figure 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 (Equations 46 and 47) for the spike history filters and the coupling filters for the histories of other neurons in the network.

The cursor position and grip data vary over hundreds of milliseconds to seconds (Figure 13C), while the spike history data vary on the order of milliseconds. Capturing effects on these disparate timescales within the same model requires some

Figure 12. Summary of the Performance of Predicted Spike Trains for Two Additional Thalamic Cells, Units 88 and 99

(A–E) The whisking stimulus (A) and predicted spike probabilities (B) for a cell with weak phase tuning (C). Yet this cell was strongly modulated by the amplitude of whisking, which changes on a slow timescale, approximately 1 s, compared with changes in phase. The predicted rate is shown for two models that perform best, i.e., STA after whitening of the stimulus and MNE. The phase tuning model performs poorly as it ignores the amplitude (D). The one standard error jack-knife was calculated for the coherence at low frequency (*), 0.5 Hz (E).

(F–J) The whisking stimulus (F) and predicted spike probabilities (G) for a cell with particularly strong phase tuning (H). The predicted rate is shown for three models that perform 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 (I). The one standard error jack-knife was calculated for the coherence at low frequency (*), 0.5 Hz (J).

care, as the values of the data have a non-Gaussian distribution and are highly temporally correlated; as noted, this correlation can result in uninterpretable high frequencies in the feature vectors. This requires some form of regularization. Here, we used 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 autocorrelation timescale (Figure 11C).

The fitting was performed only on data within the movement phases of the trials, i.e., from the beginning of the "Go signal" to the end of "Grip pressed" (gray bands in Figure 13B). 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 connections between neurons. This takes the form (Equations 42 and 45)

$$\underset{\Theta}{\operatorname{argmax}}\left(\log\mathcal{L}(\Theta)-\kappa\sum_{i\neq j}^{M}\|\Theta_{i,j}\|^{2}\right), \qquad \text{(Equation 53)}$$

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 and ensures global convergence.

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 a contiguous block of 20% of the total data, or 4 min of recording, is used for validation. We take a value $\kappa = 100$ in our network analysis (Equation 53); 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 the log-likelihood estimate (Equation 48), by averaging repeated simulations of the GLM given the same stimulus. This validation process was repeated five times, selecting a non-overlapping 20% of the data for testing each time. Combining the likelihood and coherence estimates over all the individual estimates allows the mean likelihood and the standard error of the mean likelihood to be determined for both the coupled and uncoupled models.

With respect to a representative example cell (Figure 13), we find that the history filters are the same for the coupled and uncoupled cases (Figure 14A), coupling terms are present on a variety of timescales, (Figure 14B), and the stimulus feature vectors are altered in magnitude by the coupling (Figure 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 (Figure 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 for the retina and thalamus datasets, more information can be gleaned from the coherence between the predicted rate and the observed spike train than from the log-likelihood. Significant spectral power in both coupled and uncoupled cases only occurred for low frequencies, i.e., 0–1 Hz, and the coupled model had a significantly higher coherence in this range for some cells (Figure 14E). This increase was statistically significant and particularly strong for our example cell (red ellipse in Figure 14F) and one other cell (blue ellipse in Figure 14F) and barely significant in three other cells (green ellipses in Figure 14F). Thus network interactions through the spike history terms of neighboring cells improve the ability to predict the spike trains for some cells in this dataset.

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 (Granger, 1969; Seth et al., 2015). Granger causality is designed to determine when one variable is useful in predicting another. 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 (Equation 52) to test the connection from neuron j to neuron i (Kim et al., 2011). Other ongoing attempts to use maximum entropy methods have been reviewed (Fairhall et al., 2012). More generally, the issue of disambiguating direct interactions from interactions that occur through unobserved, or latent, variables is receiving increasing attention (Pfau et al., 2013; Vidne et al., 2012; Okun et al., 2015).

Discussion

We have presented and analyzed a class of methods that summarize the response properties of neurons in terms of one or a few feature vectors and an associated nonlinear input/output function (Table 2). These methods provide a principled means to describe neuronal spike responses. However, they 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 feature vectors and the nonlinearity serve as a basis to quantify changes in computation with context (Fairhall et al., 2001; Fairhall, 2013; Geffen et al., 2007), attention (Rabinowitz et al., 2015), learning (Shulz et al., 2000), and through neuromodulation (McCormick et al., 2015).

