Dimensionality reduction in neural models:
An information-theoretic generalization of
spike-triggered average and covariance analysis

Jonathan W. Pillow

Gatsby Computational Neuroscience Unit,
University College London, London, UK

Eero P. Simoncelli

Howard Hughes Medical Institute and Center for Neural Science,
New York University, NY, USA

We describe an information-theoretic framework for fitting neural spike responses with a Linear–Nonlinear–Poisson cascade model. This framework unifies the spike-triggered average (STA) and spike-triggered covariance (STC) approaches to neural characterization and recovers a set of linear filters that maximize mean and variance-dependent information between stimuli and spike responses. The resulting approach has several useful properties, namely, (1) it recovers a set of linear filters sorted according to their informativeness about the neural response; (2) it is both computationally efficient and robust, allowing recovery of multiple linear filters from a data set of relatively modest size; (3) it provides an explicit "default" model of the nonlinear stage mapping the filter responses to spike rate, in the form of a ratio of Gaussians; (4) it is equivalent to maximum likelihood estimation of this default model but also converges to the correct filter estimates whenever the conditions for the consistency of STA or STC analysis are met; and (5) it can be augmented with additional constraints on the filters, such as space–time separability. We demonstrate the effectiveness of the method by applying it to simulated responses of a Hodgkin–Huxley neuron and the recorded extracellular responses of macaque retinal ganglion cells and V1 cells.

Keywords: neural coding, white noise analysis, reverse correlation, receptive field, information theory, neural modeling

Introduction

One of the central problems in sensory neuroscience is that of characterizing the neural code, or the mapping from sensory stimuli to neural spike responses. The problem has been investigated in a large number of sensory areas, using a variety of specialized stimuli and experimental preparations, but a general solution is intractable due to the high dimensionality of both the stimulus and response spaces (Dayan & Abbott, 2001; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997).

Recent work has explored "dimensionality-reduction" methods to simplify the problem of modeling the neural response (Aguera y Arcas & Fairhall, 2003; Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Brenner, Bialek, & de Ruyter van Steveninck, 2000; de Ruyter van Steveninck & Bialek, 1988; Paninski, 2003; Ringach, Sapiro, & Shapley, 1997; Rust, Schwartz, Movshon, & Simoncelli, 2005; Schwartz, Chichilnisky, & Simoncelli, 2002; Sharpee, Rust, & Bialek, 2004; Sharpee et al., 2006; Touryan, Lau, & Dan, 2002). The concept is intuitive and sensible: Although the space of all stimuli is enormous (e.g., the space of all images), most attributes of these stimuli do not have any effect on a given neuron's response. If we can identify a low-dimensional space in which a neuron computes its response (sometimes called a "feature space") then the neural code can be characterized by describing

responses only within that space. Note that "classical" experiments can also be viewed within this framework: Characterization with dots, bars, or grating stimuli implicitly assumes that a neuron's behavior is determined by its response to a set of canonical features.

Two basic approaches have been developed to estimate a neural feature space. The first compares the mean and covariance of the spike-triggered stimulus ensemble (i.e., the set of stimuli that elicited a spike from the neuron) with those of the full stimulus ensemble (Bialek et al., 1991; de Ruyter van Steveninck & Bialek, 1988). Significant changes in either the spike-triggered average (STA) or spike-triggered covariance (STC) can be used to determine the subspace in which the neuron computes its response. These methods are relatively efficient to compute and have been demonstrated on the H1 neuron of the fly (Bialek & de Ruyter van Steveninck, 2005; Bialek et al., 1991; Brenner et al., 2000; de Ruyter van Steveninck & Bialek, 1988), macaque retinal ganglion cells (RGCs; Schwartz et al., 2002), and V1 cells in both cat (Touryan et al., 2002) and monkey (Rust et al., 2005). These methods are appealing because of their simplicity, but they do not consider information from joint changes in mean and variance, and they provide no absolute measure of importance of the filters recovered. In addition, although filters can be recovered for subspaces of arbitrary dimensionality, the estimation of the nonlinear mapping from filter responses to firing rate becomes intractable for subspaces of more than a few dimensions.

A second approach to dimensionality reduction is to seek "maximally informative dimensions," along which the mutual information between stimulus and response is maximized (Paninski, 2003; Sharpee et al., 2004). This approach makes no explicit use of the STA or STC and has the advantage that it is sensitive (in principle) to statistical changes of any order. Unlike STA and STC analysis, it also can provide consistent estimates with nonspherical and non-Gaussian (e.g., naturalistic) stimuli. Unfortunately, accurate estimation of mutual information requires a large amount of data, and the mutual information function is rife with local maxima, making reliable automated optimization difficult to perform. As a result, these techniques are often impractical in high dimensions, and published examples have been restricted to the recovery of one-dimensional (1D) or two-dimensional (2D) feature spaces (Paninski, 2003; Sharpee et al., 2004; Sharpee et al., 2006).

In this article, we describe a new method for dimensionality reduction that occupies a middle ground between the moment-based and information-theoretic approaches. Specifically, we maximize information based only on the first and second moments of the spike-triggered stimulus ensemble. This approach provides a unifying information-theoretic framework for STA/STC analysis but remains computationally tractable even in feature spaces of relatively high dimensionality. The method is sensitive to interactions between the STA and STC components and provides an implicit or "default" model of the nonlinear function mapping the feature space to the neural response. We demonstrate the method on simulated data from a Hodgkin–Huxley (HH) model neuron, as well as physiological data from a macaque V1 neuron and a macaque retinal ganglion cell (RGC). Finally, we show an application of the framework for estimating a model with space–time separable components, which cannot be easily achieved with other estimators.

Spike-triggered analysis

Figure 1 shows the elements of a typical white noise experiment and illustrates how dimensionality reduction can be understood as a tool for fitting a neural encoding model. Figure 1A depicts a discrete Gaussian white noise stimulus (flickering bars), and Figure 1B shows a 2D representation of this stimulus (space vs. time), along with the simulated neural response. The first step of the analysis involves identifying the "spike-triggered stimulus ensemble," the collection of stimuli associated with spikes. We assume that each spike is causally associated with a "chunk" of the space–time stimulus and that spikes are generated according to a Poisson process whose rate is governed entirely by the preceding stimulus chunk, independent of the times of previous spikes.

In general, methods for dimensionality reduction of neural models proceed by looking for a linear subspace that best captures the statistical differences between the spike-

triggered ensemble and the "raw" stimulus ensemble, the collection of *all* stimulus vectors (Bialek & de Ruyter van Steveninck, 2005; Bialek et al., 1991; de Ruyter van Steveninck & Bialek, 1988; Simoncelli, Paninski, Pillow, & Schwartz, 2004). Figure 1C shows an example where we have reduced the stimulus to a 1D subspace by linear projection (i.e., by filtering the stimulus with a single linear filter). Within this subspace, the mean of the STE is significantly higher and its variance is significantly lower than that of the raw stimulus ensemble. These differences indicate that position along this axis in stimulus space (i.e., the response of this linear filter) carries information about the probability that the neuron will spike.

The neural response model is denoted $P(\text{spike}|\mathbf{x})$: the probability that a neuron will elicit a spike in response to a stimulus $\mathbf{x}$. As illustrated in Figures 1C and D, we can compute this probability directly using Bayes rule:

$$P(\text{spike}|\mathbf{x}) = \alpha \frac{P(\mathbf{x}|\text{spike})}{P(\mathbf{x})}, \tag{1}$$

where α is a constant proportional to the probability that a spike occurs, $P(\text{spike})$. The encoding model can therefore be computed as the ratio of two probability distributions, and in the simple cases (e.g., Figure 1), we can estimate it directly as a quotient of two histograms.

Unfortunately, the direct approach fails in high dimensions because of the so-called "curse of dimensionality": As the stimulus dimensionality increases, the amount of data in each histogram bin falls exponentially. In these cases, we proceed by assuming that the neuron is insensitive to a large number of dimensions of the stimulus space, meaning that $P(\mathbf{x}|\text{spike})$ and $P(x)$ do not differ except within a relatively low dimensional subspace. Our first step is thus to find the subspace that best captures these differences. We formalize this as the search for a matrix $\mathbf{B}$ for which the true conditional probability of spiking is closely approximated by the conditional probability within the subspace spanned by the columns of $\mathbf{B}$:

$$P(\text{spike}|\mathbf{x}) \approx P(\text{spike}|\mathbf{B}^T\mathbf{x}). \tag{2}$$

This dimensionality-reduction step can be regarded as the first step in fitting a Linear–Nonlinear–Poisson (LNP) model of the neural response. This model consists of (1) a bank of linear filters (i.e., the columns of $\mathbf{B}$); (2) a nonlinear combination rule, which converts the filter outputs ($\mathbf{B}^T\mathbf{x}$) to an instantaneous probability of spiking; and (3) inhomogeneous Poisson spiking.

