STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4009–4031
Published online 27 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2931

Comparison of information and variance maximization strategies for characterizing neural feature selectivity

Tatyana O. Sharpee*, †

Crick-Jacobs Center for Theoretical and Computational Biology and the Laboratory of Computational Neurobiology, The Salk Institute for Biological Studies, La Jolla, CA 92037, U.S.A.

SUMMARY

This paper compares several statistical methods for analyzing neural feature selectivity with natural stimuli. Despite the non-Gaussian character of correlations in natural stimuli, several relevant stimulus dimensions can be found by maximizing either information or, as is demonstrated here, variance. In the case of information, the relevance of each dimension is quantified by a Kullback–Leibler divergence between the full input probability distribution and that across inputs associated with positive neural responses, both projected onto that dimension. We demonstrate that least-square matching of the nonlinear prediction based on several dimensions relevant to the recorded spike trains yields an optimization scheme similar to information maximization. The relevant dimensions are found as those that capture the most variance in neural response. The variance along a stimulus dimension is given by a Rényi divergence of order 2 instead of the Kullback–Leibler divergence used for maximizing information. Statistical errors expected for the two schemes are shown to be similar through both analytical and numerical calculations. However, in the asymptotic limit of large spike numbers, maximizing information results in smaller errors than variance optimization. Numerical simulations for model cells with different noise levels show that this trend persists, and possibly increases, when the number of spikes decreases. This makes the problem of finding relevant dimensions one of the examples where information-theoretic approaches are no more data limited than the variance-based measures. Variance and information optimization also outperform methods based on the spike-triggered average for all numbers of spikes and neural noise levels. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: receptive fields; linear–nonlinear model; information theory; Rényi divergences; natural stimuli

*Correspondence to: Tatyana O. Sharpee, 10010 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.
†E-mail: sharpee@salk.edu

Contract/grant sponsor: NIMH; contract/grant number: 5K25MH068094

Received 17 April 2007
Accepted 17 April 2007

INTRODUCTION

The concept of neural feature selectivity is based on the observation that, even though input signals might be of high dimension, quite often only a small number of dimensions in the input space influence neural response. This model has been very successful in elucidating neural coding at various levels of visual [1–8], auditory [9–14], and motor [15, 16] processing. Formally, the assumption of the linear–nonlinear model specifies that the probability of a spike $P(\text{spike}|\mathbf{s})$ given a stimulus $\mathbf{s}$ depends only on stimulus projections $s_i = \hat{e}_i \cdot \mathbf{s}$ on a set of K relevant vectors $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K\}$ [1, 5, 9, 17–20]:

$$P(\text{spike}|\mathbf{s}) = P(\text{spike})g(s_1, s_2, \ldots, s_K) \tag{1}$$

where $P(\text{spike})$ is the average firing rate, and g is a potentially strongly nonlinear function of stimulus projections s_i— for example, a threshold or a sigmoid. For the sake of focus, we chose here, and in what follows, a single spike as the response of interest. All of the above arguments, as well as the optimization schemes described below, can be carried out with respect to particular patterns of spikes in time and/or across neurons [21].

The goal is then to find (i) the number K of relevant features; (ii) the features themselves; and (iii) then estimate the nonlinear gain function g. The approach is useful only if the relevant number of stimulus dimensions K will turn out to be much smaller than the overall dimensionality D of the stimulus space. In some sense, the most difficult part is to find the relevant dimensions. This is because, once they are known, the probability $P(\text{spike}|\mathbf{s})$ becomes a function of only a few parameters, and it becomes feasible to map this function experimentally, inverting the probability distributions according to Bayes' rule [1, 2, 5]:

$$g(s_1, s_2, \ldots, s_K) = \frac{P(s_1, s_2, \ldots, s_K|\text{spike})}{P(s_1, s_2, \ldots, s_K)} \tag{2}$$

Note that this does not mean that nonlinearity can be discarded when determining relevant dimensions. Only under special circumstances, such as with Gaussian inputs, can the correct features be found even if our first assumptions about the shape of nonlinearity g were wrong [5, 20, 22, 23].

All of the three tasks (finding the number of filters, the filters themselves, and the corresponding nonlinearity g) can be accomplished using the spike-triggered covariance method as long as inputs can be described by a multivariate Gaussian [1, 3–6, 20, 24]. If we assume, or are interested in, finding only one relevant dimension, then finding the relevant dimension and its corresponding nonlinearity can be accomplished by the classic reverse-correlation or spike-triggered average (STA) technique. The use of the STA technique with Gaussian inputs is possibly the single most important example, where the relevant dimension can be found by presuming that the system is fully linear and still be correct for a linear–nonlinear model (1) [9, 25, 26].

But what if input signals deviate strongly from a multivariate Gaussian distribution? This is the case for most of the signals derived from real-world stimuli, whether they are visual [27], auditory [28], or olfactory [18]. According to the method of maximally informative dimensions [22], the set of relevant dimensions can be found iteratively by maximizing Shannon information between neural response and stimuli projected onto a set of trial dimensions [22]. The relevance of each dimension is quantified by the Kullback–Leibler divergence between the probability distribution of all inputs projected onto the relevant dimension and its version computed for all inputs that elicited the neural response [22]. The use of other measures of divergence [29, 30] between these two probability distributions has also been proposed [23]. Here, we will show that optimizing one of them, a Rényi divergence [31] of order 2, corresponds to 'fitting' the (nonlinear) model (1) in the least-square sense to neural data.