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

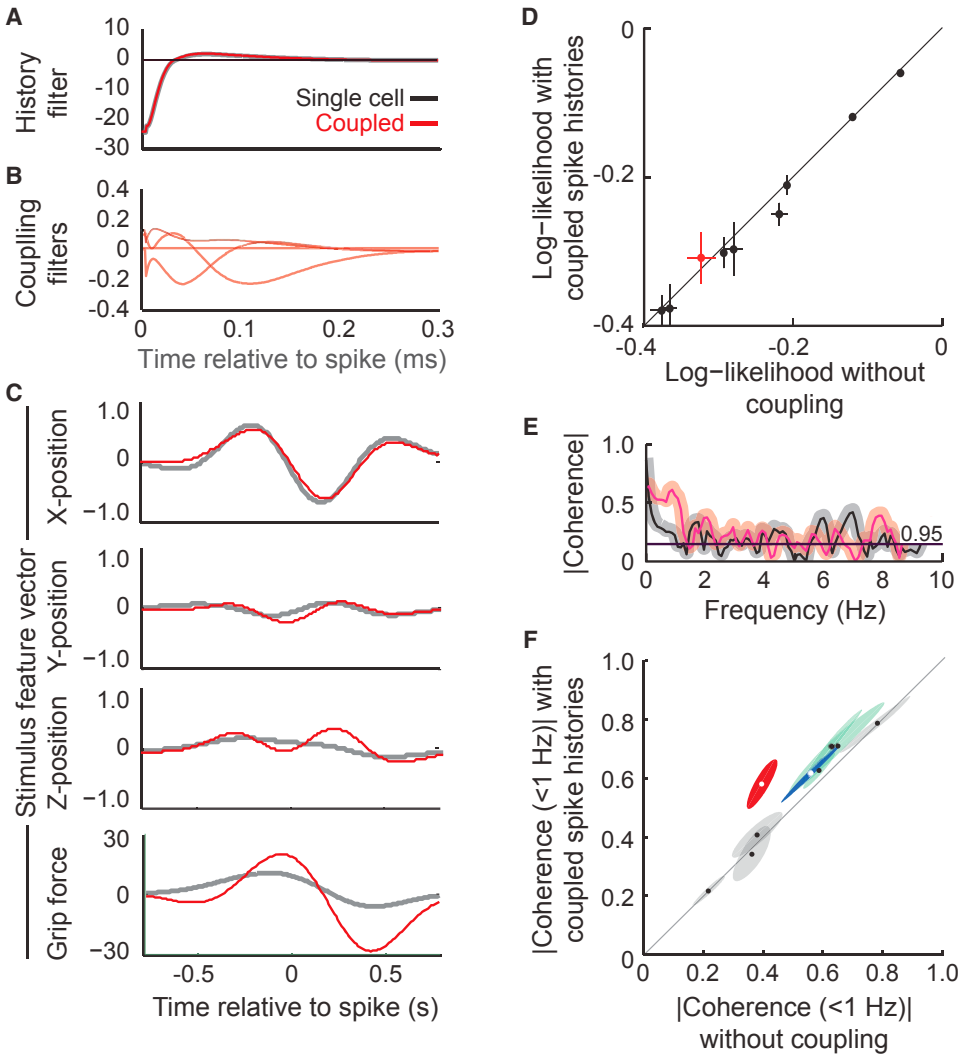


Figure 14. Network GLM Features and Validation for Monkey Reach Data Using the Interval between the Start of the Go Signal and the End of the Trial

(A) Spike history filter for sample unit 36, one of nine concurrently recorded 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. Red curve shows result for the coupled model ($\kappa = 100$ in 52), 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 ($\kappa = 100$) (Equations 52 and 53). The history filter was expanded in the basis of Equations 46 and 47 using $B = 5$, $t_0 = 0.005$ s, $t_1 = 0.4$ s, and $t_2 = 2.0$ s. The coupling terms are non-zero for three neighbors.

(C) Feature vectors calculated for the network, i.e., coupled (red), and single cell, i.e., uncoupled GLM (black).

(D) Scatterplot of log-likelihood between predicted spike rate and observed spike train for the coupled and uncoupled model. The red dot refers to the data in (A)–(C); the bars are one standard error jack-knife estimates.

(E) The spectral coherence, calculated as an average over 100 trials with a 0.5 Hz bandwidth, for the network, i.e., coupled (red), and single cell, i.e., uncoupled GLM (black). The band is one standard error.

(F) Scatterplot of the coherence between predicted spike rate and observed spike train for the coupled and uncoupled models. The ellipses are one standard error jack-knife estimates. The red ellipse refers to the data in (A)–(C) and (E). The red and blue data are significant at the level of three standard errors (0.01), the green data barely significant at two standard errors (0.05), and the gray data have no significant improvement from coupling to the network.

the model predictions. This results in an observed variance that is a continuous function of time and can be compared to a "model," in this case the observed mean rate; these values, of course, depend on the scale of smoothing applied to the data. Within early stages of the visual pathway, modeling based on repeated trials capture 80%–90% of the trial-to-trial variance in macaque retina (Pillow et al., 2008), 80%–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 (Equation 48) and the coherence (Equation 50) between the test spike train and the prediction. The log-likelihood, applied to test