Dimensionality reduction with STA and STC

The first and second moments of the STE can be used to identify a subspace that is informative about the neural
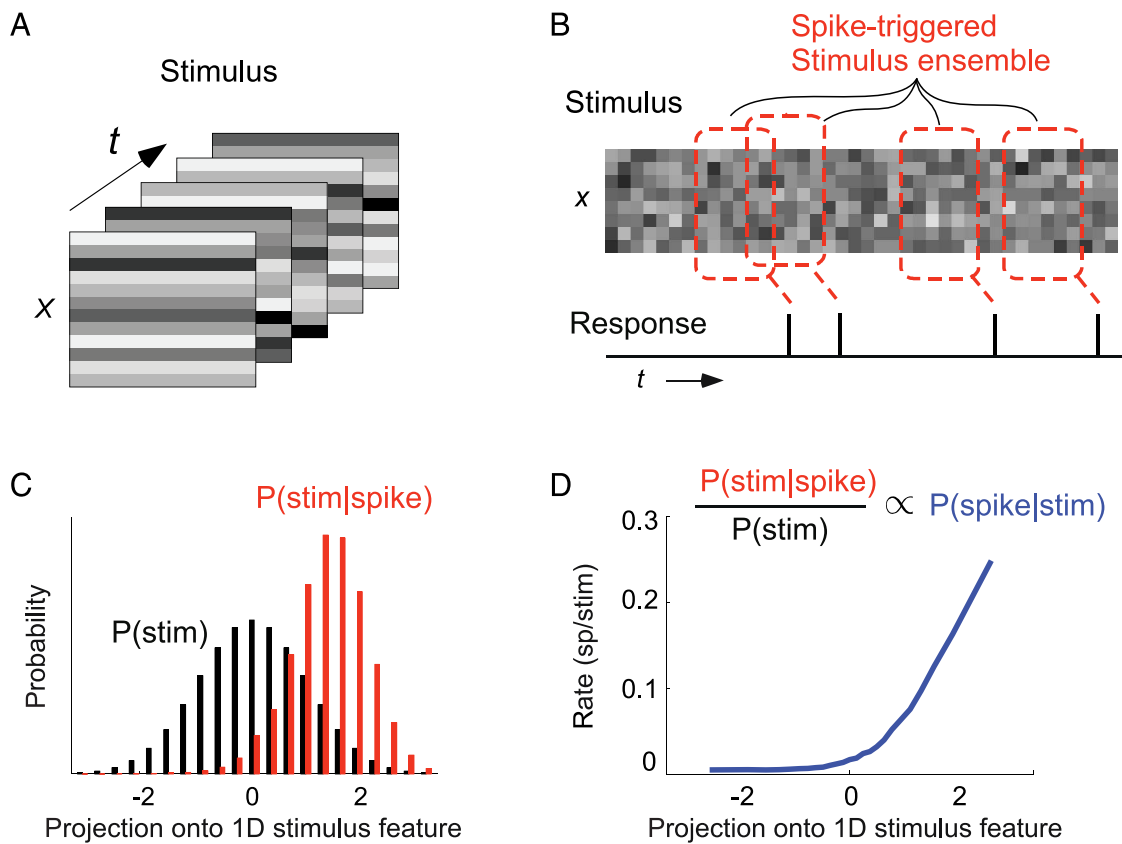
Figure 1. Illustration of spike-triggered stimulus ensemble and dimensionality reduction for a 1D neural model. (A) Depiction of a discretized white noise stimulus (e.g., flickering bars whose intensities are drawn from a Gaussian distribution on each temporal frame). (B) Samples from the spike-triggered stimulus ensemble (red boxes), vectors consisting of a space–time stimulus "chunk" preceding each spike. Here, the chunks have a duration of six frames and a spatial extent of eight bars. Each element of the spike-triggered stimulus ensemble is therefore a 48-dimensional vector. (C) By projecting the raw stimulus and the spike-triggered stimulus onto a single axis in the stimulus space, we can empirically measure the probability distributions of the raw (black) and spike-triggered (red) stimuli. (D) The 1D response model is specified by the probability of spiking conditioned on the stimulus projection along this axis, which (according to Bayes' rule) is just the ratio of the spike-triggered and raw distributions.

response. Specifically, if we assume that $P(\mathbf{x})$ has zero mean, then the STA,

$$\mu = \frac{1}{n_{\text{sp}}} \sum_{\{\mathbf{x}_i | \text{spike}\}} \mathbf{x}_i, \qquad (3)$$

gives the direction in the stimulus space along which the means of $P(\mathbf{x}|\text{spike})$ and $P(\mathbf{x})$ differ most. Similarly, the STC,

$$\Lambda = \frac{1}{n_{\text{sp}}} \sum_{\{\mathbf{x}_i | \text{spike}\}} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T, \qquad (4)$$

can be used to find the directions in the stimulus space along which the variance of $P(\mathbf{x}|\text{spike})$ and $P(\mathbf{x})$ differ maximally.

STA/STC analysis can provide a basis **B** for a reduced-dimensional model of the neural response consisting of the STA (if it differs significantly from zero) and the eigenvectors of the STC whose associated eigenvalues differ significantly from the variance of the raw ensemble. We will refer to this latter group as the "significant" eigenvectors of the STC or simply "STC axes." For spike responses generated by an LNP model, the STA and STC axes converge asymptotically to span the correct subspace (i.e., the subspace associated with **B**) if the raw stimulus distribution $P(\mathbf{x})$ is Gaussian and the instantaneous nonlinearity induces a change in the mean, variance, or both along each dimension of this subspace (Bialek & de Ruyter van Steveninck, 2005; Bussgang, 1952; Paninski, 2003).

Although the STA/STC method is simple and efficient to compute, it has several important drawbacks. Firstly, the STA axis is typically not orthogonal to the STC axes and may even lie within the span of the STC axes. Although the STC can be orthogonalized with respect to the STA (e.g., Rust et al., 2005; Schwartz et al., 2002; Simoncelli et al., 2004), this runs the risk of losing information, as we will show in the following section. Secondly, STA/STC

analysis does not easily allow us to quantify information or to know which axes of the subspace are most informative (the STA, the large-variance STC axes, or the small-variance STC axes). Finally, the basic STA/STC methodology does not specify a general means for estimating the nonlinear function taking filter outputs to spike rate. Although we can easily compute the nonlinearity along any single dimension using a ratio of two histograms (as demonstrated in Figures 1C and D), this approach is impractical for subspaces with more than a few dimensions.

Information-theoretic approach

These considerations motivate an information-theoretic framework for dimensionality reduction, based on the information contained only in the mean and covariance of the STE. Specifically, we would like to define a single objective function that incorporates the information each of these moments provides about the neural response. If we are agnostic about higher order moments, the simplest assumption we can make is that the STE is Gaussian, with mean and covariance given by the STA and STC, respectively; this is the maximum-entropy density for given mean and variance (Cover & Thomas, 1991; Levine & Tribus, 1978). A minimal-assumption model of the STE, or $P(\mathbf{x}|\text{spike})$, can therefore be written as

$$Q(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Lambda}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Lambda}^{-1}(\mathbf{x}-\mu)}, \qquad (5)$$

where n is the dimensionality of the stimulus space.

A natural choice for measuring statistical differences between the spike-triggered ensemble and the raw ensemble is the Kullback–Leibler (KL) divergence, an information-theoretic measure of the difference between two probability distributions (Cover & Thomas, 1991):

$$D(Q,P) = \int_{\mathcal{R}^n} Q(\mathbf{x}) \log \frac{Q(\mathbf{x})}{P(\mathbf{x})} d\mathbf{x}, \qquad (6)$$

where P represents the distribution of the raw stimuli.

In the present case, we use Gaussian approximations for both P and Q. P is assumed to have zero mean and has identity covariance or can be made so by subtracting the mean and "whitening" the stimulus space according to $\mathbf{x} = \mathbf{\Lambda}_0^{-\frac{1}{2}}(\mathbf{x}_0 - \mu_0)$, where μ_0 and $\mathbf{\Lambda}_0$ are the mean and covariance of the original stimulus distribution $P(\mathbf{x}_0)$. Under these assumptions, Equation 6 reduces to

$$D(Q,P) = \frac{1}{2}\big(\text{Tr}(\mathbf{\Lambda}) - \log|\mathbf{\Lambda}| + \mu^T\mu - n\big), \qquad (7)$$

where $\text{Tr}(\cdot)$ and $|\cdot|$ indicate matrix trace and determinant, respectively. The KL divergence between P and Q within a given subspace is given by:

$$D_{[\mathbf{B}]}(Q,P) = \frac{1}{2}\big(\text{Tr}\big[\mathbf{B}^T(\mathbf{\Lambda} + \mu\mu^T)\mathbf{B}\big] - \log|\mathbf{B}^T\mathbf{\Lambda}\mathbf{B}| - m\big), \qquad (8)$$

where $\mathbf{B}$ is a matrix whose m columns form an orthonormal basis for the subspace.

The most informative subspace, therefore, is given by the matrix $\mathbf{B}$ that maximizes Equation 8. If we want a 1D subspace, this objective function reduces to

$$D_{[\mathbf{b}]}(Q,P) = \frac{1}{2}\big(\mathbf{b}^T\mathbf{\Lambda}\mathbf{b} - \log\big(\mathbf{b}^T\mathbf{\Lambda}\mathbf{b}\big) + \big(\mathbf{b}^T\mu\big)^2 - 1\big), \qquad (9)$$

where $\mathbf{b}$ (the most informative filter) is a unit vector.

This function, depicted graphically in Figure 2, specifies how changes in mean and variance of the STE trade off in terms of information-theoretic significance. Figure 2A shows KL divergence as a function of both the mean ($\mathbf{b}^T\mu$) and standard deviation ($\sqrt{\mathbf{b}^T\mathbf{\Lambda}\mathbf{b}}$) of the projected stimuli. KL divergence grows as a symmetric function of the mean around 0 but is an asymmetric function of the standard deviation around 1. This asymmetry is apparent in Figure 2B, which shows a vertical slice through the function at a mean value of 0.

The objective function of Equation 8 can be computed directly from the STA and STC. That is, we can perform optimization without the computational cost of operating on the entire stimulus and spike train. Moreover, the objective function has a limited number of local maxima and connects smoothly with the results of traditional STA/STC analysis. For example, when the STC is the identity matrix, the most informative axis is the STA, and when the STA is zero, the most informative axis is either the smallest or largest eigenvector of the STC (see Appendix A for a more thorough discussion). We will refer to our approach as "information-theoretic spike-triggered average and covariance" (iSTAC) analysis.

