Published in final edited form as:
Neural Comput. 2012 September ; 24(9): 2384–2421. doi:10.1162/NECO_a_00330.

Characterizing Responses of Translation Invariant Neurons to Natural Stimuli: Maximally Informative Invariant Dimensions

Michael Eickenberg[*,1], **Ryan J. Rowekamp**[*], **Minjoon Kouh**[2], and **Tatyana O. Sharpee**
The Computational Neurobiology Laboratory and the Crick-Jacobs Center for Theoretical and Computational Biology, The Salk Institute for Biological Studies, La Jolla, CA 92037. The Center for Theoretical Biological Physics, University of California, San Diego, La Jolla

Abstract

Our visual system is capable of recognizing complex objects even when their appearances change drastically under various viewing conditions. Especially in the higher cortical areas, the sensory neurons reflect such functional capacity in their selectivity for complex visual features and invariance to certain object transformations, such as image translation. Due to the strong nonlinearities necessary to achieve both the selectivity and invariance, characterizing and predicting the response patterns of these neurons represents a formidable computational challenge. A related problem is that such neurons are poorly driven by randomized inputs, such as white noise, and respond strongly only to stimuli with complex high-order correlations, such as natural stimuli. Here we describe a novel two-step optimization technique that can characterize both the shape selectivity and the range and coarseness of position invariance from neural responses to natural stimuli. One step in the optimization involves finding the template as the maximally informative dimension given the estimated spatial location where the response could have been triggered within each image. The estimates of the locations that triggered the response are subsequently updated in the next step. Under the assumption of a monotonic relationship between the firing rate and stimulus projections on the template at a given position, the most likely location is the one that has the largest projection on the estimate of the template. The algorithm shows quick convergence during optimization, and the estimation results are reliable even in the regime of small signal-to-noise ratios. When we apply the algorithm to responses of complex cells in the primary visual cortex (V1) to natural movies, we find that responses of the majority of cells were significantly better described by translation invariant models based on one template compared with position-specific models with several relevant features.

Keywords

object recognition; translation invariance; dimensionality reduction; information theory; spike-triggered methods

1 Introduction

The ability to recognize objects despite large variations in their position relative to us is a hallmark of animal vision (DiCarlo and Maunsell, 2003; Edelman, 1999; Riesenhuber and Poggio, 1999; Rolls, 2000; Ullman, 1996). Changes in the relative position can cause large changes in the retinal image that greatly exceed the differences between retinal images

[*]Equal contribution
[1]Current address: DMA, Ecole Normale Supérieure and Département de Mathématiques, Ecole Normale Sup rieure de Cachan, France
[2]Current address: Physics Department, Drew University, Madison, NJ 07940, USA

generated by different objects presented at the same location (when differences are quantified using linear projections). The problem is further complicated by changes in the retinal images caused by scaling, rotation, as well as differences in pose, and illumination. For these reasons, understanding the computations required for mediating robust object recognition remains a challenging frontier of both computer science and neuroscience. Although this problem is solved in the brain and fast recognition within a fraction of second is possible (Thorpe et al., 1996), much remains to be understood about the underlying neural mechanisms. In this paper, we describe a spike-triggered method that can help map out how the visual stimuli are represented in the brain.

One of the obstacles for characterizing feature selectivity of these high-level visual neurons is that, because such neurons are tuned to highly specific combinations of visual features, they do not respond well to noise and other stimuli without higher-order correlations. Therefore, up to now feature selectivity of high-level visual neurons has been primarily studied with respect to reduced stimuli (Saleem et al., 1993; Wang et al., 1996, 1998) that are optimized for a particular neuron, sets of controlled naturalistic stimuli, such as isolated face images in a blank background (Desimone et al., 1984; Kobatake and Tanaka, 1994; Logothetis et al., 1995; Rust and DiCarlo, 2010; Zoccolan et al., 2007), and/or parametrized stimulus sets where orientation, curvature, and spirality have been systematically varied (Desimone et al., 1984; Desimone and Schein, 1987; Gallant et al., 1993, 1996; Hegde and Van Essen, 2000, 2007; Kobatake and Tanaka, 1994; Pasupathy and Connor, 1999, 2001). These studies reveal the complexity of the feature selectivity of neurons in the ventral stream, including selectivity to faces and hands, but leave open the possibility that the optimal stimulus for a given neuron may have never been presented. It has also been difficult to predict responses to novel stimuli that have not been used in the experiment.

Instead of using a set of reduced stimuli, an alternative approach is to take advantage of the fact that neurons throughout the ventral stream respond robustly to natural visual stimuli. Although average response rates to natural stimuli may be lower than to the optimal stimulus (Baddeley et al., 1997), they are still significantly higher than those elicited by noise inputs (Rainer et al., 2001). At the same time, natural stimulus ensembles can be made sufficiently diverse such that they sample the neural response along many directions in the stimulus space, albeit not-uniformly (Dong and Atick, 1995; Field, 1987; Ruderman and Bialek, 1994; Simoncelli and Olshausen, 2001; van Hateren and Ruderman, 1998). This approach has the potential to make it possible to characterize the feature selectivity of neurons without making prior assumptions about the actual type of optimal stimulus features. The approach is based on the linear-nonlinear (LN) model (Chichilnisky, 2001; de Boer and Kuyper, 1968; Meister and Berry, 1999; Schwartz et al., 2006; Victor and Shapley, 1980) that describes the neural response as an arbitrary nonlinear function of the stimulus components along the relevant stimulus dimensions. Each of the relevant stimulus dimensions represents a spatiotemporal filter that is applied to incoming stimuli to account for neural responses. While original methods for finding receptive fields were designed to work with noise inputs (de Boer and Kuyper, 1968; Rieke et al., 1997) analogous methods that are valid with natural stimuli have been recently developed for both linear (Gill et al., 2006; Ringach et al., 2002, 1997; Theunissen et al., 2001, 2000; Woolley et al., 2005, 2006) and LN models by several groups (Paninski, 2003; Sharpee et al., 2004; Sharpee, 2007; Sharpee et al., 2006).

Despite the success of spike-triggered methods in characterizing the selectivity of V1 neurons to multiple stimulus features (Chen et al., 2007; Felsen et al., 2005; Horwitz et al., 2007; Rapela et al., 2010, 2006; Rust et al., 2005; Schwartz et al., 2006; Touryan et al., 2005, 2002), both the models of the neural response and the statistical methods used to select them will likely need to be significantly modified in order to be useful in extrastriate

visual areas. The main reason is that, in retinotopic space, accounting for translation invariant selectivity, even to one relevant stimulus feature, requires a model with a large number of relevant dimensions. Although the relevant dimensions each represent the same image feature, they differ in their centering, leading to a high-dimensionality of the resultant LN model. Established methods that can simultaneously estimate a large number of relevant dimensions are guaranteed to work only with Gaussian stimuli, such as white noise or correlated Gaussian noise (Paninski, 2003; Schwartz et al., 2006). These include the methods of spike-triggered covariance (Bialek and de Ruyter van Steveninck, 2005; de Ruyter van Steveninck and Bialek, 1988; Schwartz et al., 2002, 2006), its information-theoretic generalization (Pillow and Simoncelli, 2006), and projection pursuit (Rapela et al., 2010, 2006). On the other hand, methods that estimate multiple filters from neural responses to natural stimuli (Sharpee et al., 2004) can only estimate a few filters because of the need to sample the joint multidimensional dependence of the spike probability on the relevant stimulus components (Rowekamp and Sharpee, 2011; Schwartz et al., 2006). It should be noted that spike-triggered covariance can be applied with natural stimuli, e.g. as in (Touryan et al., 2005), but estimation methods that take higher-order stimulus statistics into account, such as projection pursuit and maximally informative dimensions, yield models with improved predictive power in accounting for V1 neural responses (Rapela et al., 2010).

Our approach here is to reduce the dimensionality by searching for one or several, in practice 2 or 3, image templates while allowing for the possibility that they could be jointly shifted to different positions within the visual space to elicit a spike. The method we propose here combines ideas from methods that address translation invariance but work primarily with noise inputs (Dimitrov et al., 2009; Nishimoto et al., 2006; Tjan and Nandy, 2006) and ideas from methods that can characterize feature selectivity of position-specific models with natural stimuli, e.g. Sharpee et al. (2004, 2006). The overall goal is to address, to our knowledge, a previously unsolved problem of how to characterize the feature selectivity of neurons that exhibit translation invariance based on their responses to natural stimuli. We propose to search for the most relevant feature (or a conjunction of features) for a given neuron, assuming that the probability of triggering the neural response is the same for all retinotopic positions to which the neuron responds (Fig. 1). This is an approximation, because real neurons do not exhibit perfect translation invariance; rather, their responses do decline with distance from the receptive field center (Boussaoud et al., 1991; Desimone et al., 1984; Gross et al., 1969, 1972; Ito et al., 1995; Kobatake and Tanaka, 1994; Leuschow et al., 1994; Logothetis et al., 1995; Missal et al., 1999; Op de Beeck and Vogels, 2000; Richmond et al., 1983; Sary et al., 1993; Schwartz et al., 1983; Tovee et al., 1994). However, this approximation provides an approach for mapping out the receptive fields of high-level neurons that is complementary to the conventional spike-triggered approaches that work at a given retinotopic location only. At the same time, comparison of predictive power achieved by the two kinds of models – either translation invariant or non-translation invariant – could also be helpful in quantifying the emergence of invariance across the hierarchy of sensory representations.

In sum, the central goal of this work is to develop and test the computational methods that can estimate the relevant stimulus features of a neuron under the assumption that the neural response can be triggered at different positions within the visual field. We seek a method that:

- will work with arbitrary stimuli, including natural stimuli;
- should in principle be capable of recovering an arbitrary complex stimulus feature $\vec{v}$, i.e. we do not assume that relevant stimulus features can be parametrized as Gabor functions (DeAngelis et al., 1993) or curved line elements; rather, these shapes should emerge as a result of the analysis;

- will allow for nonlinearities in the neural responses in order to describe such properties as rectification and saturation;
- produce an estimate of the range and coarseness of translation invariance.

We note that the problem of characterizing translation invariant feature selectivity can be thought of as complementary to the perceptual task of representing visual scenes in the presence of uncertainties introduced by fixational eye movements (Burak et al., 2010). Here, our goal is to estimate one or a few templates based on known stimuli and inferred relevant locations for each stimulus, whereas in the perceptual task the goal is to infer unknown images based on known receptive fields of neuron and the inferred eye positions.

This paper is organized as follows: In Sec. 2 we describe an approach for characterizing feature selectivity of neurons whose responses exhibit translation invariance. Sec. 3 presents results on both model and real neurons (V1 complex cells). Sec. 4 contains concluding remarks, and Sec. 5 describes the methods.

2 Accounting for translation-invariant neural responses

A biologically plausible model that is consistent with the definition of translation-invariance is based on the combination of the responses of position-specific neurons according to a logical OR operation (Cadieu et al., 2007; Fukushima, 1980; Pelli, 1985; Riesenhuber and Poggio, 1999). Mathematically, the probability that stimulus $\vec{s}$ will elicit a spike from a neuron selective for a template $\vec{v}$ at different locations within the visual field can be written as:

$$P(\text{spike}|\vec{s}) = 1 - \prod_{\vec{z} \in G} (1 - \bar{r} f(\vec{s} \cdot T_{\vec{z}} \vec{v})). \tag{1}$$

Here, $\vec{z} \in G$ represents a set of all possible position shifts, $T_{\vec{z}}\vec{v}$ represents a particular positioning of the feature $\vec{v}$ at a center location described by shift $\vec{z}$. The function $f(\vec{s} \cdot T_{\vec{z}}\vec{v})$ describes the normalized probability of eliciting spikes from the presumed hidden units, whose responses are not translation invariant but are specific to a particular positioning $T_{\vec{z}}\vec{v}$ of the image feature $\vec{v}$; $\bar{r}$ is the average response probability of a hidden unit. We will refer to this image feature $\vec{v}$ as the template because it is the same for all hidden units.

Another biologically plausible model that yields translation invariance is based on selecting the maximal response:

$$P(\text{spike}|\vec{s}) = \bar{r} \max_{\vec{z} \in G} f(\vec{s} \cdot T_{\vec{z}} \vec{v}). \tag{2}$$

The experimental support for this model was also found, e.g. in area V4 (Gawne and Martin, 2002). Although with non-binary and probabilistic hidden units, the MAX model will yield a different predicted firing rate compared to the model based on the logical OR, below we show that the templates of both models can be estimated using the same two-step optimization procedure.