Table 2. Summary of Methods

	STA	STC + STA	MNE	GLM
Number of stimulus feature vectors	One	Typically two or three, bounded by the stimulus dimension	Bounded by the stimulus dimension	One
History dependence	No		Yes	
Network interactions	No			Yes
Fitting method	Averaging and binning	Matrix diagonalization and binning	Optimization	
Nonlinearity	Derived from expectation or Bayes' procedure		Fixed as sigmoidal	Fixed as exponential
Binning of stimulus projections	Necessary but not problematic	Necessary but problematic for multiple dimensions	Not appropriate	
Convergence on training set	Guaranteed for elliptic distributions of stimuli with a non-zero second moment		Optimization converges as fitting is convex	
Overfitting	Not a problem with appropriate binning	Not a problem since nonlinearity is smoothed in light of sparse 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 timescales
Pioneering publication(s)	Marmarelis and Naka, 1972; Eckhorn and Pöpel, 1981	de Ruyter Van Steveninck and Bialek, 1988	Fitzgerald et al., 2011b for cells; Granot-Atedgi et al., 2013 for networks	Brown et al., 1998 for cells; Pillow et al., 2008 for networks

data (Figures 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 be a shallow function that does not clearly discriminate between predictions from models that are rather distinct (Figures 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 (Figures 10E, 11E, 12D, and 12I). 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 Figure 10A and Figures 10E and 11E).

The magnitude of the coherence 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. The normalization allows one to compare results between cells in addition to comparing models of the same cell. The coherence will be less than one because of variations that are independent of the stimulus and thus are not captured by the models, as well as because of nonstationary variations, such as changes in brain states (Goris et al., 2014; McCormick et al., 2015). Further, it is always possible that an improved model could perform better on the existing data; in this regard the reported coherence should be taken as a lower limit on the predictability of the spike pattern.

Caveats on Whitening

The 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 (Figure 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 (Figure 11). The 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).

Adaptation and Dependence on Stimulus Statistics

One significant issue with the fitting of LN models is that the resulting model, including feature vector, 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 likely to be altered through 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). These effects 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 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, e.g., activity-dependent changes in kinetic parameters (Ozuysal and Baccus, 2012).

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 that occur 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; Briggman et al., 2005; Churchland et al., 2010a, 2010b). 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., 2010a, 2010b). The second strategy aims to systematically model the multi-neuronal 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, similar to the MNE approach (Equation 32), one fits a maximum entropy distribution to the joint neural responses by choosing relevant constraints on the response, such as mean firing rates and correlations. Typically these methods do not yet provide a full mapping of input to output. 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).

Non-spiking Data

We have focused on 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 subthreshold flow of current in the extracellular space that is measured by field electrodes or by magnetoencephelography, while at the microscopic level this includes the subthreshold membrane potential as well as 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 number of spikes per sample time, $n_s(t)$, by the intensity of the sampled signal (Ramirez et al., 2014). The challenge arises in inferring exact spike times from such signals, which are needed to fit some models. 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).

The generalized integrate and fire method (Pozzorini et al., 2015) provides an extension of the GLM method to account for both spiking and subthreshold dynamics, as would be obtained from an intracellular measurement of membrane potential. The generalized integrate and fire method incorporates a term that filters the membrane potential as it evolves over time that is equivalent to the stimulus feature vector in the GLM. It further incorporates two spike history filters. One is equivalent to the spike history filter in the GLM, and the second is a new term that evolves over time and shifts the threshold of the spiking nonlinearity. This approach accounts for the subthreshold dynamics of the neuron yet bypasses complicated modeling of the active membrane currents.

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 starting point 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 do 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. 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 timescale 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 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*.

Implementation

All calculations were performed using MATLAB 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 (see Supplemental Information and, for updated versions, https://github.com/NeuroInfoPrimer): 53 salamander retina sets, 7 rat thalamic sets, and 9 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 MathWorks (http://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 (http://neurophysics.ucsd.edu/software.php), Partha Mitra's Chronux Toolbox (http://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 (https://www.cs.ubc.ca/~schmidtm/Software/thesis.html), and the multidimensional histogram function histcn.m downloaded at http://www.mathworks.com/matlabcentral/fileexchange/23897-n-dimensional-histogram/content/histcn.m.

SUPPLEMENTAL INFORMATION

Supplemental Information includes a zip file containing the MATLAB materials and data and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.05.039.

AUTHOR CONTRIBUTIONS

J.A., A.L.F., and D.K. planned the primer; J.A. and B.J.L. wrote the computer code; J.A., D.K., and B.J.L. prepared the figures; and all authors contributed to the writing of the manuscript.

ACKNOWLEDGMENTS

This Primer evolved from material presented at the "Methods in Computational Neuroscience" and "Neuroinformatics" summer schools at the Marine Biological Laboratory, the program on "Emerging Techniques in Neuroscience" at the Kavli Institute for Theoretical Physics, and courses taught at the University of California San Diego and the University of Washington. 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 assistance with the computer code. We further thank the anonymous reviewers for critical comments and corrections. This effort was supported by grants from the Allen Family Foundation (to A.L.F.), the NIH (NS058668 and NS090595 to D.K.), the NSF (CRCNS IIS-1430296 to J.A.; 0928251, EEC-1028725 to A.L.F.; and EAGER 2144GA to D.K.), and the U.S.-Israel Binational Science Foundation (855DBA to D.K. and Ehud Ahissar).