An important advantage of the iSTAC approach over traditional STA/STC analysis is that it makes statistically efficient use of changes in both mean and covariance of the STE. Figure 3 illustrates this using simulations of a single-filter LNP model, with a stimulus consisting of temporal Gaussian white noise. The temporal filter $\mathbf{b}$ was a 20-dimensional vector resembling the (biphasic) temporal profile of a retinal receptive field. Simulations were performed using three different point nonlinearities (Figure 3, top row), each of which produces a change in both mean and variance in the STE. A half-wave rectified linear function (left column) shifts the mean of the STE and reduces its variance relative to the raw stimuli, meaning that both the STA and the low-variance STC axis provide consistent estimates for $\mathbf{b}$. For the sigmoidal and quadratic nonlinearities
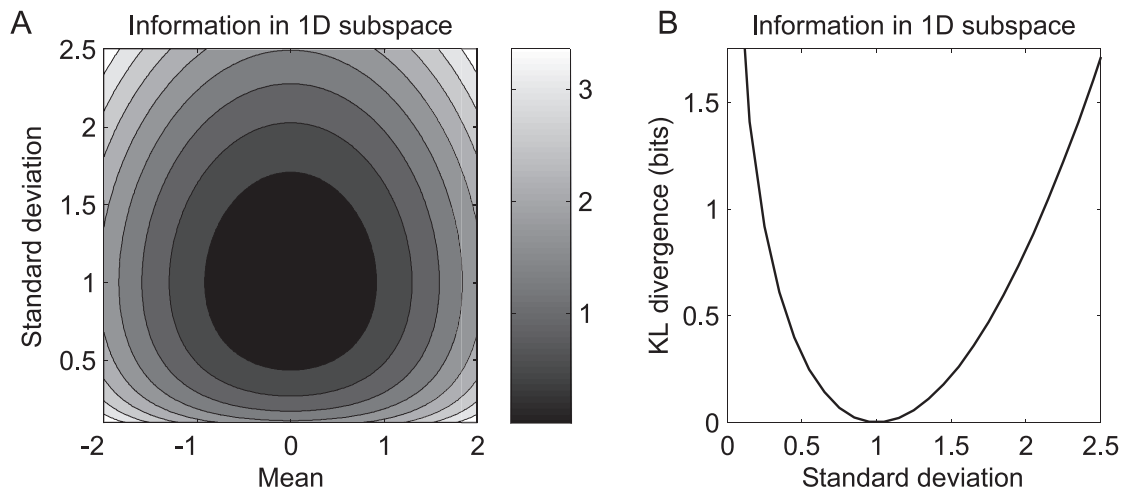
Figure 2. KL divergence between the spike-triggered and raw stimulus distributions Q and P, restricted to a 1D subspace. (A) KL divergence as a function of the mean and standard deviation of the projection of Q (Equation 9). (B) Vertical slice through the same function (at zero mean), showing asymmetry as a function of standard deviation.
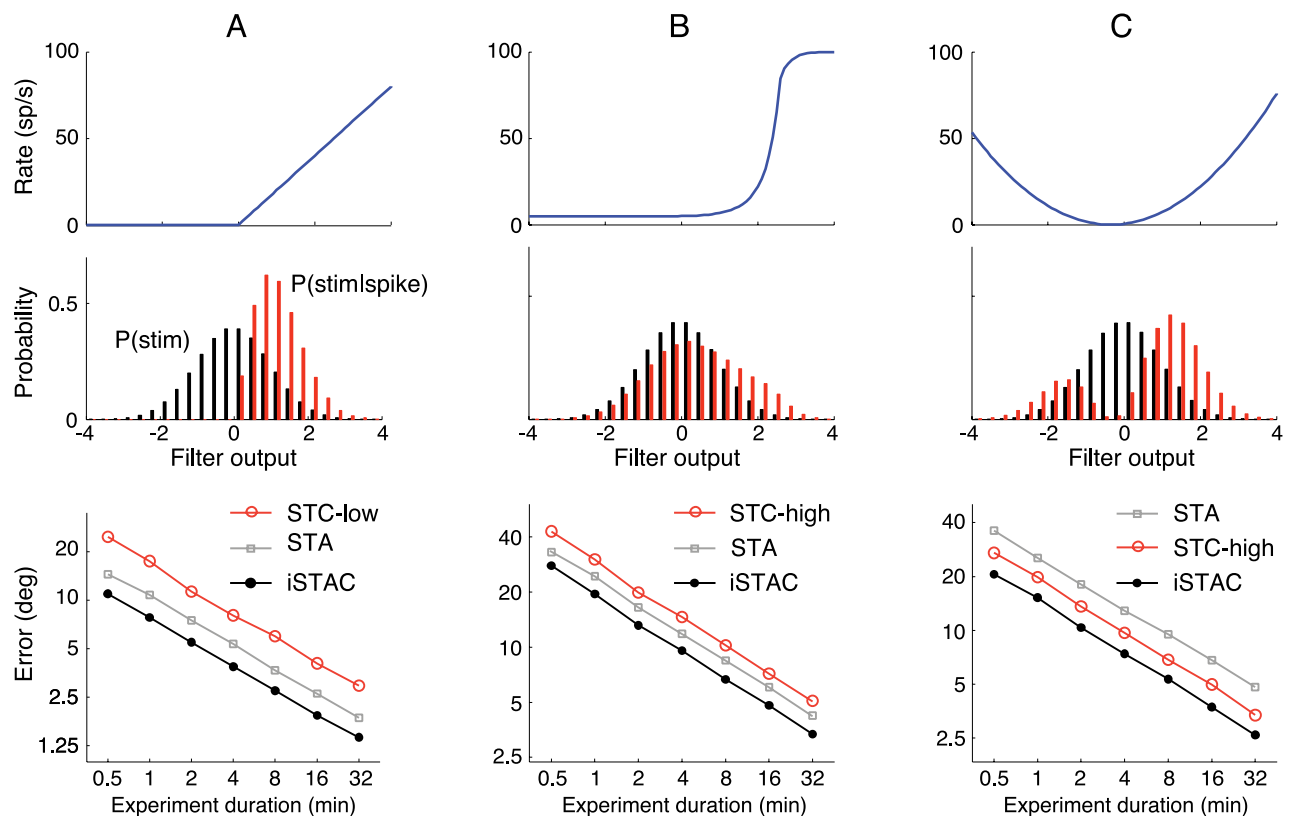

Figure 3. Comparison of STA/STC and iSTAC analysis for recovering a 1D subspace in simulation. Spike responses to a 100-Hz Gaussian white noise temporal stimulus were generated by an LNP model with a single 20-dimensional filter, followed by one of three different nonlinearities. *Top row*: Nonlinearities used for converting filter output to spike rate: (A) linear half-wave rectified, (B) sigmoidal, and (C) quadratic. *Middle row*: Distribution of raw stimuli (black) and spike-triggered stimuli (red) along the filter axis for each of these models. *Bottom row*: Error in STA, STC, and iSTAC estimates of the linear filter as a function of the number of raw stimulus samples. Error was computed as the average angle between the true and estimated filter, using 100 independent simulations at each duration.

(center and right columns, respectively), the STE has shifted mean and increased variance relative to the raw stimulus; hence, the STA and the large-variance STC axes can provide consistent estimates for **b**. The bottom row of plots in Figure 3 shows the convergence of the STA, STC, and iSTAC estimates as a function of the experiment duration. The iSTAC estimate was computed by directly maximizing Equation 9 for **b**, which gives the 1D subspace maximizing the KL divergence between P and Q. This optimization took less than 1 s for each estimate and does not depend on the amount of data, apart from the additional time required to compute the STA and STC. Although STA, STC, and iSTAC estimates all converge (i.e., are statistically consistent) for these examples, iSTAC exhibits superior performance in all three cases, due to its sensitivity to information in both mean and covariance.

Application to the HH model

Aguera y Arcas, Fairhall, and Bialek (2003) have performed an elegant dimensionality-reduction analysis of the Hodgkin-Huxley model (Hodgkin & Huxley, 1952) with STC analysis, examining how well model responses to white noise could be captured by a low-dimensional subspace. They concluded that although the HH model can be approximated with a 2D LNP model, there is no finite-dimensional space that fully captures its behavior. Here, we illustrate how the iSTAC approach can be used to supplement and extend these conclusions.

The HH model consists of a four-dimensional nonlinear differential equation that we simulated with a Gaussian white noise input current, discretized in 1-ms bins. We used a 100-ms portion of the stimulus preceding each spike to define the spike-triggered ensemble. Figure 4A shows the sorted eigenvalues of the STC matrix, computed from a simulated train with 10^8 time samples and roughly 10^6 spikes. A substantial number of eigenvalues lie above or below 1, indicating that the computation performed by the HH neuron on its input is multidimensional (Aguera y Arcas et al., 2003).

Nevertheless, if we desired a 1D approximation to the HH model, we could ask: Which dimension of the stimulus space preserves the most information about HH neuron's response? Equivalently, what filter provides the best description of the HH model's temporal receptive field? Figure 4B shows several candidate filters offered by traditional STA/STC analysis, along with the solution offered by the iSTAC approach. The iSTAC filter is intermediate between the STA and the lowest variance eigenvector—it lies in the space spanned by these two vectors, but it preserves 10% more information about spike times than either of these vectors individually and roughly eight times more information than the large-variance eigenvector.