2.1 Review of estimation methods for position-specific neurons

In developing methods for characterizing invariant feature selectivity, we will build on the statistical methods developed for characterizing neural feature selectivity in the absence of invariance, which we now briefly review. Bearing in mind that high-level neural responses are likely to be more responsive to complex stimuli, such as natural stimuli, we focus on

methods that are applicable in this case (Kouh and Sharpee, 2009; Paninski, 2003; Schwartz et al., 2006; Sharpee et al., 2004; Sharpee, 2007). Without translation invariance, Eqs. (1) and (2) simplify to the model based on just one relevant stimulus feature $\vec{v}$:

$$P(\text{spike}|\vec{s}) = P(\text{spike}) f(x), \qquad x = \vec{s} \cdot \vec{v}. \tag{3}$$

The problem of finding the relevant dimension is illustrated in Fig. 2. A stimulus can be represented as a point in a high-dimensional space. A change in the stimulus component along the relevant dimensions modulates the neural response (x_1-axis in the figure), while a change along an irrelevant dimension (i.e., one that is orthogonal to the relevant dimension) will not influence the response unless the stimulus components along the two dimensions are correlated. Thus, the relevant dimension can be found by comparing distributions $P(\vec{s})$ and $P(\vec{s}|\text{spike})$ along various dimensions, and selecting the dimension along which these distributions are most different. The intuition for this strategy is that stimulus projections x along an irrelevant stimulus dimension will be weakly correlated with the occurrence of a spike. Because the spikes will have occurred with similar or equal probability for all values of x, the distributions $P(x)$ and $P(x|\text{spike})$ will be similar to each other along the irrelevant dimensions. In contrast, these two probability distributions will be quite different along the relevant dimension $\vec{v}$. The dissimilarity between two probability distributions can be quantified by a number of divergence measures (Paninski, 2003; Sharpee, 2007). However, the smallest unbiased estimation error is obtained by maximizing the Kullback-Leibler (KL) divergence (Kouh and Sharpee, 2009):

$$D_{KL}(P(x|\text{spike}) \,||\, P(x)) = \int dx P(x|\text{spike}) \log_2 [P(x|\text{spike})/P(x)]. \tag{4}$$

In the limit of low spike probabilities, the above quantity corresponds to the mutual information between stimulus components along the relevant dimension and the arrival times of single spikes (Adelman et al., 2003; Sharpee et al., 2004). The mutual information (4) is small when projection value x and spike times are relatively independent, because in this case distributions $P(x)$ and $P(x|\text{spike})$ are similar. On the other hand, when evaluated along the relevant dimensions, the KL divergence will take its maximal value equal to the mutual information between full stimuli and single spikes (Brenner et al., 2000):

$$I_{\text{spike}} = \int d\vec{s} P(\vec{s}) \frac{P(\text{spike}|\vec{s})}{P(\text{spike})} \log_2 \left[\frac{P(\text{spike}|\vec{s})}{P(\text{spike})} \right] \tag{5}$$

This suggests that the relevant dimensions can be found by maximizing the mutual information between x and the spike probability according to Eq. (4). This approach has been implemented with a combination of line optimization and simulated annealing (Press et al., 1992) to analyze the neurons from the visual system (Sharpee et al., 2008, 2006), with subsequent extensions to recover multiple features of both visual (Rowekamp and Sharpee, 2011; Sincich et al., 2009) and auditory (Atencio et al., 2008, 2009) neurons.

2.2 Two-step optimization for finding the relevant stimulus dimensions of translation invariant neurons

Characterizing feature selectivity with invariance is significantly more challenging, because the responses of a translation invariant neuron are based on stimulus features at multiple locations within the visual field. The first step in our analysis of an invariant neuron is to determine the most likely location where the response could have been triggered. The second step then uses the stimuli centered at these locations to improve the estimate of the

template as in the non-invariant case in Section 2.1. These two steps: 1) updating the estimated locations and 2) updating the estimated template, are repeated until convergence.

To perform the first step of optimization, we will at first consider separately stimuli that elicited and did not elicit a spike from the invariant neuron. For a stimulus $\vec{s}$ that elicited a spike, one can transform the probability $P(\vec{z}|\text{spike}, \vec{s})$ that the hidden unit at a location characterized by shift $\vec{z}$ from the center has produced a spike according to the Bayes' rule:

$$P(\vec{z}|\text{spike}, \vec{s}) \propto P(\vec{z})P(\text{spike}|\vec{z}, \vec{s}) = P(\vec{z})f(\vec{s} \cdot T_{\vec{z}}\vec{v}). \tag{6}$$

Here, $P(\vec{z})$ is the prior probability that the response could have been triggered at a shift position $\vec{z}$ and function $f(\vec{s} \cdot T_{\vec{z}}\vec{v})$ is the nonlinear gain function of the hidden units, which was first defined in Eq. (1). According to Eq. (6), if the gain function $f(x)$ is monotonic and *a priori* all locations are equally likely to elicit a spike, then the most likely location to have triggered the neural response is the one that yields the greatest projection value x between the translated template $T_{\vec{z}}\vec{v}$ and the stimulus $\vec{s}$.

For stimuli that did not elicit a spike, there is no uncertainty as to what happened at each of the possible locations – we know that none of the hidden units has produced a spike. For these trials we can associate any patch with the neural response for analysis using a position-specific model. Here we also select the patch with the greatest projection value as the one that was most likely to trigger the response (according to the current model), but did not. This choice corresponds to a maximal reduction in the entropy of the current model from incorporating the measured response, and thus is an example of maximally informative data point selection (Mackay, 1992). To summarize, in the first step of optimization we determine for each stimulus the maximum projection $\max_{\vec{z} \in G} \vec{s} \cdot T_{\vec{z}}\vec{v}$ across different shifts $\vec{z}$ of the template $\vec{v}$.

To perform the second step of the optimization, we form the probability distributions $P^{\max}_{\vec{v}}(x)$ of these maximal projection values both across all stimuli:

$$P^{\max}_{\vec{v}}(x) = \int_{\mathbb{R}^D} \mathrm{d}\vec{s}\, P(\vec{s})\delta\left(x - \max_{\vec{z} \in G} \vec{s} \cdot T_{\vec{z}}\vec{v}\right), \tag{7}$$

and across all stimuli that elicited a spike:

$$P^{\max}_{\vec{v}}(x|\text{spike}) = \int_{\mathbb{R}^D} \mathrm{d}\vec{s}\, P(\vec{s}|\text{spike})\delta\left(x - \max_{\vec{z} \in G} \vec{s} \cdot T_{\vec{z}}\vec{v}\right). \tag{8}$$

In Eqs. (7) and (8), $\mathbb{R}^D$ denotes the D-dimensional stimulus space (e.g. a set of images of D pixels). Step 2 is completed by maximizing the KL divergence between these two probability distributions:

$$I(\vec{v}) = \int_{\mathbb{R}} \mathrm{d}x P^{\max}_{\vec{v}}(x|\text{spike}) \log\left(\frac{P^{\max}_{\vec{v}}(x|\text{spike})}{P^{\max}_{\vec{v}}(x)}\right) \tag{9}$$

to obtain a new estimate of template $\vec{v}$.

To aid the high dimensional optimization process of $I(\vec{v})$ with respect to $\vec{v}$, an analytical gradient function can be calculated. It reads

$$\nabla_{\vec{v}} I(\vec{v}) = \int_{\mathbb{R}} \mathrm{d}x P^{\max}_{\vec{v}}(x) \left(\langle \vec{s} | x, \text{spike} \rangle - \langle \vec{s} | x \rangle \right) \frac{\mathrm{d}}{\mathrm{d}x} \frac{P^{\max}_{\vec{v}}(x|\text{spike})}{P^{\max}_{\vec{v}}(x)} \tag{10}$$

where the response and projection value dependent averages $\langle \vec{s}|x, \text{spike} \rangle$ and $\langle \vec{s}|x \rangle$ are defined as

$$\langle \vec{s} | x, \text{spike} \rangle := \int_{\mathbb{R}^D} \mathrm{d}\vec{s} \, P(\vec{s}|\text{spike}) T_{-\vec{z}_m} \vec{s} \, \delta \left(x - \max_{\vec{z} \in G} \vec{s} \cdot T_{\vec{z}} \vec{v} \right) / P^{\max}_{\vec{v}}(x|\text{spike}) \tag{11}$$

$$\langle \vec{s} | x \rangle := \int_{\mathbb{R}^D} \mathrm{d}\vec{s} \, P(\vec{s}) T_{-\vec{z}_m} \vec{s} \, \delta \left(x - \max_{\vec{z} \in G} \vec{s} \cdot T_{\vec{z}} \vec{v} \right) / P^{\max}_{\vec{v}}(x), \tag{12}$$

where $\vec{z}_m$ is the translation with the strongest projection responses, i.e. $\vec{z}_m = \text{argmax}_{\vec{z} \in G} \vec{s} \cdot T_{\vec{z}} \vec{v} \cdot T_{-\vec{z}_m} \vec{s}$ denotes the stimulus patch that yielded the largest projection onto $\vec{v}$.

The proposed two-stage approach for estimating translation-invariant templates is reminiscent of the classic EM algorithm (Dempster et al., 1977), but differs from it both in the structure of the problem and the approach. In particular, the variables that describe which hidden units could have caused a spike in a translation-invariant neuron are not mutually exclusive, whereas the EM algorithm for mixture models (Bishop, 2004) operates with probabilities of hidden variables that are, although not directly observable, can take only one value at a time with some probability. This difference is most pronounced for stimuli that did not elicit a spike, in which case we know that inputs to none of the hidden units have exceeded the spike threshold, and thus we can choose any of them for the ensemble of stimuli/response associations from which templates will be estimated. In what follows we demonstrate, using two different implementations, that the proposed two-step optimization does converge on the correct stimulus features when tested on a series of model neurons built according to architecture of Eq. (1).

2.3 Computing Maximally Informative Invariant dimensions: direct and Fourier approaches

Two practical implementations are possible to complete step 2 – finding maximal projections between stimuli and templates shifted to different locations. Here, one possibility is to extract stimulus patches of smaller dimensionality d from different parts of the stimulus and compare them to the template, cf. Fig. 3A. The other possibility is to work with templates and stimuli of full dimensionality D. In this case one needs to use periodic boundary conditions, which is a common assumption when treating translation-invariance (Tjan and Nandy, 2006), to compute projection values with templates shifted to different positions, cf. Fig. 3B.

We refer to the first approach as the “direct approach” and to the second as the “Fourier approach.” This is to emphasize that the second approach can be conveniently implemented by taking a two-dimensional Fourier transform of each stimulus and the template. Multiplying the Fourier transform of the template by $e^{i\vec{k}\cdot\vec{z}}$ yields the Fourier transforms of the template after a shift by a vector $\vec{z}$ (here, $\vec{k}$ enumerates the Fourier components). The implicit assumption (the validity of which can be checked after the optimization is complete) is that the spatial extent of the neural template is small enough not to be affected by moderate translations from the center of the stimulus window, i.e. that the magnitude of the parts of template that wrap around is close to zero. This property is more strictly enforced in the direct method, where the size of the template by construction cannot exceed the size of

the extracted patches and needs to be specified *a priori*. The Fourier approach can provide clues as to the properties of the template, even if it is affected by some of the larger translations used during the estimation. If that happens, the configuration of the translation grid can be adjusted to a smaller range of translations or the sides of the stimulus frame can be zero padded to increase the ratio between the stimulus and template sizes.

The direct and Fourier methods offer complementary capabilities in a number of other respects, including the dimensionality of the neural template, the size of the grid over which the responses are pooled to approximate translation invariance, as well as the computational requirements. For example, the direct approach is limited by the number of patches that can be stored in the memory of a computer. Typical capabilities of current generation computers limit analysis of translation to tens of points, so that grids of at most 5×5 or 7×7 points can be analyzed. Re-extracting patches for each calculation avoids the memory limitations, but is computationally slow. In contrast, the Fourier approach can handle translational grids G of almost any size, and thus can better approximate continuous translation invariance. On the other hand, because of periodic boundary conditions assumed in the Fourier approach, the template has to spatially extend significantly beyond its non-zero part. This reduces either the signal-to-noise ratio or the resolution of templates recovered by the Fourier approach. In summary, the direct and Fourier approaches for estimating translation invariant templates allow one to choose between analyzing either finer translation grids and more coarse-grained templates using the Fourier approach or coarser translation grids and finer resolution templates using the direct approach.

We note that with both approaches the performance can be compared over different configurations of translation grids G, as we do below in Sec. 3.3, in order to determine the most appropriate configurations for each neuron under consideration.

3 Results

In this section, we present results on synthetic and real neurophysiological data. In order to validate the model in a controlled setting, we ran extensive tests on model cells. Later, the method was applied to V1 complex cells and put into contrast with methods that do not account for translation invariance.

To test the algorithms for estimating translation invariant feature selectivity, we designed a series of model translation invariant neurons according to Eq. (1) probed with natural movies (Sharpee et al., 2006). The responses of the translation invariant model cell were simulated using a logical OR operation to pool responses of a number of afferent units. The afferent units were all selective for the same preferred stimulus feature $\vec{v}$, but were centered at different positions within the visual space. The responses of the afferent units were modeled using a noisy thresholding operation: an afferent unit responded with a spike if the stimulus patch at the corresponding position had a projection onto $\vec{v}$ that exceeded a threshold value of θ in the presence of Gaussian noise with zero mean and variance σ^2. The parameter σ thus can be used as a measure of neural noise. All afferent units contributing to the translation invariant cell had the same value of θ and σ. In this way, the nonlinear gain function $f(\vec{s} \cdot T_{\vec{z}}\vec{v})$ in Eq. (1) becomes a sigmoid. Because all hidden units have the same parameters except for the centering of the relevant template, and their responses are pooled using a logical OR, the resultant model cell approximates translation invariance over the region of the visual space spanned by the centers of the hidden units. We also varied the size and coarseness of the spatial translation grid formed by the centers of stimulus patches representing the hidden units. Table 1 describes the range of parameters tested in combination with three types of translational grids.