REFERENCES

Adelman, T.L., Bialek, W., and Olberg, R.M. (2003). The information content of receptive fields. Neuron *40*, 823–833.

Agüera y Arcas, B., and Fairhall, A.L. (2003). What causes a neuron to spike? Neural Comput. *15*, 1789–1807.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In Proceeding of the Second International Symposium on Information Theory, B.N. Petrov and F. Caski, eds. (Akademiai Kiado), pp. 267–281.

Alivisatos, A.P., Chun, M., Church, G.M., Greenspan, R.J., Roukes, M.L., and Yuste, R. (2012). The brain activity map project and the challenge of functional connectomics. Neuron *74*, 970–974.

Aljadeff, J., Segev, R., Berry, M.J., 2nd, and Sharpee, T.O. (2013). Spike triggered covariance in strongly correlated gaussian stimuli. PLoS Comput. Biol. *9*, e1003206.

Archer, E.W., Koster, U., Pillow, J.W., and Macke, J.H. (2014). Low-dimensional models of neural population activity in sensory cortical circuits. In Advances in Neural Information Processing Systems, pp. 343–351.

Atick, J.J. (1992). Could information theory provide an ecological theory of sensory processing? Network *3*, 213–251.

Berry, M.J., 2nd, and Meister, M. (1998). Refractoriness and neural precision. J. Neurosci. *18*, 2200–2211.

Bialek, W. and van Steveninck, R.R. (2005). Features and dimensions: Motion estimation in fly vision. arXiv, arXiv:q-bio/0505003, http://arxiv.org/abs/q-bio/0505003.

Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W., and Wells, M.T. (2014). Akaike's Information Criterion, Cp and Estimators of Loss for Elliptically Symmetric Distributions. International Statistical Review *82*, 422–439, http://dx.doi.org/10.1111/insr.12052.

Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. Neuron *26*, 695–702.

Briggman, K.L., Abarbanel, H.D., and Kristan, W.B., Jr. (2005). Optical imaging of neuronal populations during decision-making. Science *307*, 896–901.

Brown, K.D., and Balkany, T.J. (2007). Benefits of bilateral cochlear implantation: a review. Curr. Opin. Otolaryngol. Head Neck Surg. *15*, 315–318.

Brown, E.N., Frank, L.M., Tang, D., Quirk, M.C., and Wilson, M.A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. J. Neurosci. *18*, 7411–7425.

Campagner, D., Evans, M.H., Bale, M.R., Erskine, A., and Petersen, R.S. (2016). Prediction of primary somatosensory neuron activity during active tactile exploration. Elife *5*, http://dx.doi.org/10.7554/eLife.10696.

Chen, Z., Putrino, D.F., Ba, D.E., Ghosh, S., Barbieri, R., and Brown, E.N. (2009). A regularized point process generalized linear model for assessing the functional connectivity in the cat motor cortex. Conf. Proc. IEEE Eng. Med. Biol Soc. *2009*, 5006–5009.

Chichilnisky, E.J. (2001). A simple white noise analysis of neuronal light responses. Network *12*, 199–213.

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

Churchland, M.M., Yu, B.M., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., Scott, B.B., et al. (2010b). Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nat. Neurosci. *13*, 369–378.

Cohen, M.R., and Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nat. Neurosci. *14*, 811–819.

Cunningham, J.P., and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. *17*, 1500–1509.

Curtis, J.C., and Kleinfeld, D. (2009). Phase-to-rate transformations encode touch in cortical neurons of a scanning sensorimotor system. Nat. Neurosci. *12*, 492–501.

David, S.V., and Gallant, J.L. (2005). Predicting neuronal responses during natural vision. Network *16*, 239–260.

David, S.V., Vinje, W.E., and Gallant, J.L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. J. Neurosci. *24*, 6991–7006.

de Boer, E., and Kuyper, P. (1968). Triggered correlation. IEEE Trans Biomed Eng. *15*, 169–179.

de Ruyter Van Steveninck, R., and Bialek, W. (1988). 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. Series B, Biological Sciences *234*, 379–414.

Deschênes, M., Takatoh, J., Kurnikova, A., Moore, J.D., Demers, M., Elbaz, M., Furuta, T., Wang, F., and Kleinfeld, D. (2016). Inhibition, not excitation, drives rhythmic whisking. Neuron *90*, 374–387.

Díaz-Quesada, M., and Maravall, M. (2008). Intrinsic mechanisms for adaptive gain rescaling in barrel cortex. J. Neurosci. *28*, 696–710.

Eckhorn, R., and Pöpel, B. (1981). Responses of cat retinal ganglion cells to the random motion of a spot stimulus. Vision Res. *21*, 435–443.

Engelhard, B., Ozeri, N., Israel, Z., Bergman, H., and Vaadia, E. (2013). Inducing γ oscillations and precise spike synchrony by operant conditioning via brain-machine interface. Neuron *77*, 361–375.