Close parallels can be made between the strategies of optimizing information (*via* Kullback–Leibler divergence) and variance (*via* Rényi divergence). In either case, no *a priori* assumptions on the shape of nonlinearity function g are made. In particular, it does not need to be monotonic or invertible, which is an improvement over previous methods of fitting the linear–nonlinear models as defined by equation (1) to neural data [17, 32–34]. When maximizing information, the information carried by the arrival of single spikes [21] provides the maximal amount that can be accounted for by reconstruction with a fixed number of relevant filters. It can therefore be used to judge the quality of the reconstruction of model (1). A similar quantity can be derived for variance maximization scheme, cf. equation (16) below. Algorithmically, the numerical schemes of maximizing information and variance can be identical. For example, the gradient can be computed in both cases. However, we find that maximizing information allows for more accurate reconstructions of relevant dimensions than maximizing variance.

FINDING A SINGLE RELEVANT DIMENSION

Let us first tackle the case where the neuron under study generates spikes based on just one relevant dimension. Our goal is to find this relevant dimension without making any assumptions about the probability distribution of input signals, as long as these signals are diverse enough to span the space over which we will be looking for the relevant dimension. Without the Gaussian assumption, we cannot use any methods that are based on correlation functions. An advantage of those methods is that they can be reduced to linear algebra problems, which usually are efficient computationally and can be carried out as the data are being collected. Unfortunately, with non-Gaussian stimuli, our options are likely to be limited to formulating various optimization problems based on different cost functions.

Shannon information as an objective function

In the method of maximally informative dimensions, the relevance of possible dimensions is characterized by the Shannon information between spikes and stimuli [35]. In that framework, our goal is to try to find dimension $\mathbf{v}$ in the input space such that information in the reduced model (after projecting inputs onto $\mathbf{v}$) and spikes will be equal to the Shannon information of the full, unprojected stimuli and spikes (more accurately, information carried by the arrival time of one spike about the input stimuli) [21]. The latter information provides the maximal amount of information any reduced model can achieve and is formally given by [21]

$$I_{\text{spike}} = \int \mathrm{d}\mathbf{s} P(\mathbf{s}|\text{spike}) \log_2 \left[\frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})} \right] \tag{3}$$

where d$\mathbf{s}$ denotes integration over the full D-dimensional stimulus space. A careful reader might notice that probability distribution $P(\mathbf{s}|\text{no spike})$ does not contribute to the information I_{spike}. A detailed explanation can be found in Reference [21], where it is argued that the contribution from these terms tends to 0 with increasing temporal resolution. In short, fine temporal resolution is necessary for the proper treatment of binary responses such as absence or presence of a spike. Given time bins of width Δt, $P(\mathbf{s}|\text{no spike})/P(\mathbf{s}) = P(\text{no spike}|\mathbf{s})/P(\text{no spike}) = (1 - r(t)\Delta t)(1 - \bar{r}\Delta t)$

DOI: 10.1002/sim

tends to 1 in the limit when the bin size $\Delta t \to 0$. Alternatively, as is also discussed in [21], I_{spike} can be viewed as information about the stimulus carried by the arrival *times* of single spikes, which explains the absence of terms based on 'no-spike' events.

In practice, stimuli are often presented during physiological recordings at relatively low rates, for example, 30–60 Hz. If spike trains are binned at the same time resolution, multiple spikes can occur within the same bin. However, stimulus history can always be re-binned at sufficiently fine temporal resolution (e.g. 1 ms) such that multiple spikes do not occur. While this procedure is the most valid way to apply methods for finding relevant dimensions discussed in this paper, it also offers clues as to how to deal with the case of coarse time binning. Suppose that, with coarse time binning, some stimuli elicited multiple $n_{\mathbf{s}}>1$ spikes. If re-binned at fine temporal resolution, these stimuli would contribute to the spike-conditional stimulus ensemble exactly the number of times as the number of spikes $n_{\mathbf{s}}$ they elicited. Returning to the case of coarse time binning, each stimulus should thus be counted in the spike-conditional ensemble $P(\mathbf{s}|\text{spike})$ as many times as the number of spikes it elicited, and should be counted once toward the *a priori* stimulus ensemble $P(\mathbf{s})$. This argument also underlies the procedure of computing the classic method of STA in the situation where multiple spikes are elicited.

Let us not be discouraged by the fact that the probability distribution $P(\mathbf{s}|\text{spike})$ appearing in equation (3) is never well sampled. In practice, it will not be needed because it can be inverted using Bayes' rule to $P(\text{spike}|\mathbf{s})/P(\text{spike})P(\mathbf{s})$:

$$I_{\text{spike}} = \int \mathrm{d}\mathbf{s} P(\mathbf{s}) \frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} \log_2 \left[\frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} \right] \tag{4}$$

The integral over all inputs weighted by their probability distribution $P(\mathbf{s})$ can now be replaced with a time average [21]:

$$I_{\text{spike}} = \frac{1}{T} \int \mathrm{d}t \frac{r(t)}{\bar{r}} \log_2 \frac{r(t)}{\bar{r}} \tag{5}$$

where the time-dependent spike rate $r(t) = P(\text{spike}|\mathbf{s})/\Delta t$ is measured in time bins of width Δt using multiple repetitions of the same stimulus sequence. The average firing rate $\bar{r} = P(\text{spike})/\Delta t$ is obtained by averaging $r(t)$ in time.