3.1 Position-specific model fails to estimate translation invariant feature selectivity

We begin by demonstrating that translation invariance has profound effects on the template estimation. Fig. 4 presents results of a characterization that does not take invariance into account applied to a model cell with translation invariance. The model template (the relevant image feature of hidden units) is shown in Fig. 4A. It was taken to be a curved Gabor in order to mimic properties of visual extrastriate neurons (Connor et al., 2007; Pasupathy and Connor, 1999, 2001, 2002). In Fig. 4A crosses mark the centers of 3 × 3 translation grid that was used to approximate translation invariance. Because of translation invariance, the neuron produces a spike if any of the stimulus patches taken around the nine locations of the translation grid provides a sufficiently good match to the template. This results in the response region of this neuron being much larger than the spatial extent of the template. Correspondingly, the non-translation invariant optimization algorithm seeking a single relevant feature produces a filter that is a smeared superposition of the relevant templates placed at the centers of the translation grid (which describe the hidden units). It is not possible to guess from this estimation even the rough shape of the underlying template (cf. panel C with the template in panel A). The estimation also does not capture the correct form of the nonlinear gain function. The nonlinear gain function with respect to the maximal projection value across the translation grid is a sigmoidal function (model nonlinearity, panel B). In contrast, the estimated nonlinearity shows similar sensitivity to both positive and negative projections onto the estimated filter (solid line in panel D). Furthermore, even when the nonlinear gain function is evaluated along the correct template but at a fixed position in the visual space, the increase in the spike probability is still observed for negative projection values of the stimuli onto the relevant template (dashed line in panel D). This effect is due to the overlap between templates centered at different positions of the translation grid, so that negative projection values onto the template at the center positions actually signal the presence of positive projection values onto templates centered at neighboring positions on the translation grid. In sum, ignoring translation invariance prevents the correct estimation of both the relevant template and the nonlinear gain function of translation invariant neurons.

3.2 Characterizing feature selectivity of translation invariant model neurons

We now show that both the nonlinear gain functions and the relevant template of translation invariant neurons can be estimated by searching for maximally informative invariant dimensions. Fig. 5 shows estimation results of the same translation invariant model neuron from Fig. 4A, using the direct and the Fourier methods. The estimated templates have large dot products with the model templates (the dot products are computed between normalized vectors, so that the perfect estimation without any noise yields 1): $c = 0.897 \pm 0.008$ (Fourier method, panel C, error-bars represent standard errors of the mean) and $c = 0.899 \pm 0.011$ (direct method, panel D). Another measure that can be used to quantify the estimation accuracy is the ratio of information accounted for by the estimated template to that accounted for by the model template. The latter information quantity represents the overall information that is available in single neural responses (Adelman et al., 2003; Sharpee et al., 2004). The corresponding fraction of the total information explained by the estimated filters was also close to its maximal value of 1: 0.963 ±0.006 for the Fourier method and 0.969 ± 0.008 for the direct method. These values were much larger than those obtained using a position-specific template from Fig. 4. The high predictive power is also visually obvious when one compares the post-stimulus time histograms (PSTH) of this model neuron on a novel set of natural scenes with predictions PSTHs obtained using the translation invariant estimation model (Fig. 5G). Finally, we point out that the algorithms typically converged quite rapidly. In Fig. 6 we show an example of convergence during template estimation for four jackknife data sets obtained for the same model cell as shown in Figs. 4 and 5. Within 100 line optimizations in the $D = 256$-dimensional template space for this model neuron the

algorithm converges to a value that is quite close to the final outcome. Similar behavior is observed both in terms of information explained on a novel data set (panel A) and in terms of projection onto a model template (panel B). In summary, the estimation of translation invariant templates appears to be quite robust and achievable within a number of iterations that is smaller than the template dimensionality.

The configuration of the translation grid represents another important parameter of the estimation algorithm. In the above calculation, the translation grid used during estimation coincided with that of the model translation invariant neuron. We find, however, that even when the translation grids used during estimation differed from those of the model cell, reasonably accurate estimations can be obtained. For example, panels E and F of Fig. 5 show results of the estimation using a 5 ×5 translation grid for the Fourier and the direct method. Although this translation grid is substantially different from the 3 × 3 translation grid that was used in the model, the estimated templates are visually quite similar to the model one, and have dot products of 0.789 ± 0.003 and 0.78 ± 0.02 for the Fourier (panel E) and direct (panel F). The corresponding values of the fraction of total information explained by estimated filters (that takes into account that estimated templates can be translated versions of the model template) are 0.826 ± 0.003 and 0.826 ± 0.002 for the direct and Fourier methods, respectively. Thus, the estimation of the preferred image feature appears to be robust in the presence of disparities between the translation grid of the neuron and that used during the estimation. At the same time, the predictive power of the recovered model was somewhat lower in the presence of a mismatch between the translation grid of the model and estimation. We next explore whether this observation can be used to characterize the coarseness in the translation invariance of a neuron.

3.3 Resolving the coarseness of translation invariance

In addition to determining the relevant stimulus feature for a translation invariant neuron, it is also helpful to determine the range and perhaps the coarseness that characterizes its translation invariance. The spatial extent can be measured directly by observing how the neural response to the preferred stimulus feature decreases with distance from the receptive field center, and indeed detailed measurements have shown that responses of high-level visual neurons decrease with distance from the receptive field center (Desimone and Schein, 1987; Pasupathy and Connor, 2001; Pollen et al., 2002; Rust and DiCarlo, 2010). However, one would also expect to find discrete aspects in the neural implementation of translation invariance, as can already be observed in the retina (Field et al., 2010; Liu et al., 2009; Soo et al., 2011; Soodak et al., 1991). Not knowing this coarseness *a priori*, one would like to have a method to find it. Above we have shown that the relevant stimulus feature or template can be estimated quite closely, even in the presence of a mismatch between the true translation grid and the grid used during the estimation. However, the disparity between the model and the assumed translation grids during estimation resulted in the reductions of both the dot product coefficient and the percent information explained by the estimated filter. Here, we examined whether this decrease is sufficient to determine the underlying translation invariance properties of a neuron, such as its coarseness.

Fig. 7 shows how the percent information explained changes as a function of mismatch in coarseness between the model and estimation translation grids. For example, in panel (A) we analyze the model cell with a 3 × 3 translation grid using translation grids ranging from no translation invariance (1 × 1 grid) to near perfect translation invariance (9 × 9 grid corresponds to translation by two pixels). We find a clear peak when the estimation grid matches that of the model in terms of the percent information explained (panel B). The difference of the peak value from the neighboring points is significant ($p < 10^{-4}$, t-test, panel A). Thus, the algorithm can correctly identify the coarseness of translation invariance of this model cell. Analyzing results for models with a finer 5 × 5 translation grid (panel B), we

find that it is possible to rule out a coarser grid of 3 × 3 compared to the true grid ($p < 10^{-4}$), but that a finer 9 × 9 grid gives the same predictive power as the model grid ($p = 0.16$). When the analysis is taken to the limit of perfect translation invariance (17 × 17 grid in our case), we continue to observe an increase in predictive power when the translation grid is refined from 5 × 5 to 9 × 9 ($p < 10^{-4}$). Overall, these results suggest that it is possible to determine a lower limit for the coarseness of the translation grid. For cells with a rather coarse translation grid, such as when the smallest translation is about 1/4 of the overall response region (corresponding to 3 × 3 grid in our simulations), both the upper and lower limits on the coarseness of the translation grid may be determined.

3.4 Convergence with increasing data set size

An important practical consideration is how the proposed methods perform not only in a well-sampled regime where the number of trials (and spikes) greatly exceeds the stimulus dimensionality but also in much more typical cases where the two numbers are comparable. Therefore, we have analyzed estimated templates as a function of data set size for model cells that had different intrinsic noise levels. Each of these model cells was probed by the same stimulus sequence that was repeated a different number of times, from 1 to 20 times. Simulations were done using the Fourier approach, because it permits larger stimulus dimensionality D (32 × 32 frames yield $D = 1024$). Fig. 8 describes results for model cells with different thresholds, translation grids, and different noise levels. As expected, we found that the dot product between the estimated and the model template improved with increasing number of spikes. Furthermore, the improvement was more pronounced for cells with greater levels of intrinsic noise. Typically, a steeper slope was observed for noisier cells (black, $\sigma = 1.0$) than the less noisy cells (light gray, $\sigma = 0.5$), since the reduction in uncertainty is more significant for each added repetition in the noisier cell. However, estimations with dot products greater than 0.85 were obtained in all cases. These results demonstrate the feasibility of estimating feature selectivity of translation invariant neurons for data sets containing a few thousand spikes, which is achievable with current physiological techniques.

3.5 Analysis of V1 complex cells responses to natural movies

We now use the two-step optimization to characterize feature selectivity of V1 complex cells. The responses of V1 neurons are sensitive to the presence of multiple stimulus features (Chen et al., 2007; Rust et al., 2005; Touryan et al., 2002). The complex cells are thought to implement one of the first steps in building position-invariant representations, and their responses are consistent with being triggered by spatially shifted image patterns (Rust et al., 2005). Thus, we set out to explore whether the translation invariant models can provide a better description of their responses than the position-specific models with up to three features.

For each neuron, we estimated both position-invariant models and position-specific models based on its responses to natural movies (see Methods). The templates of translation invariant models and position-specific models now also included a temporal dimension comprised of three time lags. The templates were not assumed to be separable in space and time. For both position-specific and position-invariant models, we have allowed for the possibility that the spike probability can depend on the conjunction of features. In the framework of a translation invariant model, this assumes that the output of hidden units depends on several templates, such as $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$, that are evaluated at a given position. We model the position invariant response as

$$P(\text{spike}|\vec{s}) = \bar{r} f(\vec{s} \cdot T_{\vec{z}_{max}} \vec{v}_1, \vec{s} \cdot T_{\vec{z}_{max}} \vec{v}_2, \vec{s} \cdot T_{\vec{z}_{max}} \vec{v}_3) \quad (13)$$

where $\vec{z}_{max}$ is the grid location at which $x_1 = \vec{s} \cdot T_{\vec{z}} \vec{v}_1$ is maximized for a particular stimulus $\vec{s}$. Template $\vec{v}_1$ was found first for a one-dimensional model, and the projections on $\vec{v}_1$ determined which grid location we associated with the neural response. Additional templates $\vec{v}_2$ and $\vec{v}_3$ are found subsequently to create a two- and three-dimensional model, respectively, with the projections on the stimulus at the location selected by $\vec{v}_1$ modulating the neuron's response. The templates found using the model (13) will also be valid for a logical OR model in cases where the maximum of the nonlinear gain function $f(x_1, x_2, x_3)$ of hidden units occurs along the first dimension. This is the case for classical models of contrast gain control where the response of a hidden unit is a function of one (most relevant) stimulus component normalized by signal components along other dimensions (Heeger, 1992; Schwartz et al., 2002; Schwartz and Simoncelli, 2001).

In choosing the configuration of the translational grid, we were guided by previous results that V1 neurons have ~ 10 (≤18) subunits per neuron whose center positions are closely spaced (Rust et al., 2005). The position-invariant models were computed using a 3 × 3 grid with a spacing of 1 pixel, similar to Rust et al. (2005). This translation grid would yield 9 subunits in the case of one translated template and 27 subunits for models based on three translated features, which is on the order of the range of experimentally observed numbers of subunits.

Across our population of 53 V1 complex cells, we find that both position-specific and position-invariant models could account for a larger amount of information in the neural response when more templates were included (Fig. 9, see Methods for details of information calculation). However, an interesting transition was observed with increasing the number of features. The translation invariant models with one relevant template accounted for significantly ($p < 10^{-5}$, Wilcoxon signed-rank test) more information in the neural responses than position-specific models with one relevant feature. The same comparison was true for models with two features ($p = 0.0005$, Wilcoxon signed-rank test, panel B). With three relevant templates, the position-specific models performed as well, across the population, as translation invariant models ($p = 0.06$, Wilcoxon signed-rank test, panel C). Thus, position-invariant and position-specific models offer complementary paths to approximate the neural computations observed across the population of V1 complex cells. At the same time, there were individual neurons whose responses could be predicted substantially better by either the position-invariant or the position-specific models. In Figure 10, we show the estimation of the two kinds of models for a V1 complex cell that was better described with a position-invariant three-template model than with a position-specific model. Figure 11 shows estimation results for a complex cell that was better described by a position-specific three-template model than by a position-invariant model. Tables 2 and 3 show the performance of one-, two-, and three-dimensional position-invariant and position-specific models according to a number of measures of predictive power, including the correlation coefficients of predicted firing rate with the average firing rate for example cells shown in Figure 10 and Figure 11, respectively. Comparison between these two examples suggests that relevant stimulus features estimated with a better performing model have higher signal-to-noise ratio (represented in the color map) and are also more localized in space. For the example neuron in Fig. 10 that was better described with an invariant model, the relevant stimulus features of the position-specific model are more spatially distributed than those of the position-invariant model. Similarly, for the example neuron in Fig. 11 that was better described with a position-specific model, the relevant stimulus features of the invariant models are more blurred. When fitting a position-invariant model to a position-specific unit, each of the model locations could fit the unit with a translated template. Since they use the same

template, the model template becomes the average of the translated templates, which is a blurred version of the position-specific template. Likewise, when fitting a position-specific model to a position-invariant unit, the model could fit each of the locations with a translated template which results in a blurred template. Thus, the mismatch between the structure of the underlying neural computation and the estimation model is likely to result in the blurring of the relevant stimulus features, as also shown with a model neuron in Figure 4C.