Fairhall, A. (2013). Adaptation and natural stimulus statistics. In The Cognitive Neurosciences, M.S. Gazzaniga and G.R. Mangun, eds. (MIT Press), pp. 283–293.

Fairhall, A.L., Lewen, G.D., Bialek, W., and de Ruyter Van Steveninck, R.R. (2001). Efficiency and ambiguity in an adaptive neural code. Nature *412*, 787–792.

Fairhall, A.L., Burlingame, C.A., Narasimhan, R., Harris, R.A., Puchalla, J.L., and Berry, M.J., 2nd (2006). Selectivity for multiple stimulus features in retinal ganglion cells. J. Neurophysiol. *96*, 2724–2738.

Fairhall, A., Shea-Brown, E., and Barreiro, A. (2012). Information theoretic approaches to understanding circuit function. Curr. Opin. Neurobiol. *22*, 653–659.

Famulare, M., and Fairhall, A. (2010). Feature selection in simple neurons: how coding depends on spiking dynamics. Neural Comput. *22*, 581–598.

Fitzgerald, J.D., Rowekamp, R.J., Sincich, L.C., and Sharpee, T.O. (2011a). Second order dimensionality reduction using minimum and maximum mutual information models. PLoS Comput. Biol. *7*, e1002249.

Fitzgerald, J.D., Sincich, L.C., and Sharpee, T.O. (2011b). Minimal models of multidimensional computations. PLoS Comput. Biol. *7*, e1001111.

Fox, J.L., Fairhall, A.L., and Daniel, T.L. (2010). Encoding properties of haltere neurons enable motion feature detection in a biological gyroscope. Proc. Natl. Acad. Sci. USA *107*, 3840–3845.

Ganmor, E., Segev, R., and Schneidman, E. (2011). Sparse low-order interaction network underlies a highly correlated and learnable neural population code. Proc. Natl. Acad. Sci. USA *108*, 9679–9684.

Gaudry, K.S., and Reinagel, P. (2007). Contrast adaptation in a nonadapting LGN model. J. Neurophysiol. *98*, 1287–1296.

Geffen, M.N., de Vries, S.E.J., and Meister, M. (2007). Retinal ganglion cells can rapidly change polarity from Off to On. PLoS Biology *5*, e65.

Gerhard, F., Kispersky, T., Gutierrez, G.J., Marder, E., Kramer, M., and Eden, U. (2013). Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. PLoS Comput. Biol. *9*, e1003138.

Globerson, A., Stark, E., Vaadia, E., and Tishby, N. (2009). The minimum information principle and its application to neural code analysis. Proc. Natl. Acad. Sci. USA *106*, 3490–3495.

Gollisch, T., and Meister, M. (2008). Modeling convergent ON and OFF pathways in the early visual system. Biol. Cybern. *99*, 263–278.

Golomb, D., Kleinfeld, D., Reid, R.C., Shapley, R.M., and Shraiman, B.I. (1994). On temporal codes and the spatiotemporal response of neurons in the lateral geniculate nucleus. J. Neurophysiol. *72*, 2990–3003.

Goris, R.L.T., Movshon, J.A., and Simoncelli, E.P. (2014). Partitioning neuronal variability. Nat. Neurosci. *17*, 858–865.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica *37*, 424–438.

Granot-Atedgi, E., Tkačik, G., Segev, R., and Schneidman, E. (2013). Stimulus-dependent maximum entropy models of neural population codes. PLoS Comput. Biol. *9*, e1002922.

Grienberger, C., and Konnerth, A. (2012). Imaging calcium in neurons. Neuron *73*, 862–885.

Hagiwara, S. (1954). Analysis of interval fluctuation of the sensory nerve impulse. Jpn. J. Physiol. *4*, 234–240.

van Hateren, J.H., and van der Schaaf, A. (1998). 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*, 359–366.

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

Hill, D.N., Mehta, S.B., and Kleinfeld, D. (2011b). Quality metrics to accompany spike sorting of extracellular signals. J. Neurosci. *31*, 8699–8705.

Hong, S., Lundstrom, B.N., and Fairhall, A.L. (2008). Intrinsic gain modulation and adaptive neural coding. PLoS Comput. Biol. *4*, e1000119.

Hosoya, T., Baccus, S.A., and Meister, M. (2005). Dynamic predictive coding by the retina. Nature *436*, 71–77.

Jones, L.M., Depireux, D.A., Simons, D.J., and Keller, A. (2004). Robust temporal coding in the trigeminal system. Science *304*, 1986–1989.

Kaardal, J., Fitzgerald, J.D., Berry, M.J., 2nd, and Sharpee, T.O. (2013). Identifying functional bases for multidimensional neural computations. Neural Comput. *25*, 1870–1890.

Khouri, L., and Nelken, I. (2015). Detecting the unexpected. Curr. Opin. Neurobiol. *35*, 142–147.