The Shannon information between stimuli and spikes computed according to assumption (1), with some dimension $\mathbf{v}$ as the candidate-relevant stimulus dimension, is given by the Kullback–Leibler divergence:

$$I[\mathbf{v}] = \int \mathrm{d}x \, P_{\mathbf{v}}(x|\text{spike}) \log_2 \left[\frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)} \right] \tag{6}$$

where the integral is taken over all projection values, $x = \mathbf{s} \cdot \mathbf{v}$, of stimuli $\mathbf{s}$ onto dimension $\mathbf{v}$. In equation (6), the probability distribution $P_{\mathbf{v}}(x)$ describes projection values x in the *a priori* stimulus ensemble, and is an average over all inputs varying in all dimensions except for $\mathbf{v}$:

$$P_{\mathbf{v}}(x) = \int \mathrm{d}\mathbf{s} P(\mathbf{s}) \delta(x - \mathbf{s} \cdot \mathbf{v}) \tag{7}$$

where $\delta(x)$ is a delta-function. Similarly, the probability distribution $P_{\mathbf{v}}(x|\text{spike})$ describes projection values x for stimuli that lead to spike:

$$P_{\mathbf{v}}(x|\text{spike}) = \int \mathrm{d}\mathbf{s} P(\mathbf{s}|\text{spike}) \delta(x - \mathbf{s} \cdot \mathbf{v}) \tag{8}$$

According to (2), the ratio $P_{\mathbf{v}}(x|\text{spike})/P_{\mathbf{v}}(x)$ describes the 'effective' input–output function $g(x)$ along the dimension $\mathbf{v}$. In practice, both of the averages (7) and (8) are calculated by binning the range of projection values x and computing histograms normalized to sum to 1. (Considerations for selecting the optimal bin size are discussed below, together with finite data set size effects.)

Obviously, information in the reduced model $I[\mathbf{v}]$ cannot be greater than the information between spikes and unprojected stimuli, I_{spike}. But is it possible for some other dimension $\mathbf{v}$ other than the true relevant dimension $\hat{e}_1$ to provide more information between spikes and projected stimuli? To understand why this cannot be the case, note that our assumptions of the linear–nonlinear model (1) also mean that stimuli $\mathbf{s}$, their projections s_1 onto the relevant dimension, and spikes form a Markov chain [23]: $\mathbf{s}$–s_1-spike. A data-processing inequality [35] then states that the information between two variables that are stochastically related to each other (e.g. s_1 and 'spike') can decrease only when another variable is added to the chain, remaining constant only if the added variable can be deterministically obtained from its neighbors in the chain. Thus, $I(\mathbf{s}, \text{spike}) = I(s_1, \text{spike})$, because s_1 is deterministically obtained from $\mathbf{s}$. On the contrary, information along the Markov chain $\mathbf{s}\cdot\mathbf{v}$–$s_1$-spike will be less than the full amount [20, 22, 23, 36]:

$$I(\mathbf{s}\cdot\mathbf{v}, \text{spike}) < I(s_1, \text{spike}) = I_{\text{spike}} \tag{9}$$

because stimulus projections $\mathbf{s}\cdot\mathbf{v}$ and s_1 are related to each other only probabilistically and solely due to the correlations in the input ensemble. For example, if signals are uncorrelated, $I(\mathbf{s}\cdot\mathbf{v}, \text{spike}) = 0$ for all dimensions $\mathbf{v}$ that are orthogonal to the relevant dimension. If $\mathbf{v}$ has some component along the relevant dimension, it is this component that will determine the correlation between $\mathbf{s}\cdot\mathbf{v}$ and s_1 leading to a nonzero value of $I(\mathbf{s}\cdot\mathbf{v}, \text{spike})$. For correlated inputs, this is also the reason for possible positive values of information $I[\mathbf{v}]$ even when $\mathbf{v}$ is orthogonal to the relevant dimension [22].

Within this framework, the relevant dimension $\hat{e}_1$ is found by maximizing information $I[\mathbf{v}]$ as a function of all components v_i of the dimension $\mathbf{v}$. The possible algorithms for numerical optimization of this function have been described [22, 23, 37] and include a combination of gradient ascent and simulated annealing.

Variance as an objective function

One alternative to maximizing information is to find dimension $\mathbf{v}$ that minimizes a χ^2 difference between measured and predicted spike probabilities averaged across all of the inputs:

$$\chi^2[\mathbf{v}] = \int \mathrm{d}\mathbf{s} P(\mathbf{s}) \left[\frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} - \frac{P(\text{spike}|\mathbf{s}\cdot\mathbf{v})}{P(\text{spike})}\right]^2 \tag{10}$$

which can be rewritten using Bayes' rule as:

$$\chi^2[\mathbf{v}] = \int \mathrm{d}\mathbf{s} P(\mathbf{s}) \left[\frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})} - \frac{P(\mathbf{s}\cdot\mathbf{v}|\text{spike})}{P(\mathbf{s}\cdot\mathbf{v})}\right]^2 \tag{11}$$

If neural spikes are indeed based on one relevant dimension, then this dimension will explain all the variances, so that $\chi^2 = 0$. For all other dimensions $\mathbf{v}$, $\chi^2[\mathbf{v}] > 0$.