The information-theoretic framework can also be used to examine additional dimensions. Figure 4A shows the eigenvalues of the STC, revealing a large number that deviate significantly from 1. We can use the iSTAC approach to find a set of vectors ordered according to informativeness. For each dimensionality m, we computed the optimal information-preserving subspace by maximizing Equation 8 for **B** (a $100 \times m$ matrix), constraining the columns of **B** to be orthonormal (see Appendix A for details of the optimization procedure). Figure 4C shows the KL divergence between
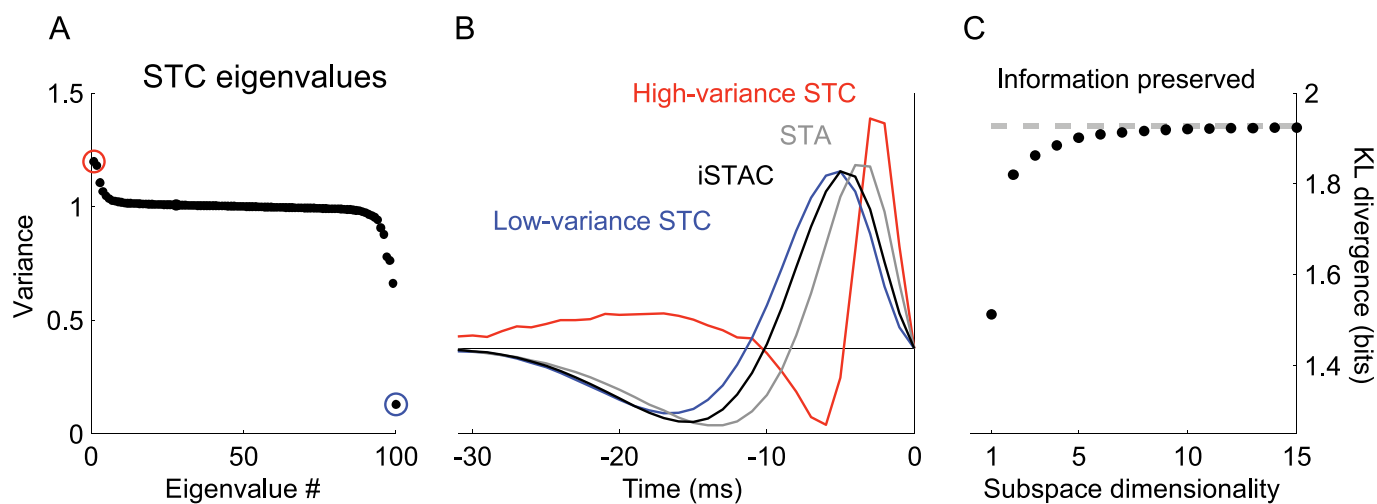

Figure 4. Comparison of STA/STC and iSTAC analyses on simulated data from an HH model. (A) Eigenvalues of the STC matrix, showing many eigenvalues larger and smaller than 1; circles highlight the largest (red) and smallest (blue) values. (B) Candidate vectors for dimensionality reduction of the HH model to a 1D subspace: the STA (gray), eigenvectors associated with the largest (red) and smallest (blue) eigenvalues, and the first iSTAC axis (black). The information preserved by these vectors is 1.52 bits (iSTAC), 1.38 bits (STA and low-variance STC), and 0.18 bits (high-variance STC). (C) Information preserved by the optimal (iSTAC) subspace as a function of dimensionality. Dashed line indicates the information available in the full 100-dimensional stimulus space.

the two distributions projected into this optimal subspace, as a function of m. Although the large number of "significant" eigenvalues (Figure 4A) reveals that the HH computation is high dimensional, the information analysis indicates that total information saturates quite rapidly with dimensionality. Specifically, as shown in Figure 4C, a model using just two linear filters captures 94% of the information available.

Application to V1

We have also applied iSTAC to data from a V1 complex cell (data published in (Rust et al., 2005)). The stimulus consisted of a set of adjacent black and white flickering bars (i.e., binary white noise), aligned with the cell's preferred orientation. Figures 5A–D show the results of STA/ STC analysis, with STA and STC eigenvectors displayed
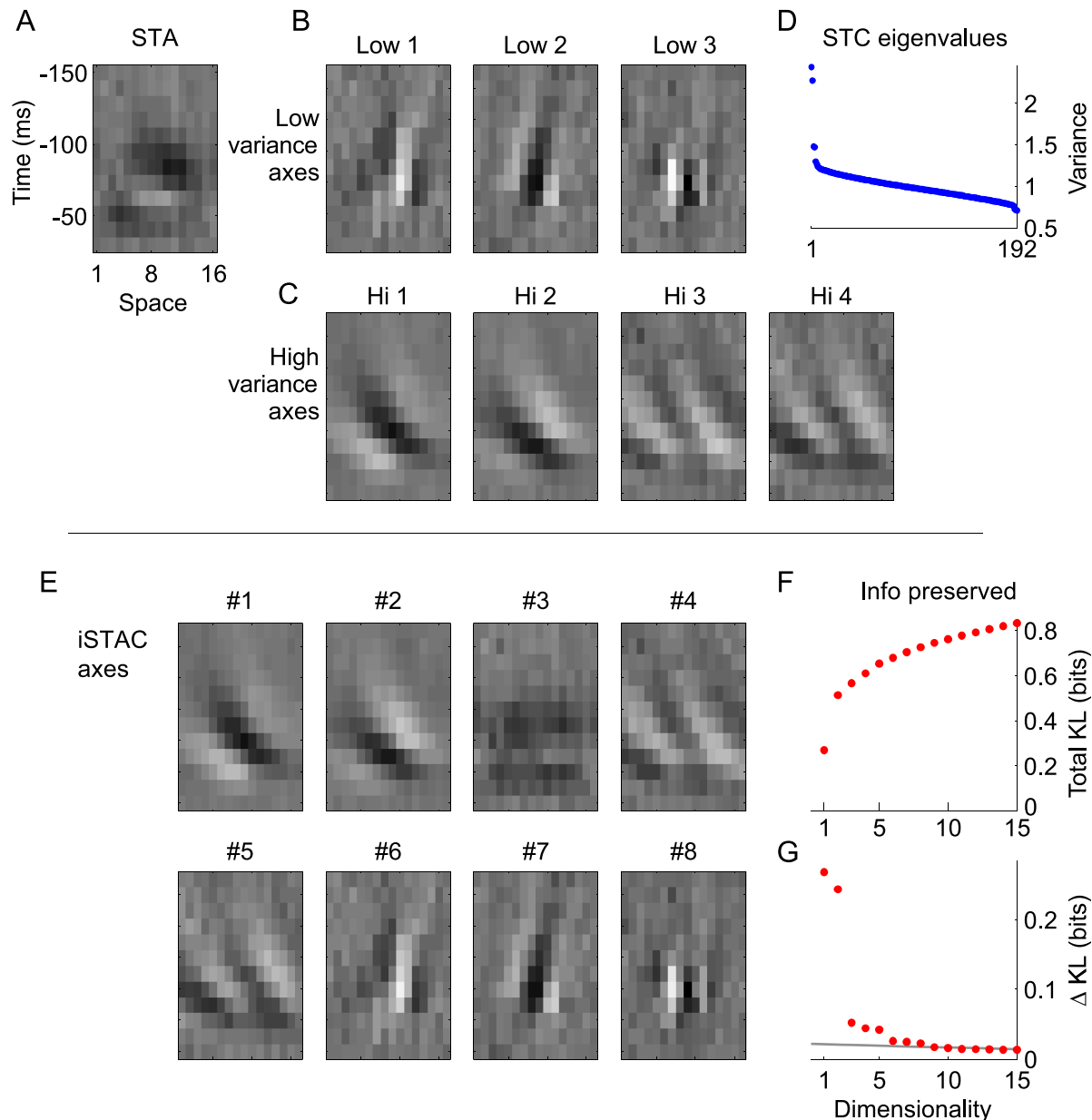

Figure 5. Analysis of a V1 complex cell. The stimulus consisted of 16 black/white bars, aligned with the cell's preferred orientation, each flickering randomly at a rate of 100 Hz. The stimulus block was chosen to cover 12 time bins. (A) The STA plotted as a 12×16 image, showing temporal (vertical) and spatial (horizontal) organization. (B and C) High- and low-variance STC eigenvectors. (D) Sorted eigenvalues of the STC matrix. Four large and three small eigenvalues were determined to be statistically different from 1 (the variance of the raw stimuli). (E) iSTAC dimensionality reduction. The iSTAC filters span nearly the same space as the STA and eigenvectors of the STC but are orthogonal and sorted by informativeness. (F) Information preserved as a function of the dimensionality of the most informative subspace. (G) Incremental information gain as a function of subspace dimensionality (difference between adjacent points in F). The gray line shows a 95% confidence interval for the increase in KL divergence due to undersampling (i.e., noise in the STA and STC), computed using nested bootstrap resampling.

as grayscale space–time images. This cell exhibits a structured STA, high-variance eigenvectors tuned for left-moving stimuli, and low-variance eigenvectors tuned for rightward-moving stimuli.

This collection of significant STA and STC eigenvectors suggests that an eight-dimensional subspace captures the cell's response to flickering bar stimuli. However, the analysis does not by itself tell us the relative significance of the axes nor does it tell us the information-theoretic cost of using a lower dimensional subspace. Figure 5E shows a collection of (orthogonal) basis vectors, sorted in order of their informativeness. The optimal k-dimensional subspace is given by the collection: $\{1, 2, …, k\}$. The first two iSTAC vectors closely resemble the first two high-variance eigenvectors of the STC. The third iSTAC vector resembles the STA (orthogonalized with respect to the first two), and the remaining vectors closely match the remaining eigenvectors of the STC (with high-variance vectors preceding the low-variance vectors in importance). Note, however, that this ordering was not obvious a priori. Other V1 cells from the same data set reveal a variety of orderings: In some cells, the STA carries more information or less information than all other filters, and in some, the low-variance axes carry more information than all high-variance axes.