Kim, S., Putrino, D., Ghosh, S., and Brown, E.N. (2011). A Granger causality measure for point process models of ensemble neural spiking activity. PLoS Comput. Biol. *7*, e1001110.

Kleinfeld, D., and Deschênes, M. (2011). Neuronal basis for object location in the vibrissa scanning sensorimotor system. Neuron *72*, 455–468.

Kleinfeld, D., and Griesbeck, O. (2005). From art to engineering? The rise of in vivo mammalian electrophysiology via genetically targeted labeling and nonlinear imaging. PLoS Biol. *3*, e355.

Kleinfeld, D., and Mitra, P.P. (2011). Applications of spectral methods in functional brain imaging. In Imaging: A Laboratory Manual, R. Yuste, ed. (Cold Spring Harbor Laboratory Press), pp. 12.11–12.17.

Kleinfeld, D., Ahissar, E., and Diamond, M.E. (2006). Active sensation: insights from the rodent vibrissa sensorimotor system. Curr. Opin. Neurobiol. *16*, 435–444.

Kulkarni, J.E., and Paninski, L. (2007). Common-input models for multiple neural spike-train data. Network *18*, 375–407.

Lee, D.N., and Kalmus, H. (1980). The optic flow field: the foundation of vision [and discussion]. Philos. Trans. R. Soc. Lond. B Biol. Sci. *290*, 169–179.

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

Linden, J.F., Liu, R.C., Sahani, M., Schreiner, C.E., and Merzenich, M.M. (2003). Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J. Neurophysiol. *90*, 2660–2675.

Lütcke, H., Gerhard, F., Zenke, F., Gerstner, W., and Helmchen, F. (2013). Inference of neuronal network spike dynamics and topology from calcium imaging data. Front. Neural Circuits *7*, 201.

Magee, J.C. (2003). A prominent role for intrinsic neuronal properties in temporal coding. Trends Neurosci. *26*, 14–16.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In Proceedings of the 6th Conference on Natural Language Learning, *Volume 20*, pp. 1–7.

Maravall, M., Petersen, R.S., Fairhall, A.L., Arabzadeh, E., and Diamond, M.E. (2007). Shifts in coding properties and maintenance of information transmission during adaptation in barrel cortex. PLoS Biol. *5*, e19.

Marčenko, V.A., and Pastur, L.A. (1967). Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik *1*, 457–483. http://stacks.iop.org/0025-5734/1/i=4/a=A01.

Marmarelis, P.Z., and Marmarelis, V.Z. (1978). The White-Noise Method in System Identification. In Analysis of Physiological Systems (Springer), pp. 131–180.

Marmarelis, P.Z., and Naka, K. (1972). White-noise analysis of a neuron chain: an application of the Wiener theory. Science *175*, 1276–1278.

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

McCormick, D.A., McGinley, M.J., and Salkoff, D.B. (2015). Brain state dependent activity in the cortex and thalamus. Curr. Opin. Neurobiol. *31*, 133–140.

McFarland, J.M., Cui, Y., and Butts, D.A. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. PLoS Comput. Biol. *9*, e1003143.

Mease, R.A., Famulare, M., Gjorgjieva, J., Moody, W.J., and Fairhall, A.L. (2013). Emergence of adaptive computation by single neurons in the developing cortex. J. Neurosci. *33*, 12154–12170.

Mease, R.A., Lee, S., Moritz, A.T., Powers, R.K., Binder, M.D., and Fairhall, A.L. (2014). Context-dependent coding in single neurons. J. Comput. Neurosci. *37*, 459–480.

Merfeld, D.M., and Lewis, R.F. (2012). Replacing semicircular canal function with a vestibular implant. Curr. Opin. Otolaryngol. Head Neck Surg. *20*, 386–392.

Moore, J.D., Deschênes, M., and Kleinfeld, D. (2015a). Juxtacellular monitoring and localization of single neurons within sub-cortical brain structures of alert, head-restrained rats. J. Vis. Exp. (98).

Moore, J.D., Mercer Lindsay, N., Deschênes, M., and Kleinfeld, D. (2015b). Vibrissa self-motion and touch are reliably encoded along the same somatosensory pathway from brainstem through thalamus. PLoS Biol. *13*, e1002253.

Nandy, A.S., and Tjan, B.S. (2012). Saccade-confounded image statistics explain visual crowding. Nat. Neurosci. *15*, 463–469, S1–S2.

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

Nelson, M.E., and MacIver, M.A. (2006). Sensory acquisition in active sensing systems. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. *192*, 573–586.

Nirenberg, S., and Pandarinath, C. (2012). Retinal prosthetic strategy with the capacity to restore normal vision. Proc. Natl. Acad. Sci. USA *109*, 15012–15017.

Okun, M., Steinmetz, N.A., Cossell, L., Iacaruso, M.F., Ko, H., Barthó, P., Moore, T., Hofer, S.B., Mrsic-Flogel, T.D., Carandini, M., and Harris, K.D. (2015). Diverse coupling of neurons to populations in sensory cortex. Nature *521*, 511–515.