DOI: 10.1002/sim

We can expand the square in (11) and average where possible over all input components, except for component $x=\mathbf{s}\cdot\mathbf{v}$ along the trial direction $\mathbf{v}$, to find

$$\chi^2[\mathbf{v}]=\int \mathrm{d}\mathbf{s}\frac{[P(\mathbf{s}|\mathrm{spike})]^2}{P(\mathbf{s})}-\int \mathrm{d}x\frac{[P_{\mathbf{v}}(x|\mathrm{spike})]^2}{P_{\mathbf{v}}(x)} \tag{12}$$

where probability distributions $P_{\mathbf{v}}(x)$ and $P_{\mathbf{v}}(x|\mathrm{spike})$ were defined in equations (7) and (8). Based on equation (12), in order to minimize χ^2, we need to maximize

$$F[\mathbf{v}]=\int \mathrm{d}x\frac{\left[P_{\mathbf{v}}(x|\mathrm{spike})\right]^2}{P_{\mathbf{v}}(x)} \tag{13}$$

which is a Rényi divergence of order 2 between probability distributions $P_{\mathbf{v}}(x|\mathrm{spike})$ and $P_{\mathbf{v}}(x)$. Rényi divergences of order α between two probability distributions $p(x)$ and $q(x) I_\alpha[p||q]$ are defined as [30, 31]

$$I_\alpha[p||q]=\frac{1}{\alpha-1}\int \mathrm{d}x q(x)\left[\frac{p(x)}{q(x)}\right]^\alpha \tag{14}$$

and are part of a family of f-divergence measures that are based on a convex function of the ratio $p(x)/q(x)$ (instead of a power α) [29, 30].

The maximal value for $F[\mathbf{v}]$ that can be achieved by any dimension $\mathbf{v}$ is:

$$F_{\max}=\int \mathrm{d}\mathbf{s}P(\mathbf{s}|\mathrm{spike})\left[\frac{P(\mathbf{s}|\mathrm{spike})}{P(\mathbf{s})}\right] \tag{15}$$

As in the case of I_{spike} above, we need not worry about being able to accurately measure $P(\mathbf{s}|\mathrm{spike})$. Similar to the transition from equation (3) to equation (5), we can use Bayes' rule and the ergodic assumption to compute $F_{\max}$ as a time average:

$$F_{\max}=\frac{1}{T}\int \mathrm{d}t\left[\frac{r(t)}{\bar{r}}\right]^2 \tag{16}$$

The fact that $F[\mathbf{v}]<F_{\max}$ can be seen either by simply noting that $\chi^2[\mathbf{v}]\geqslant 0$ or from the data-processing inequality applies not only to Kullback–Leibler divergence but also to Rényi divergences [23, 29, 30]. The same logic leads to a conclusion that information in a given dimension $I[\mathbf{v}]$ cannot be greater than the overall information I_{spike} between spikes and full, unprojected inputs applies in this case to variance. In other words, the variance in the firing rate explained by a given dimension $F[\mathbf{v}]$ cannot be greater than the overall variance in the firing rate $F_{\max}$. This is because in computing $F[\mathbf{v}]$ we average over all variations in the firing rate that correspond to inputs with the same projection value on the dimension $\mathbf{v}$ and that differ only in projections onto other dimensions.

Comparing equations (6) and (5) that describe the information-based optimization with equations (13) and (16) that describe the variance-based optimization strategies for finding relevant dimensions, one may notice that they differ only by the application of logarithm to the ratio $P(\mathbf{x}|\mathrm{spike})/P(x)$. Because the logarithm is a monotonic function, it is not surprising that both strategies produce the same relevant dimension.

Furthermore, the numerical algorithms that are good for optimizing $I[\mathbf{v}]$ will also work with $F[\mathbf{v}]$. Most notably, the gradient can be computed for both information [22] and variance functional:

$$\nabla_{\mathbf{v}} I = \int \mathrm{d}x\, P_{\mathbf{v}}(x|\text{spike})[\langle \mathbf{s}|x, \text{spike}\rangle - \langle \mathbf{s}|x\rangle] \cdot \left[\frac{\mathrm{d}}{\mathrm{d}x} \log_2 \frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)}\right] \tag{17}$$

$$\nabla_{\mathbf{v}} F = 2\int \mathrm{d}x\, P_{\mathbf{v}}(x|\text{spike})[\langle \mathbf{s}|x, \text{spike}\rangle - \langle \mathbf{s}|x\rangle] \cdot \left[\frac{\mathrm{d}}{\mathrm{d}x} \frac{P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x)}\right] \tag{18}$$

where

$$\langle \mathbf{s}|x, \text{spike}\rangle = \frac{\int \mathrm{d}\mathbf{s}\, \mathbf{s}\delta(x - \mathbf{s}\cdot\mathbf{v})P(\mathbf{s}|\text{spike})}{P(x|\text{spike})} \tag{19}$$

and similarly for $\langle \mathbf{s}|x\rangle$. The two gradients have a similar form: they are just differently weighted sums of STAs conditional on the projection values of stimuli onto the dimension $\mathbf{v}$ for which the gradient of information is being evaluated.

Both information $I[\mathbf{v}]$ and variance $F[\mathbf{v}]$ do not change with the length of the vector. Therefore, $\mathbf{v}\cdot\nabla_{\mathbf{v}} I = \mathbf{v}\cdot\nabla_{\mathbf{v}} F = 0$, as can also be seen directly from equations (17) and (18). Also, both of the two gradients are 0 when evaluated along the true receptive field. This is because, for the true relevant dimension according to which spikes were generated, $\langle \mathbf{s}|s_1, \text{spike}\rangle = \langle \mathbf{s}|s_1\rangle$, a consequence of Markov chain $\mathbf{s}$–s_1-spike property alluded to earlier. The fact that the gradients are zero for the true receptive field $\hat{e}_1$ agrees with the earlier statements that $\mathbf{v} = \hat{e}_1$ maximizes both information $I[\mathbf{v}]$ and variance $F[\mathbf{v}]$.