In addition to comparing position-specific and position-invariant models with the same number of features, one can ask whether models with a single translation invariant template can outperform position-specific models with multiple features. To carry out this comparison, we recall that models with a smaller number of dimensions are at an inherent disadvantage because adding even a random dimension to the model will almost surely improve information explained (Fairhall et al., 2006). This is because information characterizes predictive power of a given set of features up to any one-to-one transformation of the nonlinearity (the nonlinear gain function is recomputed for a given data set). The information gain from adding a random dimension is not artefactual perse, because random dimensions will always have a small component along relevant dimensions. With natural stimuli, this can lead to appreciable information gain (Sharpee et al., 2004), making it difficult to compare models with different numbers of features. Therefore, to compare models with different number of features we used a correlation coefficient between the predicted and measured firing rates on a test set, under conditions where both the features and the nonlinear gain function were computed from the training data set (see Methods). Unlike the information, this quantity should decrease if more features are added into the model than necessary to explain the responses. Here we find that majority of V1 complex cells (37/53) are better described by a translation invariant model based on just one feature than by a position-specific model with three features (Fig. 9D, signed-rank test across the population yielded $p = 0.0001$). Furthermore, across the population the mean correlation coefficient decreased with the addition of extra features to the translation invariant model ($p = 0.01$ was obtained from signed-rank test for population comparisons both between one-feature vs two-feature translation invariant models and between two-feature vs three feature translation invariant models). Panels E and F show comparisons between two- and three-feature translation invariant models compared with three-feature position specific model. In sum, the fact that position-invariant models yield improved predictive power over position-specific models serves as proof-of-principle that the proposed two-step optimization can provide useful characterization of neural responses.

4 Summary

This paper considered the problem of finding relevant image features in the situation where they can appear anywhere within the visual field to trigger the neural responses. We focused on estimating these relevant image features from neural responses to natural stimuli because neurons in the corresponding high-level visual areas typically respond poorly to randomized images, such as white noise, and require the presence of structured image features to produce robust responses. Our method characterizes translation invariant feature selectivity using an iterative two-step optimization. The first step involves obtaining the estimates of locations associated with the neural response based on the initial estimate of the relevant image feature. The second step involves updating the optimal image feature given an estimate of the location within each image responsible for triggering the neural response. We found that such a two-step optimization can produce reliable estimates of both the relevant image features and the nonlinear transformation describing how the stimulus similarity to the relevant image feature increases the neural spike rate (Fig. 5). Furthermore, the algorithm can provide estimates of the coarseness of the translation grid that is most consistent with the data. In most cases, the appropriate coarseness of the translation grid can

be determined as the coarsest that is consistent with the data. This is because considering finer than necessary translation grids did not typically lead to a decrease in predictive power compared to the model translation grid (Fig. 7B). These results mirror those reported with psychophysical data by Tjan and Nandy (2006) where the lower bound on the spatial range of stimulus uncertainty could be determined much more precisely than the upper bound. At the same time, we find that for cells with a coarse translation grid, both finer and coarser translation grids could be distinguished from the one used in the model based on the decrease in the resulting predictive power (Fig. 7A).

From a practical standpoint, we considered two approaches (direct and Fourier) for characterizing translation invariant feature selectivity. For large data set sizes and available computational resources, both the direct and Fourier approaches will yield converging estimates of the relevant image features. However, the two approaches are complementary in terms of their trade-offs between the sizes of the translation grid and the relevant image features that they can handle. The Fourier approach can typically handle finer translation grids but will yield coarser (or less reliable) estimates of the relevant image template, than the direct approach. At the same time, the convergence result of the Fourier approach (Fig. 8) is encouraging, as the projection between the model and estimated relevant image features were greater than 0.85 (for the perfect estimation, the projection value would be precisely 1), even in the regime of undersampled data sets where the number of spikes was less than the stimulus dimensionality.

Using the new algorithm to characterize responses of V1 complex cells to natural stimuli, we found that, across the population, neural responses were equally well described by both the translation invariant model and the position-specific model with three features. This suggests that the two models provide complementary approaches for characterizing responses of V1 neurons. At the same time, Mechler and Ringach (2002) noted that the standard (and so far the only available) measure for classifying simple and complex cells in V1 based on responses to moving gratings might not be appropriate. This leaves open the possibility that the set of complex cells we analyzed might be actually comprised by cells that perform different types of computations. We found that some neurons in our population were substantially better described by a position-invariant model with three templates than by a position-specific model (Fig. 9C). At the same time there were other neurons for which position-specific models worked significantly better. Furthermore, using correlation coefficients between measured and predicted firing rates on a novel data set, we find that even models with a single translation invariant feature can yield better predictive power than models with three position-specific features (Fig. 9D). In sum, translation invariant models represent an alternative and complementary way of characterizing responses of V1 neurons compared to existing methods.

The described approaches can be extended to other types of invariance, such as scaling. This can be done with the current algorithm by augmenting stimuli with those from different scales and expanding the grid of possible translations to include points corresponding to stimuli of different scales. Finally, we would like to emphasize that although we have focused on characterizing responses of visual neurons that show tolerance to translation of the preferred image feature, the described methods are statistical in nature, and can be used for analyzing responses of neurons in other sensory modalities that show invariance to appropriate transformations, such as pitch and tempo for high-level auditory neurons.

5 Methods

5.1 Analysis of V1 responses

The responses of V1 complex cells were recorded while the animal was presented with natural movies and were collected as part of a previous study (Sharpee et al., 2006). The data set for each neuron consisted of three sets of responses: responses to a relatively long sequence (~ 10 min) of different natural scenes ("unrepeated" data set), responses to a shorter stimulus sequence (~ 10 sec) repeated 55 times ("repeated" data set), and responses to moving gratings of optimal orientation and spatial frequency. In some neurons, the responses to multiple blocks of these kinds of stimuli were also available. Neurons were selected as complex if the modulation of their responses to moving gratings at the stimulus frequency $F1$ was less than the mean elicited firing rate $F0$ (Skottun et al., 1991). Natural movies were presented at 30 Hz; both stimuli and spike trains were binned into 33 msec time bins. Multiple occurrences of spikes in a bin were added (responses were not binarized).

5.2 Finding relevant stimulus features

For position-specific LN models, the relevant stimulus features were computed as dimensions in the stimulus space that accounted for the maximal amount of information in the neural response (Sharpee et al., 2004). The first maximally informative dimension (MID) was found by maximizing the KL divergence in Eq. (4) between the probability distribution

$$P_{\vec{v}}(x) = \int_{\mathbb{R}^D} \mathrm{d}\vec{s}\, P(\vec{s})\, \delta(x - \vec{s}\cdot\vec{v}) \tag{14}$$

and

$$P_{\vec{v}}(x|\mathrm{spike}) = \int_{\mathbb{R}^D} \mathrm{d}\vec{s}\, P(\vec{s}|\mathrm{spike})\, \delta(x - \vec{s}\cdot\vec{v}). \tag{15}$$

When computing the spike-conditional probability distribution $P_{\vec{v}}(x|\mathrm{spike})$, projections from a given stimulus were included as many times as the number of spikes elicited by this frame. The resulting histogram was normalized to sum to 1 by dividing by the number of spikes. This procedure is consistent with a Poisson assumption of independent spikes (Sharpee, 2007).

The optimization algorithm (Sharpee et al., 2004, 2006) consisted of a series of 1D line optimizations along the gradient of information. During each line optimization, points that led to decreases of information were occasionally accepted with probability $\exp(-\Delta I/T)$, where ΔI is the decrease in information associated with acceptance of the new estimates of relevant dimensions, and parameter T – effective temperature – controls the probability of accepting decreases in information of large magnitude. Dimensions that led to an increase in information were always accepted. The optimization procedure started with the value of effective temperature $T = 1$. The effective temperature decreased by a factor of 0.95 after each line maximization until it reached the value of 10^{-5}. After that temperature increased by a factor of 100, and the iteration continued. The maximum number of line maximizations was 1000. Performance of the current dimension was evaluated on the test set after every line maximization. Dimensions with the best performance on the test set were used as the MIDs. The search for the first MID was initialized as the spike-triggered average.

After the first MID was computed, we initialized the second dimension as a random segment of the stimulus, and optimized a pair of dimensions to capture the maximal amount of

information about the arrival times of the single spikes in this case. The corresponding optimization function is given by:

$$I(\vec{v}_1, \vec{v}_2) = \int dx_1 \int dx_2 P_{\vec{v}_1,\vec{v}_2}(x_1, x_2|\text{spike}) \log_2 \frac{P_{\vec{v}_1,\vec{v}_2}(x_1,x_2|\text{spike})}{P_{\vec{v}_1,\vec{v}_2}(x_1,x_2)}, \quad (16)$$

where x_1 and x_2 represent stimulus components along dimensions $\vec{v}_1$ and $\vec{v}_2$, respectively. The probability $P_{\vec{v}_1, \vec{v}_2}(x_1, x_2)$ represents the probability distribution of stimulus components along dimension $\vec{v}_1$ and $\vec{v}_2$, and $P_{\vec{v}_1\vec{v}_2}(x_1, x_2|\text{spike})$ is the analogous probability distribution computed by taking only stimulus segments that lead to a spike. Dimensions $\vec{v}_1$ and $\vec{v}_2$ that at the end maximize Eq. (16) correspond to MID1 and MID2. Following optimization of the second dimension, the third dimension was added to the model and optimized using the three-dimensional probability distributions $P_{\vec{v}_1\vec{v}_2\vec{v}_3}(x_1, x_2, x_3)$ and $P_{\vec{v}_1\vec{v}_2\vec{v}_3}(x_1, x_2, x_3|\text{spike})$

For position-invariant models, the first dimension was estimated by maximizing the KL divergence in Eq. (9) between probability distribution $P^{\max}_{\vec{v}}(x)$ and $P^{\max}_{\vec{v}}(x|\text{spike})$ computed according to Eqs. (7, 8) with respect to maximal projections across patches of each image. The optimization used the same algorithm as described above for the position-specific case, with the only modification that patch locations yielding maximal projections onto the current estimate of the template were updated after each line optimization. Using locations that provided the greatest match to the first template $\vec{v}_1$, we then analyze the stimulus/response pairs to estimate the additional templates.

The estimates of relevant templates for each type of a model (with or without position-invariance) were obtained from the unrepeated data set. In each case, we obtained four jackknife estimates by leaving out a different consecutive 1/4 of the unrepeated data set as a validation data set and using the remaining 3/4 of the unrepeated data set as a training data set. The results of optimization that gave the best performance on the validation data set were then averaged across the four jackknife estimates to produce the estimated templates.

5.3 Quantifying predictive power of models

Information explained—To evaluate and compare performance of different kinds of models we then used a separate repeated data set (see above) to compute the mutual information accounted for by a given type of model. The mutual information was computed in the same way as during the optimization process (see preceding subsection 5.2), but using the repeated data set instead of the unrepeated data set that was used to find the relevant features. The advantage of using information as a measure of predictive power is that it characterizes how well a given set of features can account for spike times with a flexible nonlinear gain function. We note that the information values are however dependent on the number of bins. Here, we used seven bins to discretize probability distributions along each of the relevant dimensions. The dependence on the number of bins is typically largely independent of the features themselves, so that models evaluated using the same number of bins can be directly compared to each other. However, this dependence on binning makes it difficult to compare models with different number of relevant features.

Information per spike—The values for the single-spike information captured by different types of models were then compared to the overall information carried by the arrival times of single spikes, I_{spike}. Information I_{spike} about the stimulus carried by the arrival times of single spikes can then be computed using the average firing rate $r(t)$ as (Brenner et al., 2000):

$$I_{\text{spike}} = \frac{1}{T}\int dt \frac{r(t)}{\bar{r}} \log_2 \frac{r(t)}{\bar{r}}, \tag{17}$$

where $\bar{r}$ is the average stimulus evoked firing rate. This equation corresponds to Eq. (5) following the substitution of averaging over time with averaging over the stimulus probability distribution. This information measure makes no assumptions about the number of relevant stimulus dimensions nor about the shape of the nonlinear gain function describing the dependence of spike probability on the relevant stimulus components. Therefore, it can be used to quantify the performance of any model of a reduced dimensionality, such as models with and without position invariance.

Both the overall amount of information and the information accounted by different estimated models contain a positive bias, which decreases as more data are collected (Brenner et al., 2000; Strong et al., 1998; Treves and Panzeri, 1995). To correct for this bias, we computed information values based on different fractions of the repeats (80–100%), and then used linear extrapolation to find values predicted for infinite number of repetitions. This procedure was used to correct for bias in all information values (I_{spike} and information along one or more dimensions $\vec{v}$). The amount of correction varied between 3% and 15% depending on a neuron and type of model.

Maximal explained variance—Similar to information, one can also compute the maximal amount of variance that a given set of features can account for the observed responses with a flexible nonlinear gain function (Sharpee, 2007), comparing this to the overall variance in the firing rate. The latter quantity is given by

$$F_{\text{spike}} = \frac{1}{T}\int dt \left(\frac{r(t)}{\bar{r}}\right)^2 - 1. \tag{18}$$

It provides the maximal bound on the amount of variance that can be accounted for by any model. The variance accounted for by a model with multiple dimensions $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ can be computed using the following equation (Sharpee, 2007):

$$F(\vec{v}_1, \vec{v}_2, \vec{v}_3) = \int dx_1 \int dx_2 \int dx_3 \frac{[P_{\vec{v}_1,\vec{v}_2,\vec{v}_3}(x_1,x_2,x_3|\text{spike})]^2}{P_{\vec{v}_1,\vec{v}_2,\vec{v}_3}(x_1,x_2,x_3)} - 1, \tag{19}$$

Similar to the information per spike and information explained, the values for the variance in the firing rate Eq. (18) and variance accounted for by the model Eq. (19) contain a positive bias (Machens et al., 2004; Sahani and Linden, 2003). To correct for this bias, we used the same procedures as described above in the case of information values. To refer to this quantity as “maximal explained variance” to emphasize the fact that it is based on an unconstrained nonlinear gain function.