Olshausen, B.A., and Field, D.J. (2005). How close are we to understanding v1? Neural Comput. *17*, 1665–1699.

Ozuysal, Y., and Baccus, S.A. (2012). Linking the computational structure of variance adaptation to biophysical mechanisms. Neuron *73*, 1002–1015.

Pang, R., Lansdell, B.J.L., and Fairhall, A.L. (2016). Dimensionality reduction in neuroscience. Curr. Biol. *26*, R1–R5.

Paninski, L. (2003). Convergence properties of three spike-triggered analysis techniques. Network *14*, 437–464.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. Network *15*, 243–262.

Pasupathy, A., and Connor, C.E. (2002). Population coding of shape in area V4. Nat. Neurosci. *5*, 1332–1338.

Penrose, R. (1955). A generalized inverse for matrices. In Proceedings of the Cambridge Philosophical Society, *Volume 51* (Cambridge University Press), pp. 406–413.

Pfau, D., Pnevmatikakis, E.A., and Paninski, L. (2013). Robust learning of low-dimensional dynamics from large neural ensembles. In Advances in Neural Information Processing Systems, pp. 2391–2399.

Pillow, J.W., Paninski, L., Uzzell, V.J., Simoncelli, E.P., and Chichilnisky, E.J. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. J. Neurosci. *25*, 11003–11013.

Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., and Simoncelli, E.P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature *454*, 995–999.

Podvigin, N.F., Cooperman, A.M., and Tchueva, I.V. (1974). The space-time properties of excitation and inhibition and wave processes in cat's corpus geniculatum lateralis. Biophysics *19*, 341–346.

Powers, R.K., and Binder, M.D. (1996). Experimental evaluation of input-output models of motoneuron discharge. J. Neurophysiol. *75*, 367–379.

Pozzorini, C., Mensi, S., Hagens, O., Naud, R., Koch, C., and Gerstner, W. (2015). Automated High-Throughput Characterization of Single Neurons by Means of Simplified Spiking Models. PLoS Comput. Biol. *11*, e1004275.

Prescott, T.J., Diamond, M.E., and Wing, A.M. (2011). Active touch sensing. Philos. Trans. R. Soc. Lond. B Biol. Sci. *366*, 2989–2995.

Rabinowitz, N.C., Goris, R.L., Cohen, M., and Simoncelli, E.P. (2015). Attention stabilizes the shared gain of V4 populations. eLife *4*, e08998, http://dx.doi.org/10.7554/eLife.08998.

Ramirez, A., Pnevmatikakis, E.A., Merel, J., Paninski, L., Miller, K.D., and Bruno, R.M. (2014). Spatiotemporal receptive fields of barrel cortex revealed by reverse correlation of synaptic input. Nat. Neurosci. *17*, 866–875.

Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., and Ballard, D.H. (2002). Eye movements in iconic visual search. Vision Res. *42*, 1447–1463.

Rowekamp, R.J., and Sharpee, T.O. (2011). Analyzing multicomponent receptive fields from neural responses to natural stimuli. Network *22*, 45–73.

Rudd, M.E., and Brown, L.G. (1997). Noise adaptation in integrate-and fire neurons. Neural Comput. *9*, 1047–1069.

Ruderman, D.L., and Bialek, W. (1994). Statistics of natural images: Scaling in the woods. Phys. Rev. Lett. *73*, 814–817.

Rust, N.C., Schwartz, O., Movshon, J.A., and Simoncelli, E.P. (2005). Spatiotemporal elements of macaque v1 receptive fields. Neuron *46*, 945–956.

Saleh, M., Takahashi, K., and Hatsopoulos, N.G. (2012). Encoding of coordinated reach and grasp trajectories in primary motor cortex. J. Neurosci. *32*, 1220–1232.

Samengo, I., and Gollisch, T. (2013). Spike-triggered covariance: geometric proof, symmetry properties, and extension beyond Gaussian stimuli. J. Comput. Neurosci. *34*, 137–161.

Schneidman, E., Berry, M.J., 2nd, Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. Nature *440*, 1007–1012.

Schroeder, C.E., Wilson, D.A., Radman, T., Scharfman, H., and Lakatos, P. (2010). Dynamics of Active Sensing and perceptual selection. Curr. Opin. Neurobiol. *20*, 172–176.

Schwartz, O., Pillow, J.W., Rust, N.C., and Simoncelli, E.P. (2006). Spike-triggered neural characterization. J. Vis. *6*, 484–507.

Segev, R., Goodhouse, J., Puchalla, J., and Berry, M.J., 2nd (2004). Recording spikes from a large fraction of the ganglion cells in a retinal patch. Nat. Neurosci. *7*, 1154–1161.

Segev, R., Puchalla, J., and Berry, M.J., 2nd (2006). Functional organization of ganglion cells in the salamander retina. J. Neurophysiol. *95*, 2277–2292.