Illustration

As an illustration of both schemes, let us consider a model visual neuron that responds to stimuli derived from natural scenes. Visual stimuli were derived from movies of walks through a wooded area [37], converted to an 8-bit gray scale. Our goal is to demonstrate that even though the correlations present in the ensemble are non-Gaussian, they can be removed successfully from the estimate of relevant dimensions. The example model neuron is taken to mimic the properties of simple cells found in the primary visual cortex. It has a single relevant dimension $\hat{e}_1$, which is phase and orientation sensitive and is shown in Plate 1(a). A given stimulus $\mathbf{s}$ leads to a spike if the projection $s_1 = \mathbf{s}\cdot\hat{e}_1$ reaches a threshold value θ in the presence of noise:

$$\frac{P(\text{spike}|\mathbf{s})}{P(\text{spike})} \equiv f(s_1) = \langle H(s_1 - \theta + \xi)\rangle \tag{20}$$

where a Gaussian random variable ξ of variance σ^2 models additive noise, and the function $H(x) = 1$ for $x>0$, and zero otherwise. Together with the relevant dimension $\hat{e}_1$, the parameters θ for threshold and the noise variance σ^2 determine the input–output function. In what follows, we will always measure σ in units of standard deviation of stimulus projections onto the relevant dimensions and use it as a measure of signal to noise.

Plate 1 shows that it is possible to obtain a good estimate of the relevant dimension $\hat{e}_1$ by maximizing information, as shown in panel (b), or variance, as shown in panel (c). The final value of projection depends on the size of the data set, as discussed below. In the example shown in Plate 1, there were $\approx 50\,000$ spikes with average probability of spike ≈ 0.05 per frame, and the reconstructed vector has a projection $\hat{v}_{\max} \cdot \hat{e}_1 = 0.98$ when maximizing either information or variance. Having estimated the relevant dimension, one can proceed to sample the nonlinear input–output function. This is done by constructing histograms for $P(\mathbf{s} \cdot \hat{v}_{\max})$ and $P(\mathbf{s} \cdot \hat{v}_{\max}|\text{spike})$ of projections onto vector $\hat{v}_{\max}$ found by maximizing either information or variance, and taking their ratio, as in equation (2). In Plate 1(d), the spike probability for the reconstructed neuron $P(\text{spike}|\mathbf{s} \cdot \hat{v}_{\max})$ (crosses and circles) is compared with the probability $P(\text{spike}|s_1)$ used in the model (solid line). A good match is obtained.

In actuality, reconstructing even just one relevant dimension from neural responses to correlated non-Gaussian inputs, such as those derived from real world, is not an easy problem. This fact can be appreciated by considering the estimates of relevant dimension obtained from the STA, shown in panel (e). Correcting the STA by second-order correlations of the input ensemble through a multiplication by the inverse covariance matrix results in a very noisy estimate, shown in panel (f). It has a projection value of 0.25. Attempt to regularize the inverse of covariance matrix results in a closer match to the true relevant dimension [7, 10–12, 38] and has a projection value of 0.8, as shown in panel (g). In these simulations, the regularization was performed by setting aside $\frac{1}{4}$ of the data as a test data set, and choosing a cutoff on the eigenvalues of the input covariances matrix that would give the maximal information value on the test data set [7, 12, 39]. Despite the less noisy appearance of the regularized decorrelated STA and its closer match to the true relevant dimension compared to decorrelated STA, it is also known [7, 22, 37, 40, 41] to have systematic deviations. These deviations stem from two separate causes: neural noise and non-Gaussian stimulus correlations (for nonlinear neurons) [22]. Regularization aims to compensate for effects due to neural noise. Because regularization is done with respect to just one parameter—the overall smoothness in the relevant dimension—the result is often biased toward lower spatial frequencies, and can eliminate some of the genuine structure of the relevant dimensions, cf. Plates 1 and 2. Incorporating more adjustable parameters makes the algorithm closer to the full optimization using information or variance. In Plate 2, we show the spatial frequency tuning along the preferred orientation for the receptive fields of Plate 1. Despite the relatively large number of spikes used in these simulations (50 000), the spatial frequency tuning remains altered in the decorrelated STA before and after regularization.

In the absence of neural noise, there is no justification for the regularization. Even in this case, the decorrelated STA will differ from the true relevant dimensions, as has been demonstrated analytically [22, 40, 41] and with numerical simulations [22].

MULTIPLE RELEVANT DIMENSIONS

So far we have considered how to find a single relevant dimension. Typically, neural spiking depends on more than one relevant dimension. If that is the case, the maximal information accounted for by the single most informative dimension will be smaller than the overall value I_{spike}, and similarly the variance $F[\mathbf{v}]$ accounted for by this dimension will be smaller than the maximal value $F_{\max}$. Both measures would then indicate that we need to look for more relevant dimensions.

In the previous section, we considered how the relevance of single dimensions in the input space can be quantified by the amount of Shannon information (6) or variance (13) they account for. By analogy, the relevance of several dimensions $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ can be computed based on the