Correlation coefficients—To characterize predictive power of both the estimated features and nonlinear gain functions, we also computed correlation coefficients between predicted and measured firing rates. Here, we use both the filter and the nonlinear gain functions derived from the unrepeated data to predict the neuron’s firing rate for the repeated stimulus data set. The nonlinear gain function was estimated in a binless manner using Gaussian kernel density:

$$f(\vec{x}) = \frac{\sum_j r_j e^{-(\vec{x}_j - \vec{x})^2/(2\nu^2)}}{\sum_j e^{-(\vec{x}_j - \vec{x})^2/(2\nu^2)}},$$

where $\vec{x}$ describes projections along the relevant dimensions measured in their standard deviations, index j enumerates stimuli in the training data set, r_j is the measured spike rate for training stimulus j with projections onto relevant dimensions $\vec{x}_j$. The width ν of the Gaussian kernel was 0.1. We note that correlation coefficients are also linearly related to percent explained variance by the full model (features and nonlinear gain function estimated from the training data set and applied to test data set) up to the rescaling in the mean evoked firing rate. The mean evoked firing rate could be different between the training and test data sets, and previous studies have sought to compensate for this effect when evaluating the predictive power (Fairhall et al., 2006). Correlation coefficients were also extrapolated to infinite data set limit using the same procedure as described above for information and maximal variance explained.

Acknowledgments

We thank Adrian Wanner for helpful suggestions on improving the optimization procedure, and Jeffrey Fitzgerald, James Jeanne, and Anirvan Nandy for comments on the manuscript. This research was supported by grants R01EY019493 and K25MH068904 from the National Institutes of Health, grant 0712852 from the National Science Foundation, Alfred P. Sloan Research Fellowship, Searle Funds, the McKnight Scholarship, the Ray Thomas Edwards Career Development Award in Biomedical Sciences, and the W. M. Keck Foundation Research Excellence Award. Computing resources were provided by the National Science Foundation through TeraGrid resources provided by supercomputer resources at the San Diego Supercomputer Center, Argonne National Laboratory, University of Illinois National Center for Supercomputing Applications, and Texas Advanced Computing Center. Additional resources were provided by the Center for Theoretical Biological Physics (NSF PHY-0822283).

References

Adelman TL, Bialek W, Olberg RM. The information content of receptive fields. Neuron. 2003; 40:823–833. [PubMed: 14622585]

Atencio CA, Sharpee TO, Schreiner CE. Cooperative nonlinearities in auditory cortical neurons. Neuron. 2008; 58:956–966. [PubMed: 18579084]

Atencio CA, Sharpee TO, Schreiner CE. Hierarchical computation in the canonical auditory cortical circuit. PNAS. 2009; 106:21894–21899. [PubMed: 19918079]

Baddeley R, Abbott LF, Booth MCA, Sengpiel F, Freeman T, Wakeman EA, Rolls ET. Responses of neurons in primary and inferios temporal visual cortices to natural scenes. Proc R Soc Lond B. 1997; 264:1775–1783.

Bialek, W.; de Ruyter van Steveninck, RR. Features and dimensions: Motion estimation in fly vision. 2005. q-bio/0505003

Bishop, CM. Neural networks for pattern recognition. Oxford University Press; New York: 2004.

Boussaoud D, Desimone R, Ungerleider L. Visual topography of area teo in the macaque. J Comp Neurol. 1991; 306:554–575. [PubMed: 1712794]

Brenner N, Strong SP, Koberle R, Bialek W, de Ruyter van Steveninck RR. Synergy in a neural code. Neural Computation. 2000; 12:1531–1552. See also physics/9902067. [PubMed: 10935917]

Burak Y, Rokni U, Meister M, Sompolinsky H. Bayesian model of dynamic image stabilization in the visual system. PNAS. 2010; 107:19525–19530. [PubMed: 20937893]

Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T. A model of V4 shape selectivity and invariance. J Neurophysiol. 2007; 98:1733–1750. [PubMed: 17596412]

Chen X, Han F, Poo M, Dan Y. Excitatory and supressive receptive field subunits in awak monkey primary visual cortex (V1). PNAS. 2007; 104:19120–19125. [PubMed: 18006658]

Chichilnisky EJ. A simple white noise analysis of neuronal light responses. Network: Comput Neural Syst. 2001; 12:199–213.

Connor CE, Brincat SL, Pasupathy A. Transformation of shape information in the ventral pathway. Curr Opin Neurobiol. 2007; 17:140–147. [PubMed: 17369035]

de Boer E, Kuyper P. Triggered correlation. IEEE Trans Biomed Eng BME-. 1968; 15:169–179.

de Ruyter van Steveninck RR, Bialek W. Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. Proc R Soc Lond B. 1988; 265:259–265.

DeAngelis GC, Ohzawa I, Freeman RD. Spatiotemporal organization of simple- cell receptive fields in the cat's striate cortex. II. linearity of temporal and spatial summation. J Neurophysiol. 1993; 69:1118–1135. [PubMed: 8492152]

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc. 1977; 39:1–38.

Desimone R, Albright TD, Gross CG, Bruce C. Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci. 1984; 4:2051–2062. [PubMed: 6470767]

Desimone R, Schein SJ. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. J Neurophysiology. 1987; 57:835–868.

DiCarlo JJ, Maunsell JH. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. J Neurophysiol. 2003; 89:3264–3278. [PubMed: 12783959]

Dimitrov AG, Sheiko MA, Baker J, Yen SC. Spatial and temporal jitter distort estimated functional properties of visual sensory neurons. J Computational Neuroscience. 2009; 27:309–319.

Dong DW, Atick JJ. Statistics of natural time-varying images. Network: Comput Neural Syst. 1995; 6:345–358.

Edelman, S. Representation and recognition in vision. MIT Press; 1999.

Efron, B.; Tibshirani, RJ. An Introduction to the bootstrap. Chapman and Hall; 1998.

Fairhall AL, Burlingame CA, Narasimhan R, Harris RA, Puchalla JL, Berry M II. Selectivity for multiple stimulus features in retinal ganglion cells. J Neurophysiol. 2006; 96:2724–2738. [PubMed: 16914609]

Felsen G, Touryan J, Han F, Dan Y. Cortical sensitivity to visual features in natural scenes. PLoS Biol. 2005; e342:1819–1828.

Field DJ. Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A. 1987; 4:2379–2394. [PubMed: 3430225]

Field GD, Gauthier JL, Sher A, Greschner M, Machado T, Shlens J, Cunning DE, Mathieson K, Dabrowski W, Paninski L, Litke AM, Chichilnisky E. Functional connectivity in the retina at the resolution of photoreceptors. Nature. 2010; 467:673–677. [PubMed: 20930838]

Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980; 36:193–202. [PubMed: 7370364]

Gallant JL, Braun J, Van Essen DC. Selectivity for polar, hyperbolic, and cartesion gratings in macaque visual cortex. Science. 1993; 259:100–103. [PubMed: 8418487]

Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC. Neural responsese to polar, hyperbolic, and cartesion gratings in area V4 of the macaque monkey. J Neurophysiol. 1996; 76:2718–2739. [PubMed: 8899641]

Gawne TJ, Martin JM. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. J Neurophysiol. 2002; 88:1128–1135. [PubMed: 12205134]

Gill P, Zhang J, Woolley SM, Fremouw T, Theunissen FE. Sound representation methods for spectrotemporal receptive field estimation. J Comput Neurosci. 2006; 21:5–20. [PubMed: 16633939]

Gross CG, Bender DB, Rocha-Miranda CE. Visual receptive fields of neurons in the inferotemporal cortex of the monkey. Science. 1969; 166:1303–1307. [PubMed: 4982685]

Gross CG, Rocha-Miranda CE, Bender DB. Visual properties of neurons in inferotemporal cortex of the macaque. J Neurophysiol. 1972; 35:96–111. [PubMed: 4621506]

Heeger DJ. Normalization of cell responses in cat visual cortex. Vis Neurosci. 1992; 9:181– 197. [PubMed: 1504027]

Hegde J, Van Essen DC. Selectivity for complex shapes in primate visual area V2. J Neurosci. 2000; 20:1–6. [PubMed: 10627575]

Hegde J, Van Essen DC. A comparative study of shape representation in macaque visual areas V2 and V4. Cereb Cortex. 2007; 17:1100–1116. [PubMed: 16785255]

Horwitz GD, Chichilnisky EJ, Albright TD. Cone inputs to simple and complex cells in v1 of awake macaque. J Neurophysiol. 2007; 97:3070–3081. [PubMed: 17303812]

Ito M, Tamura H, Fujita I, Tanaka K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. J Neurophysiol. 1995; 73:218–226. [PubMed: 7714567]

Kobatake E, Tanaka K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol. 1994; 71:856–867. [PubMed: 8201425]

Kouh M, Sharpee TO. Estimating linear-nonlinear models using Renyi divergences. Network: Comput Neural Syst. 2009; 20:49–68.

Leuschow A, Miller EK, Desimone R. Inferior temporal mechanisms for invariant object recognition. Cereb Cortex. 1994; 5:523–531.

Liu YS, Stevens CF, Sharpee TO. Predictable irregularities in retinal receptive fields. PNAS. 2009; 106:16499–16504. [PubMed: 19805327]

Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. Curr Biol. 1995; 5:552–563. [PubMed: 7583105]

Machens CK, Wehr MS, Zador AM. Linearity of cortical receptive fields measured with natural sounds. J Neurosci. 2004; 24:1089–1100. [PubMed: 14762127]

Mackay DJC. Information-based objective functions for active data selection. Neural Comput. 1992; 4:590–604.

Mechler F, Ringach DL. On the classification of simple and complex cells. Vision Research. 2002; 42:1017–1033. [PubMed: 11934453]

Meister M, Berry M II. The neural code of the retina. Neuron. 1999; 22:435–450. [PubMed: 10197525]

Missal M, Vogels R, Li C, Orban GA. Shape interactions in macaque inferior temporal neurons. J Neurophysiol. 1999; 82:131–142. [PubMed: 10400942]

Nishimoto S, Ishida T, Ohzawa I. Receptive field properties of neurons in the early visual cortex revealed by local spectral reverse correlation. J Neurosci. 2006; 26:3269–3280. [PubMed: 16554477]

Op de Beeck H, Vogels R. Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol. 2000; 426:505–518. [PubMed: 11027395]

Paninski L. Convergence properties of three spike-triggered average techniques. Network: Comput Neural Syst. 2003; 14:437–464.

Pasupathy A, Connor CE. Responses to contour features in macaque area V4. J Neurophysiol. 1999; 82:2490–2502. [PubMed: 10561421]

Pasupathy A, Connor CE. Shape representation in area V4: position-specific tuning for boundary conformation. J Neurophysiol. 2001; 86:2505–2519. [PubMed: 11698538]

Pasupathy A, Connor CE. Population coding of shape in area V4. Nat Neurosci. 2002; 5:1332–1338. [PubMed: 12426571]

Pelli DG. Uncertainty explains many aspects of visual cortex detection and discrimination. J Opt Soc Am A. 1985; 2:1508–1532. [PubMed: 4045584]

Pillow JW, Simoncelli EP. Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis. Journal of Vision. 2006; 6:414–428. [PubMed: 16889478]

Pollen DA, Przybyszewski AW, Rubin MA, Foote W. Spatial receptive field organization of macaque V4 neurons. Cereb Cortex. 2002; 12:601–616. [PubMed: 12003860]

Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press; Cambridge: 1992.

Rainer G, Augath M, Trinath T, Logothetis NK. Nonmonotonic noise tuning of bold fMRI signal to natural images in the visual cortex of the anesthetized monkey. Curr Biol. 2001; 11:846–854. [PubMed: 11516645]

Rapela J, Felsen G, Touryan J, Mendel JM, Grzywacz NM. ePPR: a new strategy for the characterization of sensory cells from input/output data. Network: Computation in Neural Systems. 2010; 21:35–90.

Rapela J, Mendel JM, Grzywacz NM. Estimating nonlinear receptive fields from natural images. Journal of Vision. 2006; 16:441–474. [PubMed: 16889480]

Richmond BJ, Wurtz RH, Sato T. Visual responses of inferior temporal neurons in awake rhesus monkey. J Neurophysiol. 1983; 50:1415–1432. [PubMed: 6663335]

Rieke, F.; Warland, D.; de Ruyter van Steveninck, RR.; Bialek, W. Spikes: Exploring the neural code. MIT Press; Cambridge: 1997.

Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999; 2:1019–1025. [PubMed: 10526343]

Ringach DL, Hawken MJ, Shapley R. Receptive field structure of neurons in monkey visual cortex revealed by stimulation with natural image sequences. Journal of Vision. 2002; 2:12–24. [PubMed: 12678594]

Ringach DL, Sapiro G, Shapley R. A subspace reverse-correlation technique for the study of visual neurons. Vision Res. 1997; 37:2455–2464. [PubMed: 9381680]

Rolls ET. Functions of the primate temporal lobe cortical visual areas in invariant-object recongnition. Neuron. 2000; 27:205–218. [PubMed: 10985342]

Rowekamp RJ, Sharpee TO. Analyzing multicomponent receptive fields from neural responses to natural stimuli. Network: Comput Neural Syst. 2011; 22:45–73.

Ruderman DL, Bialek W. Statistics of natural images: Scaling in the woods. Phys Rev Let. 1994; 73:814–817. [PubMed: 10057546]

Rust NC, DiCarlo JJ. Selectivity and tolerance (”invariance”) both increase as visual information propagates from cortical area V4 to IT. J Neuroscience. 2010; 30:12978–12995.