Seth, A.K., Barrett, A.B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. J. Neurosci. *35*, 3293–3297.

Sharpee, T.O. (2013). Computational identification of receptive fields. Annu. Rev. Neurosci. *36*, 103–120.

Sharpee, T., Rust, N.C., and Bialek, W. (2003). Maximally informative dimensions: analyzing neural responses to natural signals. Adv. Neural Inf. Process. Syst. 277–284.

Sharpee, T., Rust, N.C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput. *16*, 223–250.

Sharpee, T.O., Sugihara, H., Kurgansky, A.V., Rebrik, S.P., Stryker, M.P., and Miller, K.D. (2006). Adaptive filtering enhances information transmission in visual cortex. Nature *439*, 936–942.

Sharpee, T.O., Miller, K.D., and Stryker, M.P. (2008). On the importance of static nonlinearity in estimating spatiotemporal neural filters with natural stimuli. J. Neurophysiol. *99*, 2496–2509.

Shoham, S., Paninski, L.M., Fellows, M.R., Hatsopoulos, N.G., Donoghue, J.P., and Normann, R. (2005). Statistical encoding model for a primary motor cortical brain-machine interface. IEEE Trans. Biomed. Eng. *52*, 1312–1322.

Shulz, D.E., Sosnik, R., Ego, V., Haidarliu, S., and Ahissar, E. (2000). A neuronal analogue of state-dependent learning. Nature *403*, 549–553.

Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annu. Rev. Neurosci. *24*, 1193–1216.

Slee, S.J., Higgs, M.H., Fairhall, A.L., and Spain, W.J. (2005). Two-dimensional time coding in the auditory brainstem. J. Neurosci. *25*, 9978–9988.

Softky, W.R., and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J. Neurosci. *13*, 334–350.

Sokal, R.R., and Rohlf, F.J. (1995). The principles and practice of statistics in biological research, Third Edition (WH Freeman).

Srinivasan, M.V., Laughlin, S.B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. Lond. B Biol. Sci. *216*, 427–459.

Stevenson, I.H., London, B.M., Oby, E.R., Sachs, N.A., Reimer, J., Englitz, B., David, S.V., Shamma, S.A., Blanche, T.J., and Mizuseki, K. (2012). Functional connectivity and tuning curves in populations of simultaneously recorded neurons. PLoS Comput. Biol. *8*, e1002775.

Svoboda, K., Denk, W., Kleinfeld, D., and Tank, D.W. (1997). In vivo dendritic calcium dynamics in neocortical pyramidal neurons. Nature *385*, 161–165.

Szulborski, R.G., and Palmer, L.A. (1990). The two-dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. Vision Res. *30*, 249–254.

Tabot, G.A., Dammann, J.F., Berg, J.A., Tenore, F.V., Boback, J.L., Vogelstein, R.J., and Bensmaia, S.J. (2013). Restoring the sense of touch with a prosthetic hand through a brain interface. Proc. Natl. Acad. Sci. USA *110*, 18279–18284.

Takahashi, K., Pesce, L., Iriarte-Diaz, J., Best, M., Kim, S., Coleman, T.P., Hatsopoulos, N.G., and Ross, C.F. (2012). Granger causality analysis of state dependent functional connectivity of neurons in orofacial motor cortex during chewing and swallowing. In 2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), pp. 1067–1071.

Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network *12*, 289–316.

Thomson, D.J. (1982). Spectrum estimation and harmonic analysis. Proc. IEEE *70*, 1055–1096.

Touryan, J., Lau, B., and Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. J. Neurosci. *22*, 10811–10818.

Trenholm, S., and Roska, B. (2014). Cell-type-specific electric stimulation for vision restoration. Neuron *83*, 1–2.

Truccolo, W., Eden, U.T., Fellows, M.R., Donoghue, J.P., and Brown, E.N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. J. Neurophysiol. *93*, 1074–1089.

Vidal-Gadea, A.G., and Belanger, J.H. (2009). Muscular anatomy of the legs of the forward walking crab, Libinia emarginata (Decapoda, Brachyura, Majoidea). Arthropod Struct. Dev. *38*, 179–194.

Vidne, M., Ahmadian, Y., Shlens, J., Pillow, J.W., Kulkarni, J., Litke, A.M., Chichilnisky, E.J., Simoncelli, E., and Paninski, L. (2012). Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. J. Comput. Neurosci. *33*, 97–121.

Vogelstein, J.T., Packer, A.M., Machado, T.A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. J. Neurophysiol. *104*, 3691–3704.

Wark, B., Lundstrom, B.N., and Fairhall, A. (2007). Sensory adaptation. Curr. Opin. Neurobiol. *17*, 423–429.

Werner, G., and Mountcastle, V.B. (1963). The variability of central neural activity in a sensory system, and its implications for the central reflection of sensory events. Journal of Neurophysiology *26*, 958–977.

Yu, B.M., Cunningham, J.P., Santhanam, G., Ryu, S.I., Shenoy, K.V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J. Neurophysiol. *102*, 614–635.

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