DOI: 10.1002/sim

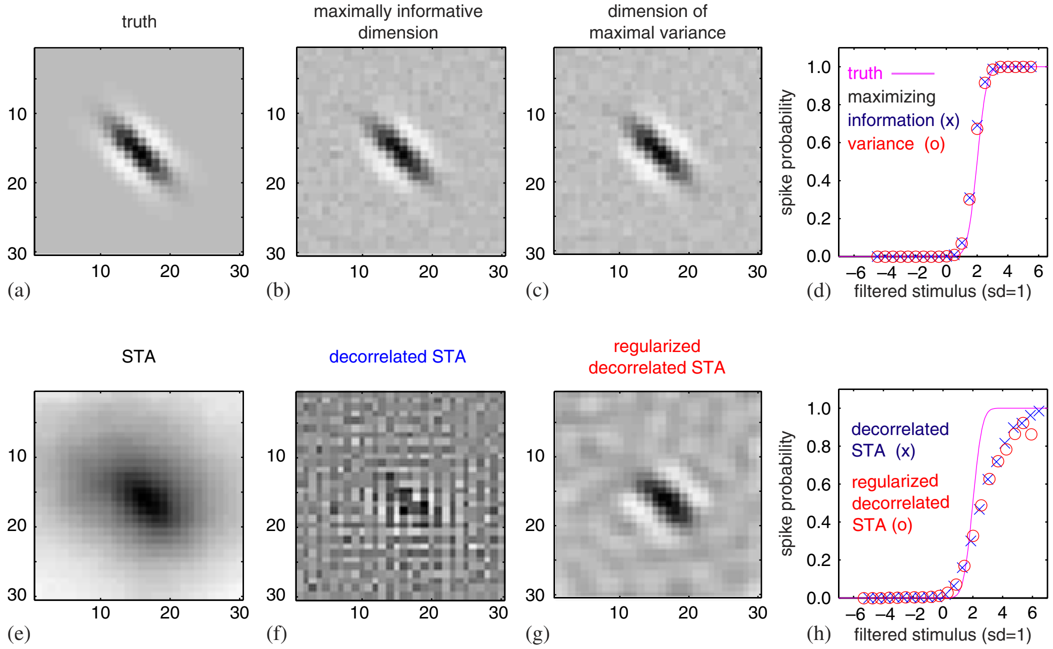

Plate 1. Analysis of a model visual neuron with one relevant dimension shown in (a). Panels (b) and (c) show normalized vectors $\hat{v}_{\max}$ found by maximizing information and variance, respectively; (d) the probability of a spike $P(\text{spike}|\mathbf{s}\cdot\hat{v}_{\max})$ (blue crosses—information maximization, red circles-variance maximization) is compared to $P(\text{spike}|s_1)$ used in generating spikes (solid line). Parameters of the model are $\sigma=0.5$ and $\theta=2$, both given in units of standard deviation of s_1, which are also the units for the x-axis in panels (d, h). The spike-triggered average (STA) is shown in (e). An attempt to remove correlations according to the reverse correlation method, $C^{-1}_{a\,priori}\mathbf{v}_{\text{sta}}$ (decorrelated STA), is shown in panel (f) and in panel (g) with regularization, as described in the text and in Reference [37]. In panel (h), the spike probabilities given as projections onto the dimensions obtained as decorrelated STA (blue crosses) and regularized decorrelated STA (red circles) are compared with a spike probability used to generate spikes (solid line).

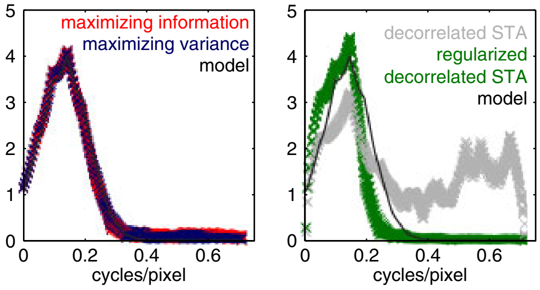

Plate 2. Spatial frequency profiles along the preferred orientation for relevant dimensions from Plate 1. The left panel shows spatial frequency profiles for the relevant dimensions obtained by maximizing variance (blue line, left panel) and information (red line, left panel). The right panel shows the same analysis for relevant dimensions obtained as decorrelated STA before (gray) and after regularization (green). The black line in both panels shows the spatial frequency profile for the true relevant dimension. Despite the rather large number of spikes (50 000) used in these simulations, the decorrelated STA remains biased toward higher spatial frequencies, while the regularized decorrelated STA remains biased to low spatial frequencies.

DOI: 10.002/sim

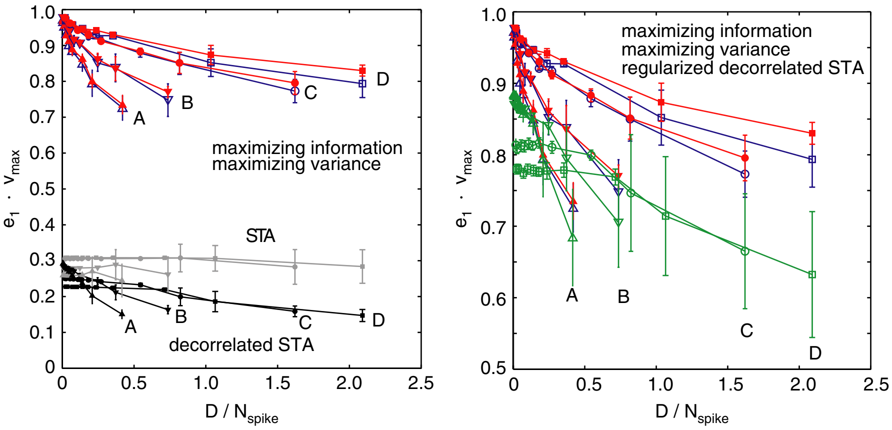

Plate 3. Projection of vector $\hat{v}_{\max}$ obtained by maximizing information (red filled symbols) or variance (blue open symbols) on the true relevant dimension $\hat{e}_1$ is plotted as a function of ratio between stimulus dimensionality D and the number of spikes N_{spike}, with $D = 900$. Simulations were carried out for model visual neurons with one relevant dimension from Plate 1(a) and the input–output function described by threshold $\theta = 2.0$ and noise standard deviation $\sigma = 1.5, 1.0, 0.5$, and 0.25 for groups labeled A (△-symbols), B (▽-symbols), C (◯-symbols), and D (□-symbols), respectively. All parameter values are given in units of standard deviation of stimulus projection values along the relevant dimension. The left panel also shows results obtained using spike-triggered average (STA, gray) and decorrelated spike-triggered average (dSTA, black). In the right panel, we replot results for information and variance optimization together with those for regularized decorrelated spike-triggered average (RdSTA, green open symbols). All error bars show standard deviations. Information and variance optimization outperform the methods of STA and dSTA (panel A) by a large margin, and also provide better or equal (for very low spike number) results compared with RdSTA, cf. right panel. As expected, none of the three methods based on STA converge to the true relevant dimension for large spike numbers. While maximizing information or variance produces similar results, information maximization achieves significantly smaller errors than the variance maximization when compared across all simulations for the four different model cells and spike numbers ($p < 10^{-4}$, paired t-test).