Note also that the ordering of filters is not necessarily the same as if we sorted them by the amount of information preserved in a 1D projection. The information-theoretic criterion of Equation 8 takes into account correlations between the projection of spike-triggered stimuli onto the $(k + 1)$th dimension and the previous k dimensions. Such correlations are important when the STA is not geometrically aligned with the eigenvectors of the STC matrix. For example, we often find that the second iSTAC filter carries less information by itself, as compared with the third or fourth filters, but gives rise to the most informative 2D subspace when grouped with the first.

Figure 5F shows the amount of information preserved by the optimal k-dimensional subspace. Moving from a 1D to a 2D representation increases information nearly as much as moving from a zero-dimensional to a 1D representation (increases of 0.27 and 0.24 bits, respectively), but moving to larger dimensional subspaces does not contribute nearly as much additional information, as illustrated in Figure 5G. Note that although information continues to increase as a function of subspace dimensionality, this increase can be attributed to data limitations. The covariance matrix Λ is an *estimate* of the true covariance, computed from a finite set of samples. These samples have, by chance, slightly smaller or greater variance than the raw stimuli along most dimensions, resulting in an apparent increase in information with each dimension included in the model. This same phenomenon is responsible for the spread of "nonsignificant" eigenvalues around 1 in Figure 5B.

Thus, we use a statistical test (specifically, a nested hypothesis test) to determine when the information increase is significant compared with that due to undersampling.

Details of this procedure are given in Appendix A. The gray line in Figure 5G shows the result of this nested test performed for each dimensionality. Although total information continues to increase with dimensionality (Figures 5F and G), for $m > 8$, the amount added does not exceed the 95% confidence level for the information increase we would expect due to statistical error in estimation of mean and covariance with this number of samples, and thus we conclude that the cell's response is captured by an eight-dimensional model.

Application to retina

We also applied the iSTAC method to spiking data from macaque RGCs (Chander & Chichilnisky, 2001; Chichilnisky, 2001). The stimulus again consisted of flickering bars, with intensities drawn i.i.d. from a Gaussian distribution (i.e., Gaussian white noise). Figure 6 shows a comparison of STA/STC and iSTAC dimensionality reduction for a sample cell. Figure 6A shows the STA of the neuron, which exhibits canonical ON-type RGC receptive field: center-surround spatial organization and a biphasic temporal profile. STC analysis (Figures 6B–D) indicates that the neuron's response is inherently multidimensional, with two significant high-variance eigenvectors and three significant low-variance eigenvectors.

Figure 6E shows the significant iSTAC features (sorted by informativeness). Although the basis provided by STA/STC analysis contains six dimensions, the information-theoretic analysis finds only five that are significant. The first of these closely resembles both the STA and the lowest-variance STC features, but the axis corresponding to the difference between the two does not contribute meaningful information. This is not the case for all RGC cells we examined: Several exhibited the same dimensionality under STA/STC and iSTAC analysis. Note also that iSTAC Axes 3 and 5 resemble the sum and difference of low-variance STC Axes 2 and 3, respectively. The information preserved in a subspace, as a function of subspace dimensionality, is shown in Figures 6F and G.

Modeling the nonlinear response

Dimensionality reduction provides a linear mapping of the stimulus x to a feature vector $x^* = \mathbf{B}^T x$, where $\mathbf{B}$ is a basis for the feature space. For a complete model, we also need a mapping from the feature vector x^* to the probability of observing a spike, $P(\text{spike}|x^*)$. In low-dimensional (1D or 2D) spaces, the nonlinearity can be estimated by computing the quotient of (e.g., histograms of) the densities $P(x^*|\text{spike})$ and $P(x^*)$ (see Figures 1C and D). For higher dimensional feature spaces, however, this is infeasible due to the difficulty of estimating densities.
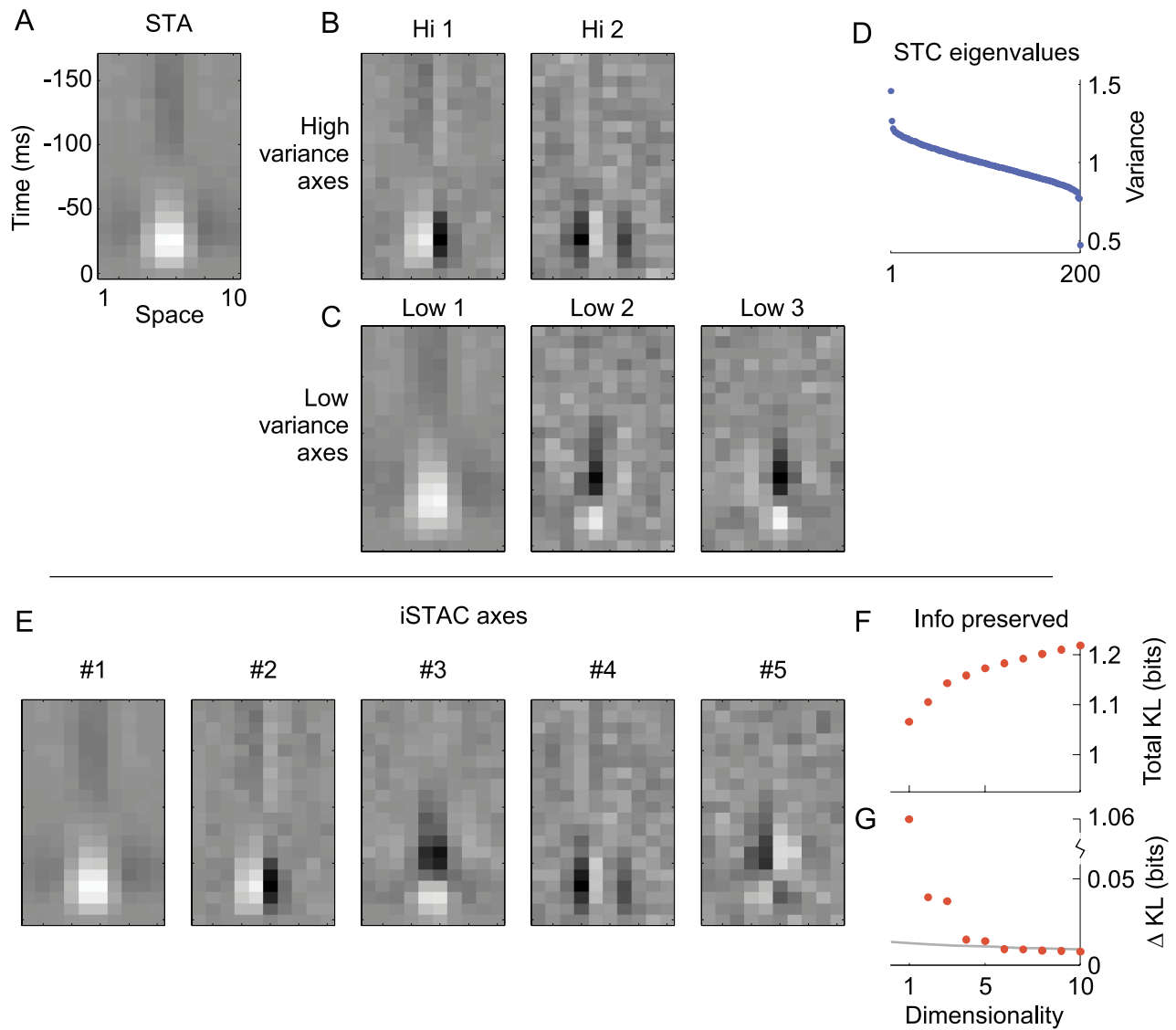
Figure 6. STA/STC analysis and iSTAC analysis of an ON RGC in macaque retina. The stimulus consisted of 10 spatially adjacent bars, with intensities on each frame drawn from a Gaussian white noise source, presented at a frame rate of 120 Hz. The stimulus vector for our analysis was chosen to include 20 time bins, producing a stimulus space of dimensionality 200. (A) STA. (B) Eigenvalues of STC. To the right, STC eigenvectors associated with high (C) and low (D) variance. (E) Basis vectors of the iSTAC dimensionality reduction, sorted in order of their informativeness. (F) Total information preserved and (G) incremental information added as a function of subspace dimensionality.

In such situations, the information-theoretic iSTAC framework provides a default model of the nonlinearity in the form of a "ratio of Gaussians" (ROG):

$$P(\text{spike}|x) = \alpha \frac{Q(x)}{P(x)}, \tag{10}$$

where $Q(x)$ is a Gaussian density with mean and covariance matching that of the spike-triggered ensemble, $P(x)$ is the prior distribution over raw stimuli, and α is a proportionality constant equal to $P(\text{spike}) = n_{\text{sp}}/n_{\text{stim}}$. Reducing dimensionality by a linear projection onto **B** preserves Gaussianity of both numerator and denominator distribu-

tions; thus, the reduced-dimensional model of the neural response, specified in feature space, is given by

$$P(\text{spike}|x^*) = \alpha \frac{\hat{Q}(x^*)}{\hat{P}(x^*)}, \tag{11}$$

where $\hat{Q}$ is Gaussian with mean $\hat{\mu} = \mathbf{B}^T\mu$ and covariance $\hat{\Lambda} = \mathbf{B}^T\Lambda\mathbf{B}$. A bit of algebra reduces this to an exponential form:

$$P(\text{spike}|x^*) = ae^{x^{*T}Mx^* + b^T x^*}, \tag{12}$$

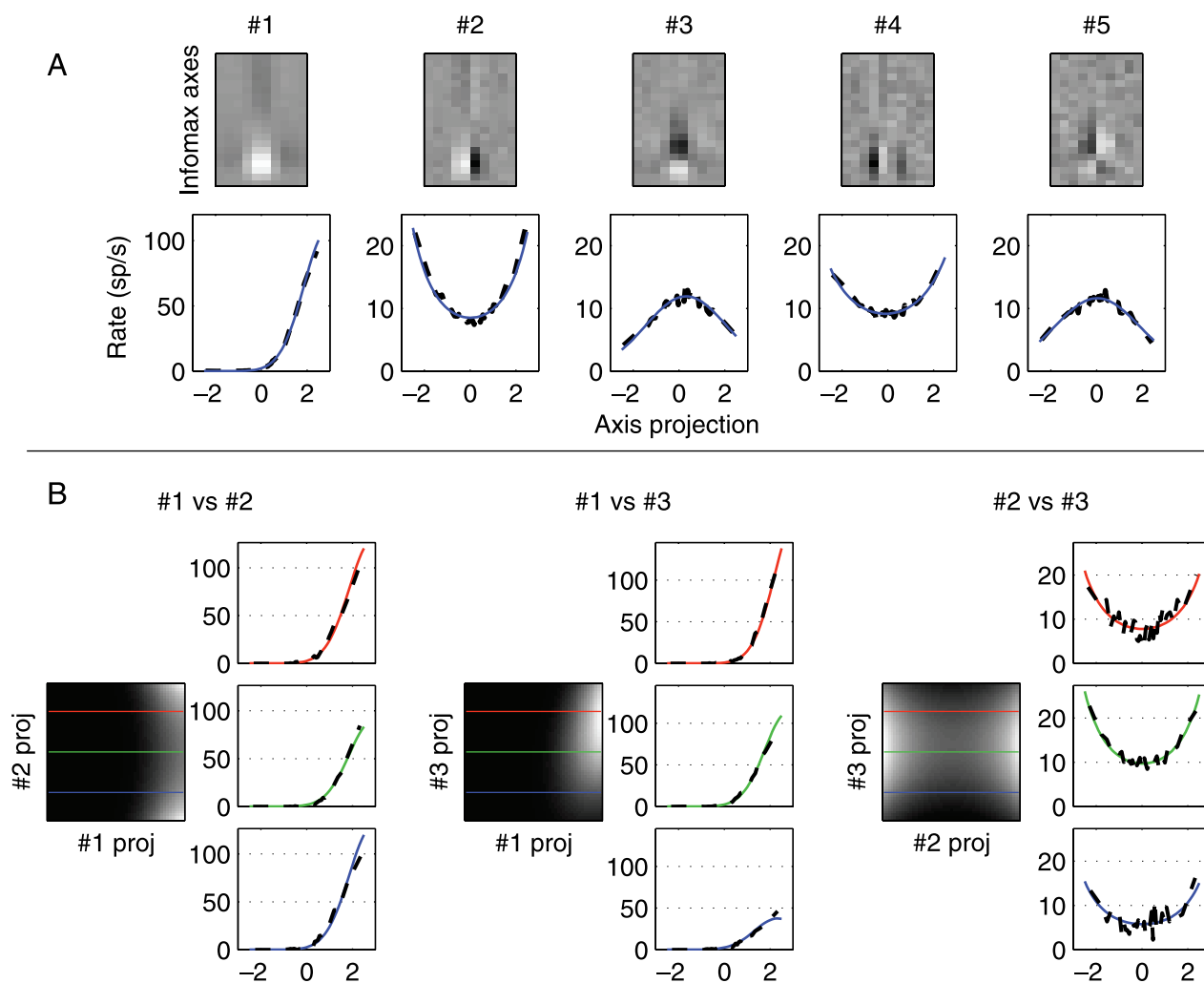where $M = \frac{1}{2}(I - \hat{\Lambda}^{-1})$ and $b = \hat{\Lambda}^{-1}\hat{\mu}$.

Figure 7. Reconstruction of 1D (marginal) and 2D nonlinearities using the ROG model for the RGC shown in Figure 6. (A) Five vectors discovered using iSTAC (above) and the 1D nonlinearity obtained by projecting the stimulus onto each of these axes (below). Blue traces show the prediction of the ROG model, and dashed lines show the histogram-based estimate. (B) 2D nonlinearities obtained by projecting onto pairs of feature vectors. The three grayscale images show the probability of spiking under the ROG model as a function of the projection onto Feature Vectors 1 and 2 (left), 1 and 3 (middle), and 2 and 3 (right). Red, green, and blue lines indicate conditional "slices" through at 1.5, 0, and −1.5, respectively. These slices are plotted to the right of each image: Solid lines show ROG prediction, and dashed lines show histogram estimates computed using stimuli whose projection onto the ordinate feature vector was within ±0.1 of the slice value.

Although an ROG might initially seem like a strange choice of parametric model, we find that it provides a surprisingly good fit to the data of many cells. Figure 7A shows a comparison between the ROG model and a histogram estimate of the nonlinearity along each dimension of the (five-dimensional) feature space. For each plot, the stimuli were projected onto feature vector $\mathbf{b}_j$, and the nonlinearity was computed using the ROG (blue) and a ratio of densities estimated using histograms (dotted black; Chichilnisky, 2001). Note that, for each plot, the numerator of the ROG model is a Gaussian with mean ($\mathbf{b}_j^T\mu$) and variance ($\mathbf{b}_j^T\Lambda\mathbf{b}_j$), whereas the histogram estimate is computed using the 1D projection of the stimuli: $\{\mathbf{b}_j^T x\}$. The model is seen to be

equally adept at describing the asymmetric, symmetric excitatory, and symmetric suppressive behaviors found along different axes.

Figure 7B shows an analysis of the nonlinearity as projected onto pairs of axes of the feature space. The grayscale images represent the probability of spiking under the ROG model. Conditional slices through these plots show comparisons of the model to histograms of the data (dashed lines). The example in the center column (Feature 1 vs. Feature 3) shows the most striking change in the nonlinearity because of conditioning. For large positive projections onto Feature 3 (red line), the nonlinearity is steep and has a high threshold, whereas for negative projections (blue line), the

nonlinearity is shallow with a lower threshold. Note that these dependencies in spike probability could not have been predicted from the marginal spike probabilities shown in Figure 7A for Features 1 and 3.

Agreement across multiple 1D and 2D projections and slices suggests that the ROG model provides a reasonably good approximation of RGC responses. We have examined similar projections onto linear combinations of these axes (i.e., rotations of the feature space) and found similarly good agreement. In other cell types (such as the V1 cells described previously), the ROG model may prove less accurate, and we may need to introduce a different model of the nonlinear mapping from feature space to spike rate. Even in such cases, the ROG model provides a first approximation of the nonlinearity, with the advantage that it is completely determined by the filters used for dimensionality reduction (i.e., it requires no additional parameter fitting). This is particularly useful when the subspace dimensionality is greater than 2 and nonparametric nonlinearity reconstruction is not feasible.

As we will show in Appendix B, dimensionality reduction using iSTAC analysis is asymptotically optimal if the response nonlinearity is an ROG. Specifically, for responses generated by an ROG model, iSTAC analysis performs a maximum likelihood (ML) estimate of the model parameters. This generalizes the optimality conditions for STA and STC analysis: The STA is an ML estimate when the response nonlinearity is exponential (i.e., P and Q have identical covariance but differ in mean; Paninski, 2004) and, as shown here, STC analysis corresponds to ML estimation when the nonlinearity is the ratio of two zero-mean Gaussians (P and Q have the same mean but differing covariance). Thus, iSTAC analysis generalizes the optimality of STA and STC. It is also consistent and unbiased under the same conditions as STA and STC analysis, meaning that it converges to the correct subspace whenever the raw stimulus is Gaussian and the nonlinearity affects STE's mean, variance, or both. The ROG description also provides an important litmus test for the possible suboptimality of these moment-based approaches. If the estimated nonlinearity is poorly fit by an ROG, there may be a significant statistical advantage to dimensionality-reduction techniques that are sensitive to higher order moments (e.g., Paninski, 2003; Sharpee et al., 2004).

Extension: Analysis of space–time separable models

Finally, the iSTAC framework can be extended to incorporate additional constraints on the filters recovered, which we illustrate with an application to a model with space–time separable elements. Figures 8D and E shows an example of this approach: The spatial and temporal sections of the iSTAC filters estimated for the RGC cell (Figure 6, blue traces) exhibit only a small number of distinct profiles, suggesting that we can reduce model complex-

ity by using only a small number of spatial and temporal waveforms.

A space–time separable filter is one that can be specified as the outer product of a temporal filter $\mathbf{h}$ and a spatial filter $\mathbf{g}$:

$$\mathbf{hg}^T = \begin{bmatrix} h_1 g_1 & h_1 g_2 & \cdots & h_1 g_{n_g} \\ h_2 g_1 & h_2 g_2 & \cdots & h_2 g_{n_g} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n_h} g_1 & h_{n_h} g_2 & \cdots & h_{n_h} g_{n_g} \end{bmatrix}, \quad (13)$$

where n_h and n_g are the number of spatial and temporal elements of the raw stimulus, respectively. Note that this greatly reduces the number of filter parameters from that of stimulus dimensionality n (equal to $n_h \times n_g$) to $n_h + n_g$. By stacking the columns of $\mathbf{hg}^T$ to form a single column vector,

$$\mathbf{b} = \begin{bmatrix} \mathbf{h}g_1 \\ \mathbf{h}g_2 \\ \vdots \\ \mathbf{h}g_{n_g} \end{bmatrix} = \begin{bmatrix} \mathbf{h} & & & \\ & \mathbf{h} & & \\ & & \ddots & \\ & & & \mathbf{h} \end{bmatrix} \mathbf{g} = [\mathbf{L}_h]\mathbf{g}, \quad (14)$$

where $\mathbf{L}_h$ is an $n \times n_g$ block-diagonal matrix, with each block given by the column vector $\mathbf{h}$.