Rust NC, Schwartz O, Movshon JA, Simoncelli EP. Spatiotemporal elements of macaque V1 receptive fields. Neuron. 2005; 46:945–956. [PubMed: 15953422]

Sahani, M.; Linden, JF. How linear are auditory cortical responses?. In: Becker, S.; ST; Obermayer, K., editors. Advances in Neural Information Processing Systems. Vol. 15. MIT Press; Cambridge, MA: 2003. p. 109-116.

Saleem KS, Tanaka K, Rockland KS. Specific and columnar projection from area teo to te in the macaque inferotemporal cortex. Cereb Cortex. 1993; 3:454–464. [PubMed: 8260813]

Sary G, Vogels R, Orban GA. Cue-invariant shape selectivity of macaque inferior temporal neurons. Science. 1993; 260:995–997. [PubMed: 8493538]

Schwartz EL, Desimone R, Albright TD, Gross CG. Shape recognition and inferior temporal neurons. Proc natl Acad Sci U S A. 1983; 80:5776–5778. [PubMed: 6577453]

Schwartz, O.; Chichilnisky, EJ.; Simoncelli, E. Characterizing neural gain control using spike-triggered covariance. In: Dietterich, TG.; Becker, S.; Ghahramani, Z., editors. Advances in Neural Information Processing. Vol. 14. 2002.

Schwartz O, Pillow J, Rust N, Simoncelli EP. Spike-triggered neural characterization. Journal of Vision. 2006; 6:484–507. [PubMed: 16889482]

Schwartz O, Simoncelli EP. Natural signal statistics and sensory gain control. Nat Neurosci. 2001; 4:819–825. [PubMed: 11477428]

Sharpee T, Rust N, Bialek W. Analyzing neural responses to natural signals: Maximally informatiove dimensions. Neural Computation. 2004; 16:223–250. [PubMed: 15006095] See also physics/0212110, and a preliminary account in Becker S, Thrun S, Obermayer K. Advances in Neural Information Processing. 15:261–268.MIT PressCambridge2003;

Sharpee TO. Comparison of information and variance maximization strategies for characterizing neural feature selectivity. Statistics in Medicine. 2007; 26:4009–40031. [PubMed: 17597484]

Sharpee TO, Miller KD, Stryker MP. On the importance of static nonlinearity in estimating spatiotemporal neural filters with natural stimuli. J Neurophysiol. 2008; 99:24962–2509.

Sharpee TO, Sugihara H, Kurgansky AV, Rebrik SP, Stryker MP, Miller KD. Adaptive filtering enhances information transmission in visual cortex. Nature. 2006; 439:936– 942. [PubMed: 16495990]

Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annu Rev Neurosci. 2001; 24:1193–1216. [PubMed: 11520932]

Sincich LC, Horton JC, Sharpee TO. Preserving information in neural transmission. J Neurosci. 2009; 29:6207–6216. [PubMed: 19439598]

Skottun BC, De Valois RL, Grosof DH, Movshon JA, Albrecht DG, Bonds AB. Classifying simple and complex cells on the basis of response modulation. Vision Res. 1991; 31:1079–1086. [PubMed: 1909826]

Soo FS, Schwartz GW, Sadeghi K, Berry M II. Fine spatial information represented in a population of retinal ganglion cells. J Neurosci. 2011; 31:2145–2155. [PubMed: 21307251]

Soodak RE, Shapley RM, Kaplan E. Fine structure of receptive field centers of X and Y cells of the cat. Visual Neuroscience. 1991; 6:621–628. [PubMed: 1883766]

Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. Entropy and information in neural spike trains. Phys Rev Lett. 1998; 80:197–200.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network. 2001; 3:289–316. [PubMed: 11563531]

Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci. 2000; 20:2315–2331. [PubMed: 10704507]

Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature. 1996; 381:520–522. [PubMed: 8632824]

Tjan BS, Nandy AS. Classification images with uncertainty. Journal of Vision. 2006; 6:387–413. [PubMed: 16889477]

Touryan J, Felsen G, Dan Y. Spatial structure of complex cell receptive fields measured with natural images. Neuron. 2005; 45:781–791. [PubMed: 15748852]

Touryan J, Lau B, Dan Y. Isolation of relevant visual features from random stimuli for cortical complex cells. J Neurosci. 2002; 22:10811–10818. [PubMed: 12486174]

Tovee MJ, Rolls ET, Azzopardi P. Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. J Neurophysiol. 1994; 72:1049–1060. [PubMed: 7807195]

Treves A, Panzeri S. The upward bias in measures of information derived from limited data samples. Neural Comp. 1995; 7:399–407.

Ullman, S. High level vision. MIT Press; 1996.

van Hateren JH, Ruderman DL. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proc R Soc lond B Biol Sci. 1998; 265:2315–2320.

Victor JD, Shapley R. A method of nonlinear analysis in the frequency domain. Biophys J. 1980; 29:456–483.

Wang G, Tanaka K, Tanifuji M. Optical imaging of functional organization in the monkey inferotemporal cortex. Science. 1996; 272:1665–1668. [PubMed: 8658144]

Wang G, Tanifuji M, Tanaka K. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. Neurosci Res. 1998; 32:33–46. [PubMed: 9831250]

Woolley SMN, Gill PR, Theunissen FE. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat Neurosci. 2005; 8:1371–1379. [PubMed: 16136039]

Woolley SMN, Gill PR, Theunissen FE. Stimulus-dependent auditory tuning results in synchronous population coding of vocalization in the songbird midbrain. J Neurosci. 2006; 26:2499–2512. [PubMed: 16510728]

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci. 2007; 27:12292–12307. [PubMed: 17989294]

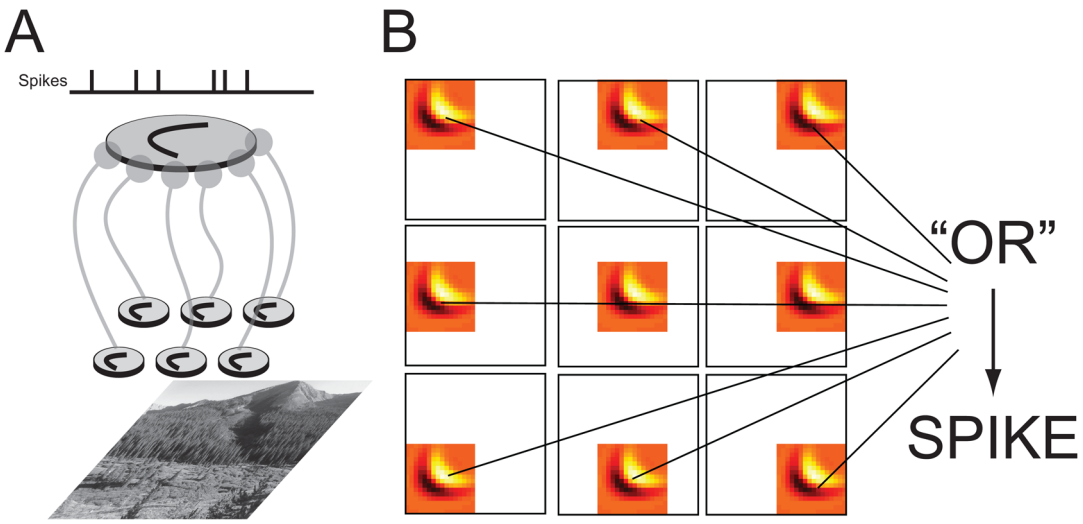

Figure 1.

(**A**) Model of neural response based on one translation invariant stimulus feature. The spike probability represents a logical OR combination of responses from hidden, position-specific units that are selective for the same stimulus feature centered at different retinotopic coordinates. (**B**) An example of a discrete 3 × 3 grid approximation that can be used to model invariance of neural responses to image translation. The shaded square denotes the spatial extent of the preferred image feature; nine possible ways of centering the preferred template within the overall stimulus are shown.

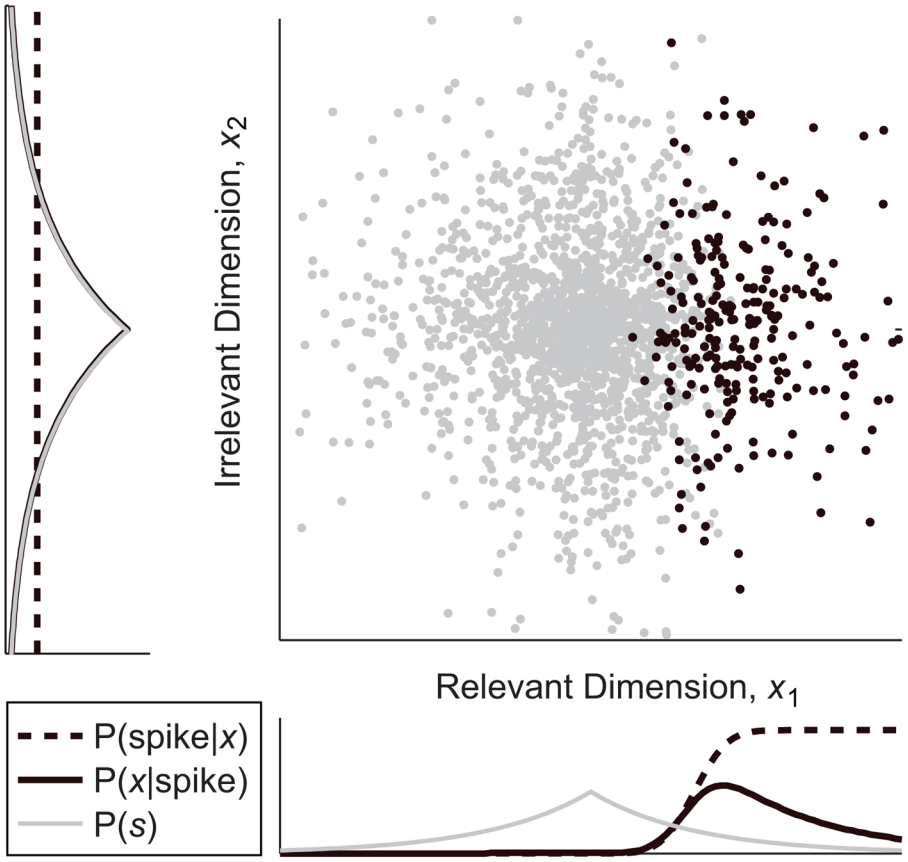

Figure 2. Statistical description of neural responses along relevant and irrelevant dimensions

In the framework of the position-specific model, some images elicit spikes (black) and others do not (gray). Here each of the images $\vec{s}$ is represented as a point in a two-dimensional space, although it is a point in a high d-dimensional space (each axis may correspond to the luminance of a pixel). Because the vertical dimension (x_2) does not affect the spike probability, the probability distribution of stimuli along that dimension $P(x_2)$ is similar to the distribution of stimuli given a spike $P(x_2|\text{spike})$. On the other hand, the horizontal dimension x_1 can account for the spiking behavior, because the spikes are observed whenever the stimulus component x_1 exceeds a certain value.

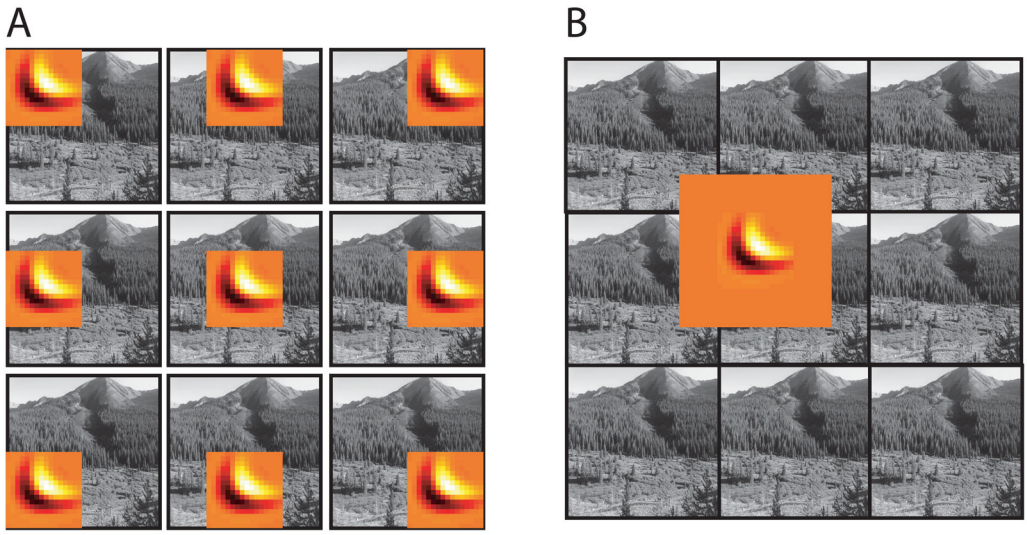

Figure 3. Two approaches for characterizing translation invariant feature selectivity

(**A**) In the direct approach, we seek a template whose spatial extent is smaller than the overall stimulus that covers the response region of a neuron. The spike probability is examined by translating the candidate template to different locations of the translation grid (shown here for a 3 × 3 grid). (**B**) In the Fourier approach, in order to account for the translation invariance, the template is shifted to different locations of the translation grid assuming periodic boundary conditions. Compared to the direct approach, the Fourier approach can typically handle finer translation grids (due to memory restrictions in the direct approach), but it yields coarser estimates of the template because of the need to leave larger margins when using periodic boundary conditions.