DOI: 10.002/sim

following two multi-point probability distributions of projection values $x_1, x_2, \ldots, x_n$ along them:

$$\begin{aligned} P_{\mathbf{v}_1,\ldots,\mathbf{v}_n}(\{x_i\}|\text{spike}) &= \int \mathrm{d}\mathbf{s} \prod_{i=1}^{n} \delta(x_i - \mathbf{s}\cdot\mathbf{v}_i) P(\mathbf{s}|\text{spike}) \\ P_{\mathbf{v}_1,\ldots,\mathbf{v}_n}(\{x_i\}) &= \int \mathrm{d}\mathbf{s} \prod_{i=1}^{n} \delta(x_i - \mathbf{s}\cdot\mathbf{v}_i) P(\mathbf{s}) \end{aligned} \tag{21}$$

For example, in the case of two dimensions, the information along the two dimensions $\mathbf{v}_1$ and $\mathbf{v}_2$ considered simultaneously is given by [22]

$$I[\mathbf{v}_1, \mathbf{v}_2] = \int \mathrm{d}x_1\, \mathrm{d}x_2\, P(x_1, x_2|\text{spike}) \log_2 \frac{P(x_1, x_2|\text{spike})}{P(x_1, x_2)} \tag{22}$$

Contribution to the variance by these two dimensions can be written in terms of the same two multi-point probability distributions $P_{\mathbf{v}_1,\mathbf{v}_2}(x_1, x_2)$ and $P_{\mathbf{v}_1,\mathbf{v}_2}(x_1, x_2|\text{spike})$ given by equation (21):

$$F[\mathbf{v}_1, \mathbf{v}_2] = \int \mathrm{d}x_1\, \mathrm{d}x_2\, P(x_1, x_2|\text{spike}) \frac{P(x_1, x_2|\text{spike})}{P(x_1, x_2)} \tag{23}$$

For a neuron whose spikes are based on two relevant dimensions $\hat{e}_1$ and $\hat{e}_2$, these dimensions can be recovered by maximizing either the Shannon information (22) or variance (23) as a function of two dimensions $\mathbf{v}_1$ and $\mathbf{v}_2$. The data-processing inequality in this case states that the overall information between (unprojected) stimuli $\mathbf{s}$ and spikes equals information between spikes and stimuli projected onto the plane of relevant dimensions $\hat{e}_1$ and $\hat{e}_2$, $I(\mathbf{s}; \text{spike}) = I(s_1, s_2; \text{spike})$. Information will be smaller for all pairs of dimensions $\{\mathbf{v}_1, \mathbf{v}_2\}$ that deviate from the relevant plane, because projections onto such dimensions will not be deterministically related to stimulus projections onto the relevant plane. The choice of basis is not restricted in this formulation, any two nondegenerate linear combinations of relevant dimensions will account for the same amount of information or variance. Information or variance will not increase if more dimensions are added to the relevant ones. This can be used as a criterion for determining the number of relevant dimensions. If either I_{spike} or $F_{\max}$ is known, then the number of relevant dimensions could be determined by iteratively adding the relevant dimensions until these maximal values are reached within experimental uncertainty.

To find, for example, two most relevant dimensions, one would optimize either $I(\mathbf{v}_1, \mathbf{v}_2)$ or $F(\mathbf{v}_1, \mathbf{v}_2)$ with respect to both components describing dimension $\mathbf{v}_1$ and dimension $\mathbf{v}_2$. In practice, this could be done by searching for one dimension while keeping the second one fixed, and then alternating the dimensions being optimized and the one being held fixed [22, 37]. If stimuli are uncorrelated, then the first most relevant dimension can be found separately, i.e. assuming that no other dimensions are relevant. The second dimension can be found by maximizing either (22) or (23) with the first dimension fixed [22]. This can be seen by noting that the gradients of information or variance along one of the most relevant dimensions (denoted as $\hat{e}_1$) are

$$\nabla I(\hat{e}_1) = \int \mathrm{d}s_1\, \mathrm{d}s_2\, P(s_1, s_2) \langle \mathbf{s}|s_1, s_2\rangle \frac{P(\text{spike}|s_1, s_2) - P(\text{spike}|s_1)}{P(\text{spike})} \frac{\mathrm{d}}{\mathrm{d}s_1} \log_2 \frac{P(s_1|\text{spike})}{P(s_1)} \tag{24}$$

$$\nabla F(\hat{e}_1) = \int \mathrm{d}s_1\, \mathrm{d}s_2\, P(s_1, s_2) \langle \mathbf{s}|s_1, s_2\rangle \frac{P(\text{spike}|s_1, s_2) - P(\text{spike}|s_1)}{P(\text{spike})} \frac{\mathrm{d}}{\mathrm{d}s_1} \frac{P(s_1|\text{spike})}{P(s_1)} \tag{25}$$

which can be obtained starting from the expression for the gradient (17) valid for any dimension and expanding the conditional averages as

$$\langle \mathbf{s}|s_1\rangle = \frac{\int \mathrm{d}s_2 P(s_1, s_2)\langle \mathbf{s}|s_1, s_2\rangle}{P(s_1)} \tag{26}$$

$$\langle \mathbf{s}|s_1, \text{spike}\rangle = \frac{\int \mathrm{d}s_2 P(s_1, s_2|\text{spike})\langle \mathbf{s}|s_1, s_2\rangle}{P(s_1|\text{spike})} \tag{27}$$