Suppose now that we wanted to find the temporal filter $\mathbf{h}$ that preserves maximal information about the response. Filtering each spatial element of the stimulus with $\mathbf{h}$ is equivalent to projecting each stimulus x onto the columns of $\mathbf{L}_h$; this operation produces an n_g-dimensional vector, $\mathbf{L}_h^T x$, with one dimension for each spatial element of the stimulus. From this derivation, it is obvious that $\mathbf{L}_h$ is a special form of the dimensionality-reducing matrix $\mathbf{B}$ that we considered previously. Therefore, we can find the most informative $\mathbf{h}$ simply by maximizing KL divergence, using $\mathbf{L}_h$ in place of $\mathbf{B}$ in Equation 8. Note that we could not directly estimate such a temporal filter using STA or STC analysis: Although we could find a space–time separable fit to either the STA or the eigenvectors of the STC, this provides no unique solution nor does it combine information from all of the filters (STA and the significant eigenvectors) simultaneously. The more general information-theoretic estimators also cannot be tractably applied to this problem. The matrix $\mathbf{L}_h$ reduces dimensionality to an n_g-dimensional feature space, which is too high-dimensional (assuming the stimulus contains more than a few spatial elements) for estimating mutual information directly.

If one temporal filter does not suffice to describe the response, we can find multiple filters using the same approach: Each temporal filter $\mathbf{h}_i$ is inserted into a dimensionality-reducing matrix $\mathbf{L}_{hi}$, and the concatenation of these matrices, $[\mathbf{L}_{h1} \ldots \mathbf{L}_{hk}]$, preserves more dimensions of the original stimulus space and can be inserted into Equation 8 in place of $\mathbf{B}$. Figure 8 shows an application of this approach to the RGC shown in Figures 6 and 7. Figure 8A shows the two most informative temporal filters, which were found to be
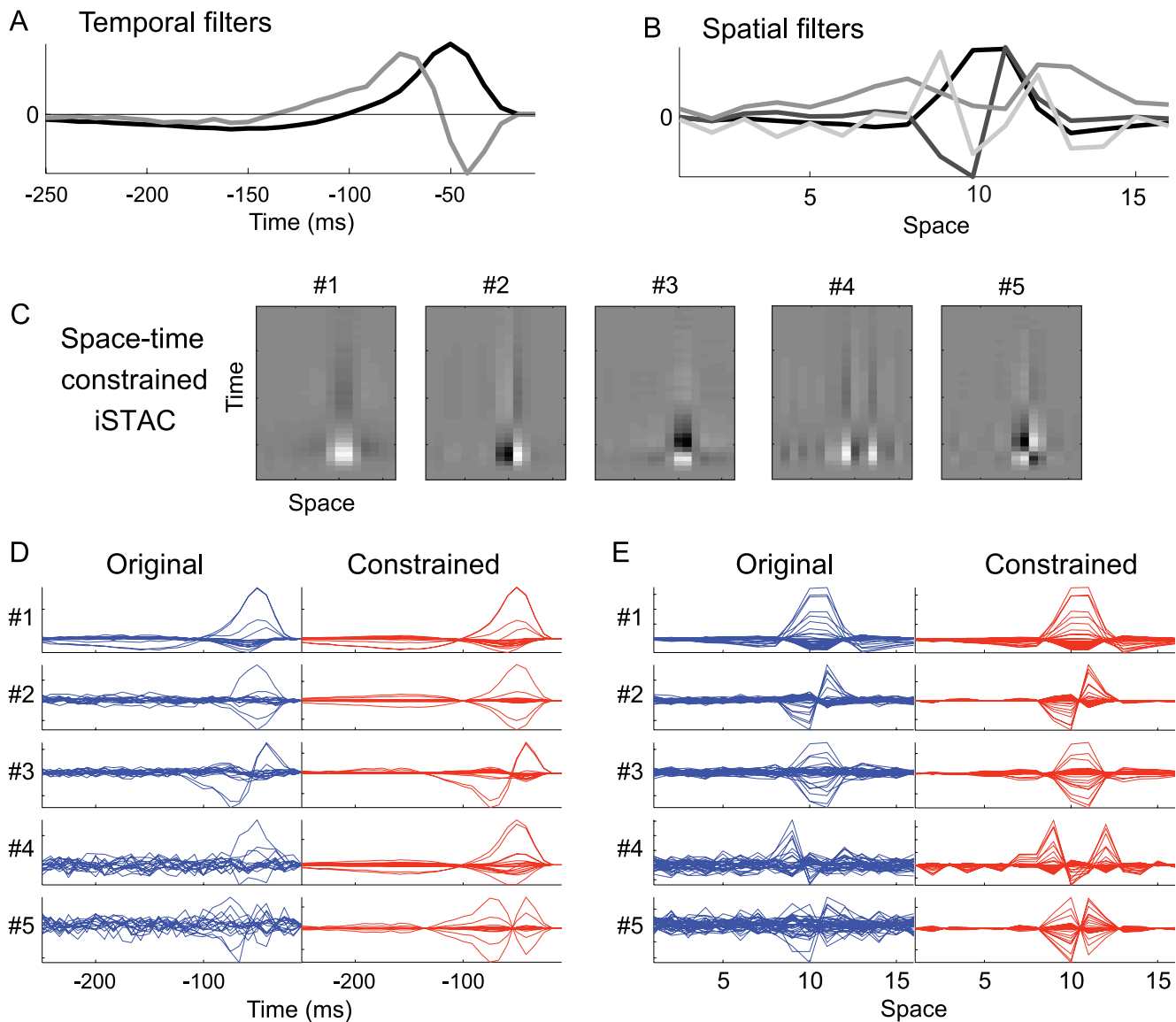
Figure 8. Dimensionality reduction with space/time-constrained iSTAC. (A) Temporal dimensionality reduction finds that the temporal dependence of the response can be well characterized using two filters. The black line represents the first (more informative) filter. (B) Spatial dimensionality reduction indicates that four filters can characterize spatial dependence. The informativeness of each filter is indicated by grayscale level (lighter traces = less informative). (C) iSTAC filters constrained to live in the space spanned by the spatial and temporal filters in A and B. These compare directly with the iSTAC filters in Figure 6E. (D) Temporal sections of the original iSTAC filters (Figure 6) and the space/time-constrained filters shown in C. (E) Spatial sections of the original and the space/time-constrained iSTAC filters. Spatial and temporal similarity across sections and across filters suggests that the small number of spatial and temporal filters discovered provides a parsimonious description of the neural feature space.

sufficient for preserving the information about the response. We then performed an identical analysis to find a set of maximally informative spatial filters (i.e., by exchanging the roles of **h** and **g** in the previous analysis) and found that four filters were required to preserve spatial information about the response, as shown in Figure 8B.

Finally, we can combine these temporal and spatial filters to obtain a set of *constrained* iSTAC filters; these should resemble the original filters but obey the additional constraint that they are composed only of the spatial and temporal filters obtained from the space–time separable analysis. This constraint implies that each filter **b** can written as

$$\mathbf{b} = \sum_{i=1}^{n_h} \sum_{j=1}^{n_g} \left(\mathbf{h}_i \mathbf{g}_j^T \right) w_{ij}, \tag{15}$$

where $\mathbf{h}_i \mathbf{g}_j^T$ indicates the outer product of the i temporal filter with the jth spatial filter, and $\{w_{ij}\}$ is a set of linear weights,

which we fit by maximizing KL divergence. Figure 8C shows the set of constrained iSTAC filters obtained for this cell, each of which is constructed using the spatial and temporal components shown in Figures 8A and B and a set of weights. For comparison, we can plot these alongside the original filters obtained with iSTAC analysis (Figure 6). Figures 8D and E show temporal and spatial sections of the original iSTAC filters (blue) and the space/time-constrained filters (red), indicating basic agreement between the two methods. The constrained filters, however, are much smoother and require far fewer parameters to describe: A set of five 300-element filters (20 temporal × 15 spatial dimensions) was reduced to a set of filters constructed from two temporal filters and four spatial filters and a set of weights (eight for each of the five filters), resulting to a reduction from 1,500 to 140 parameters.

Discussion

We have described a methodology for fitting an LNP model to extracellular data, in which we maximize the information between stimulus and response, assuming an ROG response model. The model parameters are fully constrained by the mean and covariance of the raw and spike-triggered stimulus ensembles, and thus it is seen to occupy a middle ground between STA/STC and information maximization methods for dimensionality reduction while providing some advantages over each. In addition, our method provides a default model, in the form of an ROG, for the nonlinearity that maps linear responses to firing rates.

As with STA/STC analysis, the restriction to a model that is characterized by first and second moments is what guarantees tractability of the fitting procedure. However, this also means that the method is blind to variations that manifest themselves only in higher order moments. It should be possible to augment the method to include higher order moments, but these will necessarily increase the data required for accurate estimation and are also likely to increase the chances of getting stuck in local minima during fitting.

Although we found that the ROG model provides a surprisingly good account of data from RGCs, it is unlikely to provide an accurate description for all cells. For example, we find that the fits are not nearly as good for V1 responses. However, we note that the ROG model is not essential to our analysis, and it is easy to envision generalizations in this regard. For example, the ROG model can be raised to an unknown power, allowing the model to fit nonlinearities that accelerate more steeply than those shown in Figure 7. More generally, one could introduce any parametric nonlinearity to operate on the feature space, as long as the parameter-fitting problem is tractable.

Appendix A: Information maximization

iSTAC analysis provides a natural generalization of STA and STC analysis, which we will show by proving that iSTAC reduces either to the STA or to the STC analysis in the case that either the STC or the STA provides no information about the response. Specifically:

- When the STC is the identity matrix, iSTAC recovers the same subspace as STA. The proof is simple: If STC Λ is the identity matrix, then the first two terms in Equation 8 are constant, and we obtain maximal KL divergence by taking $\mathbf{B}$ as a unit vector proportional to μ, the STA. Note that in this case, all information is captured by this 1D subspace; hence, there is no advantage to using a higher dimensional feature space.