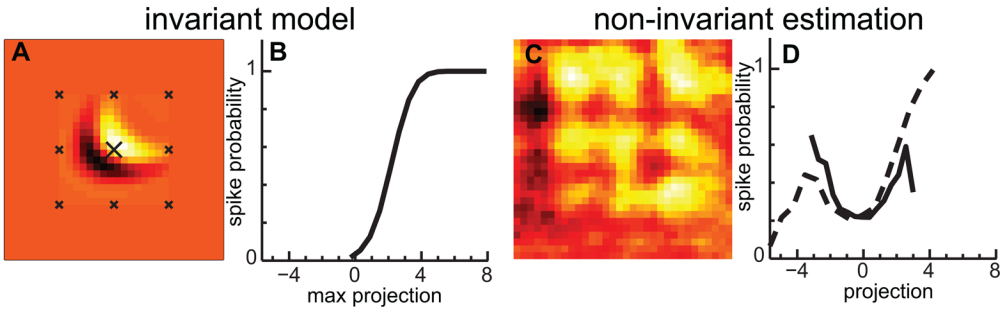

Figure 4. Feature selectivity of translation invariant neurons cannot be characterized without taking this invariance into account

(**A**) The relevant feature of a model neuron with translation invariant responses. The centers of the 3 × 3 translation grid are marked with crosses. (**B**) The nonlinear gain function of the translation invariant model cell evaluated at the location producing a maximal projection with the model template ($\theta = 2.5$, $\sigma = 1.0$, stimulus repeated 20 times). (**C**) The estimated template without taking into account translation invariance. (**D**) Comparison of the nonlinear gain functions with respect to the estimated filter (solid line) with the nonlinear gain function with respect to the model template at the central location of the translation grid (dashed line). Both functions are computed without translation invariance. The observed increase in the nonlinear gain function for negative projection values is due to the overlap between the templates centered at neighboring positions of the translation grid.

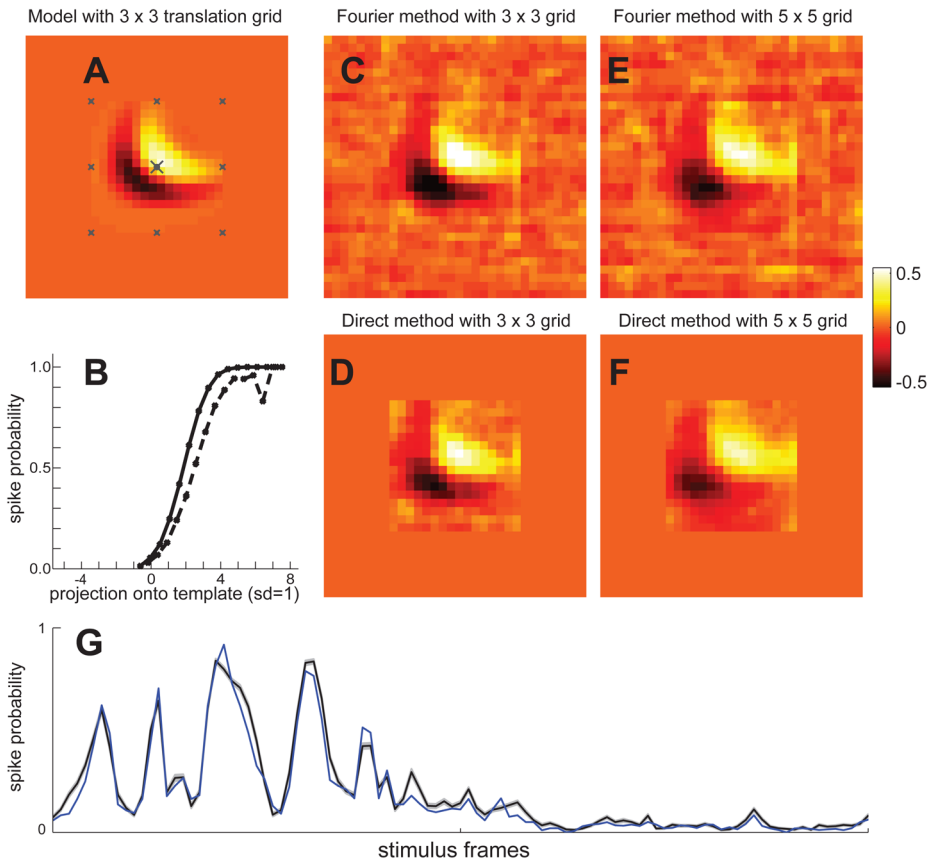

Figure 5. Estimation of translation invariant models

(**A**) The relevant template of the model neuron overlayed with the 3×3 translation grid (whose points are marked by crosses). (**B**) Comparison between the nonlinear gain function of the model cell (solid line) and the translation invariant estimation (dashed line). In contrast to the case of estimation without translation invariance, cf. Fig. 4, this estimation does reproduce the correct, sigmoidal form of the nonlinear gain function. (**C**) The Fourier method estimation using the 3×3 translation grid (same grid as in the model) yields a dot product of $c = 0.897 \pm 0.008$ and a fraction of information explained $I_{expl} = 0.963 \pm 0.006$ (1 is the maximum). (**D**) Analogous estimation using the direct method yielded $c = 0.899 \pm 0.011$ and $I_{expl} = 0.969 \pm 0.008$. Assuming a mismatched 5 ×5 translation grid (compared to the model) still leads to reasonable estimation results using either the Fourier method (**E**), $c = 0.790 \pm 0.004$ and $I_{expl} = 0.826 \pm 0.002$, or the direct method (**F**), $c = 0.78 \pm 0.02$, $I_{expl} = 0.826 \pm 0.003$. (**G**) Comparison of model spike probability (black line, gray area shows standard errors of the mean) and the predicted spike probability (blue) using the template and model from panel **D**. Predictions were made for a novel set of frames not used in estimating the model.

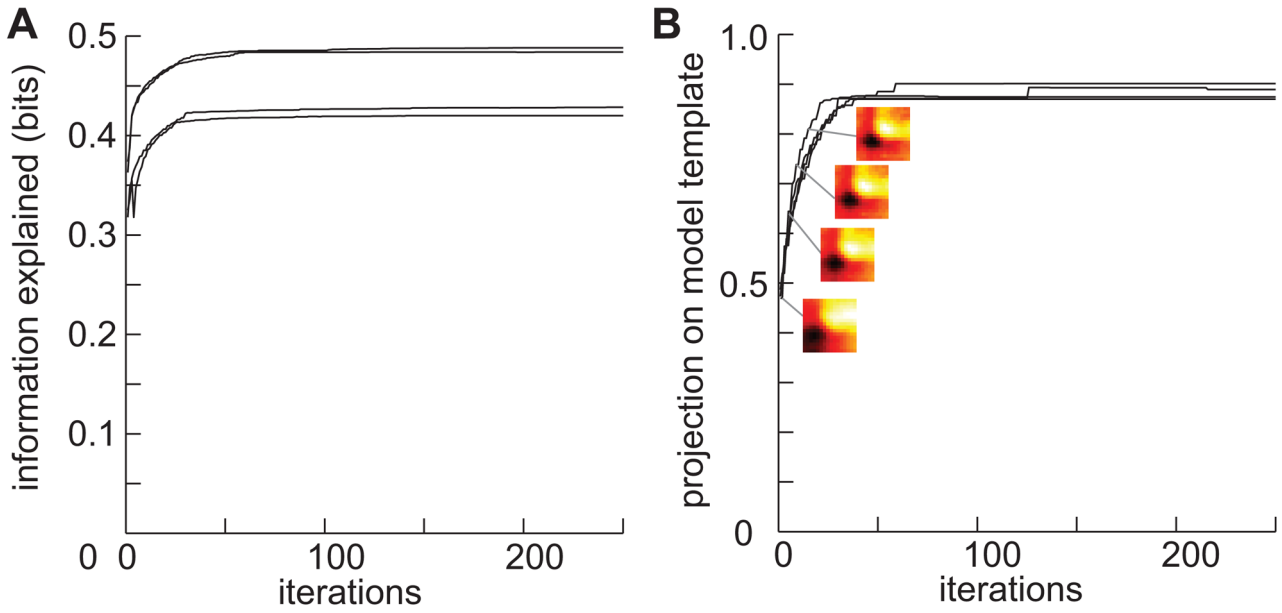

Figure 6. Example of algorithm convergence

(**A**) Convergence in terms of information explained on the test data set by the candidate template. (**B**) Convergence in terms of projection of the candidate template onto the model template. In both panels, the four different lines correspond to four different jackknife analyses of the same model neuron. In the case of information, different final values are due to differences in the overall information per spike in a particular test data set. According to both parameters the algorithm converges in all cases within 100 iterations, less than $d = 256$ of the template space. Insets in (**B**) show the estimated templates after 1, 5, 10, and 15 line optimizations.

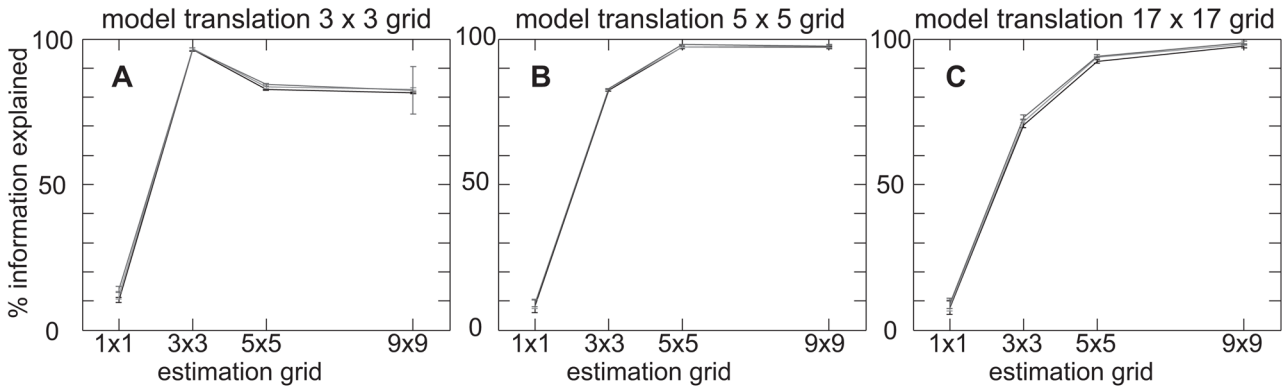

Figure 7. Recovering the coarseness of translation invariance

The percent of information explained is plotted as a function of the translation grid size assumed during estimation. (**A**) Model cells with a 3 × 3 translation grid, $\sigma = 1.0$, $\theta = 2.5$, 2.75, 3.0 analyzed from 20 repeats of the whole stimulus sequence (16, 384 frames). The best predictive power is obtained when the same grid is used during estimation. Significant t-tests are obtained for the difference between the peak value and the value for 1 × 1 grid and 5 × 5 grid ($p < 10^{-4}$, t-test). (**B**) Model cells with 5 × 5 translation grid and $\theta = 3.0$, 3.25, 3.75 (other parameters are the same as in (**A**)). The use of coarser translation grids results in significantly worse performance ($p < 10^{-4}$); however finer translation grid results in the same performance ($p = 0.16$). (**C**) Model cells with 17 × 17 translation grid, $\theta = 4.0$, 4.25, 4.50. This is the case of perfect translation invariance with the grid spacing of 1 pixel. We find that the performance of the estimation algorithm continues to improve from 5 × 5 to 9 × 9 grids ($p < 10^{-4}$). In all cases, therefore, the algorithm could disambiguate coarser translation grids from the true ones.

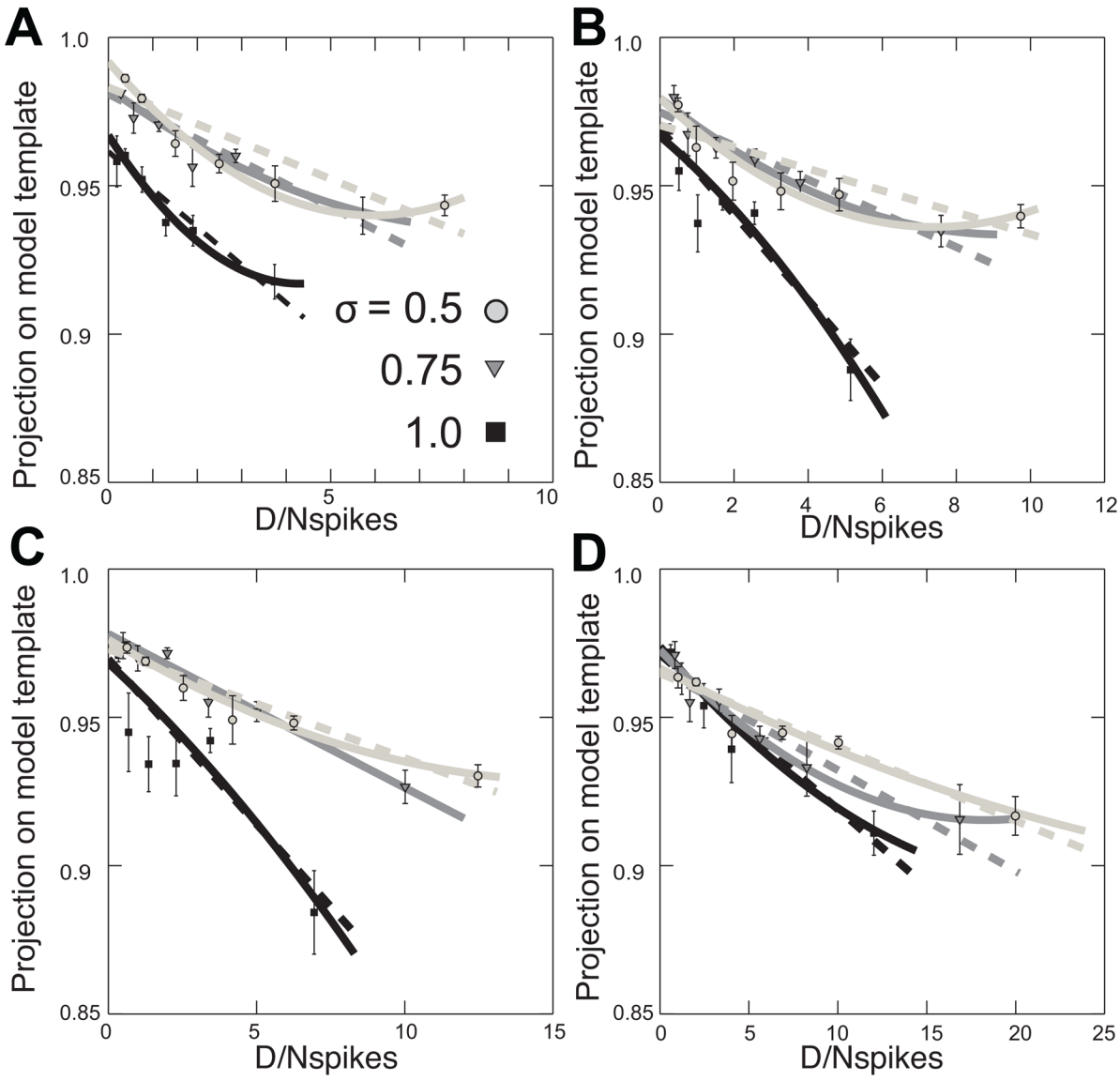

Figure 8. Projection between estimated and model dimensions as a function of the number of spikes