- When the STA is 0, iSTAC recovers the same subspace as the significant eigenvectors of the STC. To prove this, let $\mathbf{A} = \mathbf{B}^T\Lambda\mathbf{B}$, and the KL divergence reduces to $\mathrm{Tr}[\mathbf{A}] - \log|A|$ plus a constant. Note that the first term of this expression is the sum of the eigenvalues of $\mathbf{A}$ and that the second is the negative sum of the log eigenvalues of $\mathbf{A}$. These eigenvalues, in turn, represent the variance of the STE along each major axis preserved by the basis $\mathbf{B}$. If we diagonalize Λ using its eigenvectors, it is easy to show that the function is maximized by setting $\mathbf{B}$ to contain the eigenvectors of Λ for which the corresponding eigenvalues σ_i are greater than or less than 1. The information contributed by each eigenvector is equal to $\sigma_i - \log(\sigma_i) - 1$, which is the function plotted in Figure 2B, and monotonically increases as σ_i moves away from a value of 1. This means that extrema (high or low eigenvalues) of the STC will also be maxima in the iSTAC analysis, and thus, the iSTAC basis will be the same as the STC basis. Moreover, if we wish to preserve only j axes of the stimulus space, the most informative j-dimensional subspace is generated by the eigenvectors whose corresponding eigenvalues σ_i give the j largest values of $\sigma_i - \log(\sigma_i)$.

When both the STA and STC contain meaningful information about the neural response, as in all real and simulated examples presented here, we desire a basis that maximizes the full objective function of Equation 8. The objective function can be rewritten simply as:

$$f(\mathbf{B}) = \mathrm{Tr}[\mathbf{B}^T(\Lambda + \mu\mu^T)\mathbf{B}] - \log|\mathbf{B}^T\Lambda\mathbf{B}|. \tag{A1}$$

Note that the first term is the sum of the eigenvalues of $(\Lambda + \mu\mu^T)$ and that the second term is the sum of the log eigenvalues of Λ within the subspace preserved by $\mathbf{B}$. A simple intuition (verified by hand inspection in low dimensions) suggests that this objective function has only n

local maxima and/or saddle points on the unit hypersphere, which are intermediate between the eigenvectors of $(\Lambda + \mu\mu^T)$ and Λ.

For numerical optimization, we also made use of the analytic gradient of the objective function, which is given by

$$\frac{\partial f}{\partial \mathbf{B}} = 2\left(\Lambda + \mu\mu^T - \Lambda \mathbf{B}\mathbf{B}^T \Lambda^{-1}\right)\mathbf{B}. \tag{A2}$$

We performed the optimization by growing $\mathbf{B}$ incrementally, starting with the maximally informative 1D basis and adding columns so that KL divergence is maximized for each dimensionality. The optimal k-dimensional basis includes the $(k-1)$-dimensional basis provided by the previous step of the algorithm; thus, it is possible to think of the optimal basis as consisting of the first k vectors from a fixed, ordered set (as shown in Figures 5, 6, and 8). To ensure that the optimization converges to the true global optimum at each step, we use several initialization points, selected from a set of the significant eigenvectors of Λ and $(\Lambda + \mu\mu^T)$.

To determine the number of significant subspace dimensions, we performed a nested bootstrap test, analogous to that described for STC analysis in Rust et al. (2005), Schwartz et al. (2002), and Simoncelli et al. (2004). The test at step k examines whether the incremental information that arises from increasing dimensionality from $k - 1$ to k is significantly above that expected from random sampling. To quantify the latter, we performed 1,000 bootstrap resamplings of the STE by randomly time shifting the spike train relative to the stimulus (removing stimulus dependence of the response but preserving spike train statistics) and computed the STA and STC of the shifted samples. We then computed the KL divergence of the most informative k-dimensional subspace while setting the mean and covariance in the first $k - 1$ dimensions to be those given by the true STA and STC. We use these 1,000 estimates to generate an empirical distribution of the incremental information provided by the kth dimension and compute a 95% confidence level (gray line plotted in Figures 5 and 6). If the incremental information computed from the actual data fails to surpass this significance level, we conclude that the neural response is captured by the first $k - 1$ dimensions. Otherwise, we proceed by repeating the whole test for $k + 1$ dimensions.

Appendix B: Relationship to ML

It is interesting to note that maximizing KL divergence between Q and P is asymptotically equivalent to finding the ROG model parameters that maximize the likelihood of the spike train given the stimuli. We assume that spikes are generated according to an inhomogeneous Poisson process,

and thus, the likelihood of observing k spikes for a stimulus x is given by

$$p(k|x) = \frac{1}{k!} r(x)^k e^{-r(x)}, \tag{B1}$$

where $r(x)$ is the instantaneous firing rate. The average log-likelihood of set of spike data $\{k_i, x_i\}$, for $i \in [1, N]$ is given by

$$L(\{k_i, x_i\}) = \frac{1}{N} \sum_i [k_i \log r(x_i) - r(x_i)] + c, \tag{B2}$$

where c is a constant that does not depend on $r(x)$.

We assume that the spike rate is determined by the ROG:

$$r(x) = \alpha \frac{Q(x)}{P(x)}. \tag{B3}$$

We substitute this into Equation B2 and take the limit as the amount of data goes to infinity:

$$L_N \rightarrow \int Q(x) \log \frac{Q(x)}{P(x)} dx - \int P(x) \frac{Q(x)}{P(x)} dx \tag{B4}$$

$$= D(Q, P) - 1. \tag{B5}$$

Therefore, any parameter of Q and P that maximizes KL divergence will also (in the limit of large data) maximize the Poisson likelihood of the data under the model.

One corollary of this result is that STC analysis is asymptotically optimal (i.e., equivalent to ML) when the response function $r(x)$ is a ratio of two zero-mean Gaussians. This follows from the conjunction of the second point in the previous section (equivalence of STC and iSTAC when the expected mean of the STE is zero) and the optimality of iSTAC under an ROG model. The corollary that the STA is asymptotically optimal when $r(x)$ is exponential has been shown previously (Paninski, 2004) but can be derived similarly from the fact that the ratio of two Gaussians with identical covariance but shifted means is exponential.

Although it is not optimal for other nonlinearities, iSTAC analysis is both unbiased and consistent whenever the raw stimulus distribution is Gaussian, and the nonlinearity affects the STE's mean, variance, or both. This follows directly from the unbiasedness and consistency of the STA and STC eigenvectors under the same conditions, which have been shown previously (Bialek & de Ruyter van Steveninck, 2005; Bussgang, 1952; Paninski, 2003). If we have an LNP neuron with a set of linearly independent filters and a nonlinearity that affects the mean, variance, or both along each axis of the subspace spanned by these filters, then the expected STA and STC eigenvectors span the same space

(unbiasedness) and converge asymptotically to this subspace (consistency). Unbiasedness and consistency of iSTAC analysis result from the fact that the expected and asymptotic KL divergence along axes outside this subspace is zero, meaning that the expected and asymptotic maximizer of Equation 8 is indeed the correct subspace.

References

Aguera y Arcas, B., & Fairhall, A. L. (2003). What causes a neuron to spike? *Neural Computation, 15,* 1789–1807. [PubMed]

Aguera y Arcas, B., Fairhall, A. L., & Bialek, W. (2003). Computation in a single neuron: Hodgkin and Huxley revisited. *Neural Computation, 15,* 1715–1749. [PubMed]

Bialek, W., & de Ruyter van Steveninck, R. (2005). *Features and dimensions: Motion estimation in fly vision.* arXiv:q-bio.NC/0505003. [Article]

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Scienc*e, 252, 1854–1857. [PubMed]

Brenner, N., Bialek, W., & de Ruyter van Steveninck, R. (2000). Adaptive rescaling optimizes information transmission. *Neuron, 26,* 695–702. [PubMed]

Bussgang, J. (1952). Crosscorrelation functions of amplitude-distorted gaussian signals. *RLE Technical Reports, 216.* Cambridge, MA: Research Laboratory of Electronics.

Chander, D., & Chichilnisky, E. J. (2001). Adaptation to temporal contrast in primate and salamander retina. *The Journal of Neuroscience, 21,* 9904–9916. [PubMed] [Article]

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

Cover, T., & Thomas, J. (1991). *Elements of information theory.* New York: Wiley.

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience.* Cambridge, MA: MIT Press.

de Ruyter van Steveninck, R., & Bialek, W. (1988). Real-time performance of a movement-sensitive neuron in the blowfly visual system: Coding and information transmission in short spike sequences. *Proceedings of the Royal Society of London. Series B, 234,* 379–414.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology, 117*(4), 500–544. [PubMed]

Levine, R. D., & Tribus, M. (Eds.). (1978). *The maximal entropy formalism.* Cambridge, MA: MIT Press.

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

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

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code.* Cambridge, MA: MIT Press.

Ringach, D. L., Sapiro, G., & Shapley, R. (1997). A subspace reverse-correlation technique for the study of visual neurons. *Vision Research, 37,* 2455–2464. [PubMed]

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

Schwartz, O., Chichilnisky, E. J., & Simoncelli, E. P. (2002). Characterizing neural gain control using spike-triggered covariance. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Adv. neural information processing systems* (Vol. 14, pp. 269–276). Cambridge, MA: MIT Press.

Sharpee, T., Rust, N. C., & Bialek, W. (2004). Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Computation, 16,* 223–250. [PubMed]

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

Simoncelli, E., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 327–338). Cambridge, MA: MIT Press.

Touryan, J., Lau, B., & Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *The Journal of Neuroscience, 22,* 10811–10818. [PubMed] [Article]