N_{spikes}. Improvement in performance with increasing number of spikes is shown for 12 model cells. All of the model cells had the same relevant template as in Fig. 5*A* and translation grid 3 × 3, but different noise levels and thresholds $\theta = 2.5$ (**A**), $\theta = 2.75$ (**B**)), $\theta = 3.0$ (**C**), and $\theta = 3.5$ (**D**). Within each panel, model cells have $\sigma = 0.5$ (light gray, ○), 0.75 (dark gray, ▽), and 1.0 (black, □). The solid and dashed lines represent results of quadratic and linear regressions. Stimulus dimensionality $D = 1024$, corresponding to frames with 32 × 32 pixels. Results were obtained using the Fourier approach. Good performance is obtained for all models cells even in the severely undersampled regime with $D > N_{spikes}$. As expected, the improvements with increasing the number of spikes are more pronounced for neurons with larger noise levels.

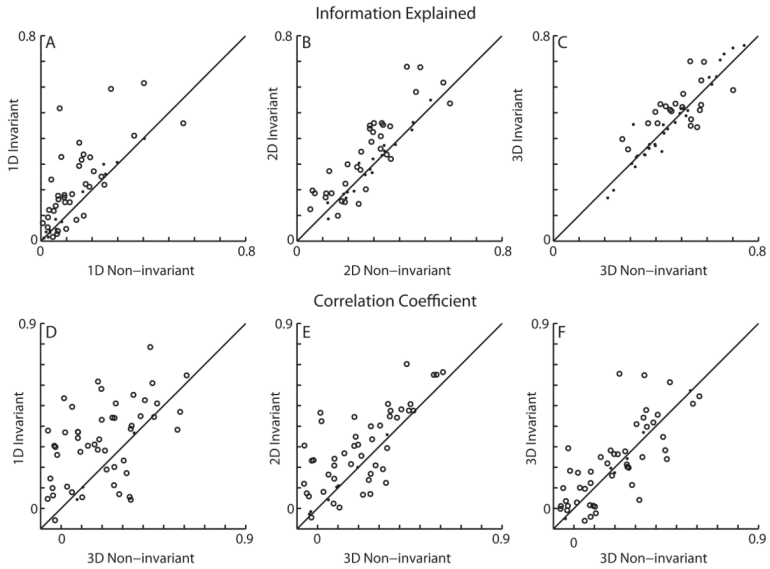

Figure 9. Population analysis of predictive power of position-specific and position-invariant models for V1 complex cells

Fraction of information explained by models with one (A), two (B) and three (C) features. Correlation coefficients between measured and predicted firing rates are compared for models with one (D), two (E), and three (F) translation invariant features with correlation coefficients obtained with the three-feature position-specific model. Models with the same number of features can be compared according percent information values (A–C), whereas models with different number of features can be compared according to correlation coefficients (D–F). Across the population, position-invariant models with one or two features outperformed their position-specific counterparts. Furthermore, significant improvements were observed for some of the neurons considered individually (points marked with empty circles, $P < 0.05$ t-test), where even the models with single translation invariant template outperformed the models with three position-specific features (D).

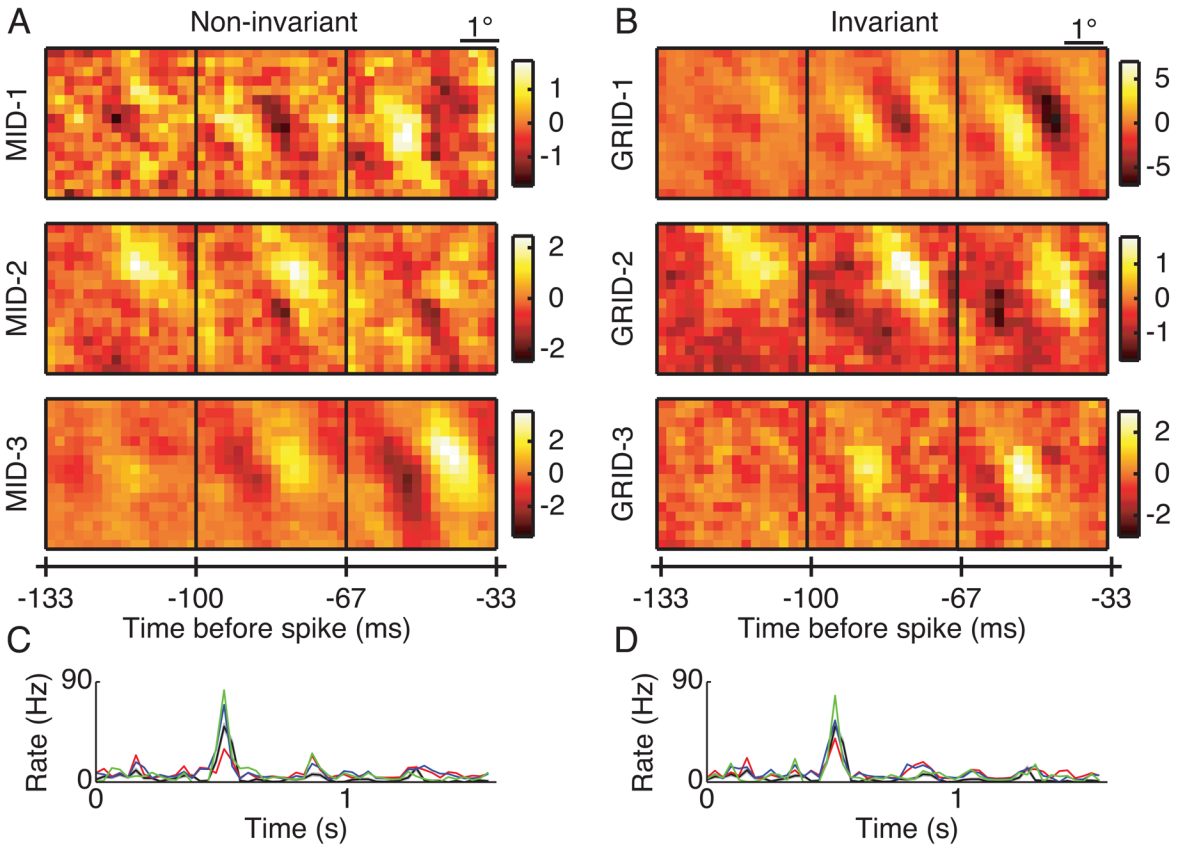

Figure 10. Example V1 complex cell that was better described with a position-invariant model

(**A**) Three relevant spatiotemporal features for a position-specific LN model are shown. Each feature is shown in a separate row and represents a spatiotemporal profile covering three time lags from −132 to −33 msec before the spike arrival time. Results are shown as averages over four jackknife estimates of each feature. The color scale denotes signal-to-noise ratio relative to the variance across the jackknife estimates, which was corrected for overlapping data in the jackknife estimates (Efron and Tibshirani, 1998). (**B**) Three relevant spatiotemporal templates of a position-invariant LN model, notations are as in (**A**). Firing rate predictions were made using these models for a novel, repeated data set. Predictions using the position-specific models (**C**) and position-invariant models (**D**) are shown using red, blue, and green lines for models based on one, two and three features, respectively. The measured firing rate (black line) is shown together with its standard error of the mean (gray shading), Neuron 883-2.

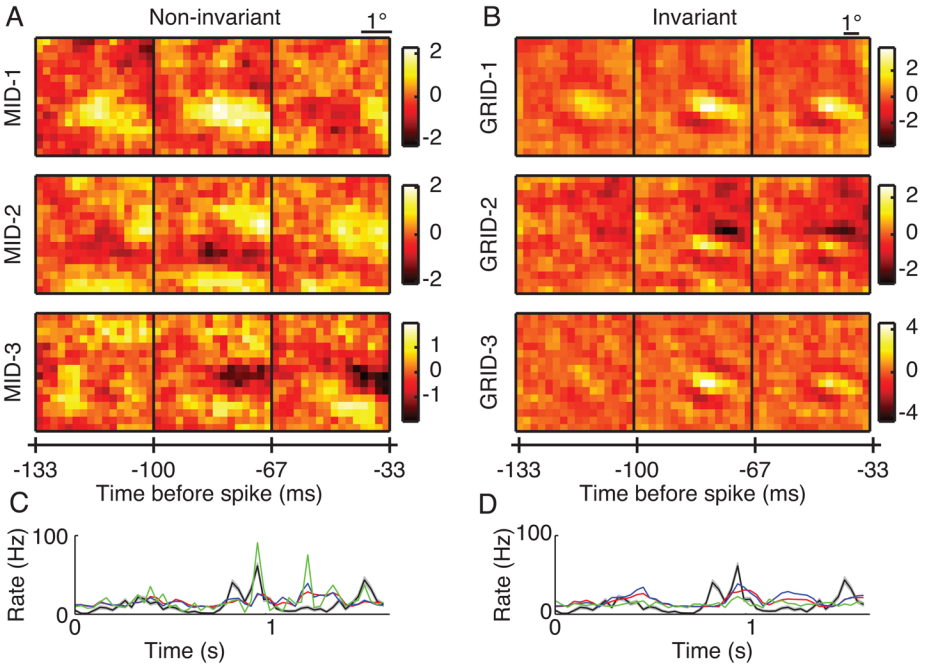

Figure 11. Example V1 complex cell that was better described with a position-specific model
Notations are as in Figure 10. Neuron 772-2.

Table 1

The combinations of parameters used to generate the model cells. The values of threshold θ and noise level σ are measured in units of the standard deviation of the stimulus projections onto the relevant template. A total of 45 different model cells were analyzed, each of which analyzed for six different numbers of repeats of the whole stimulus sequence. The stimulus length was 16, 384 frames. Given the frame size 32 × 32, and the template size of 16 × 16, the 17 × 17 translation grid corresponds to full translation invariance (all patches are considered). To maintain the average spike rate of the translation invariant model cell within a reasonable range, we had to adjust the spike thresholds θ for hidden units depending on the translation grid. Finally, we also explored how results of the estimation improved with an increasing number of spikes for a given model cell, by simulating several batches of responses to the same repeated stimulus sequences (see Sec. 3.4 below).

Spike threshold				
3×3	5×5	17×17	Noise level	Number of repeats
2.5	3.0	4.0	0.5	1
2.75	3.25	4.25	0.75	2
3.00	3.50	4.50	1.0	3
3.25	3.75	4.75		5
3.50	4.00	5.00		10
				20

Table 2

Measures of predictive power of position-specific and position-invariant models for an example neuron from Figure 10 that was best described by a position-invariant model. Models with one, two, and three features are denoted as 1D, 2D, and 3D, respectively. The means and standard deviations are reported.

		Information Fraction	Max Variance Fraction	Correlation coefficients
Position-specific	1D	0.159 ± 0.003	0.121 ± 0.011	0.367 ± 0.002
	2D	0.336 ± 0.005	0.31 ± 0.02	0.477 ± 0.003
	3D	0.525 ± 0.008	0.72 ± 0.04	0.468 ± 0.002
Position-invariant	1D	0.287 ± 0.005	0.197 ± 0.009	0.511 ± 0.003
	2D	0.453 ± 0.007	0.42 ± 0.02	0.476 ± 0.003
	3D	0.614 ± 0.011	0.81 ± 0.04	0.614 ± 0.002

Table 3

Measures of predictive power of position-specific and position-invariant models for an example neuron from Figure 11 that was best described by a position-specific model.

		Information Fraction	Max Variance Fraction	Correlation coefficients
Position-specific	1D	0.190 ± 0.006	0.114 ± 0.009	0.329 ± 0.002
	2D	0.350 ± 0.007	0.39 ± 0.03	0.423 ± 0.003
	3D	0.532 ± 0.013	0.79 ± 0.05	0.453 ± 0.002
Position-invariant	1D	0.211 ± 0.006	0.094 ± 0.006	0.445 ± 0.004
	2D	0.337 ± 0.008	0.204 ± 0.012	0.507 ± 0.003
	3D	0.518 ± 0.010	0.47 ± 0.03	0.240 ± 0.004