In the last expression, $\langle \mathbf{s}|s_1, s_2, \text{spike}\rangle = \langle \mathbf{s}|s_1, s_2\rangle$, because knowledge of the two relevant variables $\{s_1, s_2\}$ determines the spike probability.

Let us consider first the case of uncorrelated inputs. In this case,

$$\langle \mathbf{s}|s_1, s_2\rangle = \mathbf{a} + s_1\hat{e}_1 + s_2\hat{e}_2 \tag{28}$$

where **a** is some constant vector. The term with the constant vector **a** integrates to zero, because $\int \mathrm{d}s_2 P(s_1, s_2)(P(\text{spike}|s_1, s_2) - P(\text{spike}|s_2)) = 0$. Contributions from the terms linear in s_1 and s_2 of equation (28) in equations (17) and (18) describe the amplitude of the gradient of information along the dimensions $\hat{e}_1$ and $\hat{e}_2$, which would be zero for the most informative dimension *within* the relevant subspace. In other words, for uncorrelated inputs the most informative dimension within the relevant subspace is also the most relevant overall.

The above argument also extends to correlated Gaussian inputs. This is because the relevant dimensions can always be described in the coordinate system that corresponds to independent axes of the multivariate Gaussian. Rescaling by the standard deviation along these axes converts the distribution to those of independent inputs where the conditional average (28) is a vector that depends linearly on s_1 and s_2. Because rescaling introduces only constant factors that do not depend on s_1 and s_2, the linear dependence (28) is unaltered. Thus, the argument based on equation (28) applies not only to uncorrelated inputs, but also to correlated Gaussian inputs. We can search for fewer relevant dimensions than are actually relevant without the need to update the already found relevant dimensions when additional dimensions are introduced, provided stimulus distribution is a correlated Gaussian one. With natural stimuli, however, it is necessary to allow for the previously found relevant dimensions to change when searching for additional ones. This is because, in the presence of non-Gaussian correlations, the single most informative dimension found assuming that no other relevant dimensions exist might deviate from a plane of two most relevant dimensions [22]. This statement applies whether we are maximizing information or variance.

COMPARISON OF PERFORMANCE WITH FINITE DATA

When either of the two objective functions is calculated from a finite data set, the optimal vector $\hat{v}$ will deviate from the true relevant dimension $\hat{e}_1$. The deviation $\delta\mathbf{v} = \hat{v} - \hat{e}_1$ arises because the probability distributions (7) and (8) are estimated from experimental histograms and differ from the distributions found in the limit of infinite data size. The effects of noise on the reconstruction can be characterized by taking the dot product between the relevant dimension and the optimal vector for a particular data sample: $\hat{v}\cdot\hat{e}_1 = 1 - (1/2)\delta\mathbf{v}^2$, where both $\hat{v}$ and $\hat{e}_1$ are normalized, and $\delta\mathbf{v}$ is by definition orthogonal to $\hat{e}_1$. The deviation $\delta\mathbf{v} = -H^{-1}\nabla F$, where H is the Hessian of the objective function (either information or variance), and both H and ∇F are evaluated for $\mathbf{v} = \hat{e}_1$.

DOI: 10.1002/sim

In the limit of infinite data, $\nabla F(\hat{e}_1)=0$. However, for a data set of finite size, the gradient will not be zero, and its magnitude together with the Hessian of the optimization function determines (in the quadratic approximation) the deviation of the relevant dimension computed for a particular instantiation of the noise from the true dimension. Just like in the case of optimizing information [22], the Hessian of variance H_{ij} when evaluated along the optimal dimension $\hat{e}_1$ is a weighted average of covariance matrices $C_{ij}(x)$:

$$H_{ij}=2\int \mathrm{d}x\, P(x)C_{ij}(x)\left[\frac{\mathrm{d}}{\mathrm{d}x}\,\frac{P(x|\text{spike})}{P(x)}\right]^2 \tag{29}$$

where

$$C_{ij}(x)=\langle s_i s_j|x\rangle-\langle s_i|x\rangle\langle s_j|x\rangle \tag{30}$$

is the covariance matrix of all inputs that have projection x along the optimal dimension.

Let us show that the expected value of the gradient is zero for the optimal direction. We start by substituting back expression (19) and its analog for $\langle \mathbf{s}|x\rangle$ into expression (18) for the gradient:

$$\nabla F=2\int \mathrm{d}x\int \mathrm{d}\mathbf{s}\,\mathbf{s}\delta(x-\mathbf{s}\cdot\mathbf{v})[P_N(\mathbf{s}|\text{spike})-P(\mathbf{s})P_{\mathbf{v}N}(x|\text{spike})/P_{\mathbf{v}}(x)]\frac{\mathrm{d}}{\mathrm{d}x}\,\frac{P_{\mathbf{v}N}(x|\text{spike})}{P_{\mathbf{v}}(x)} \tag{31}$$

where subscript N was added to emphasize that probability distributions $P_N(\mathbf{s}|\text{spike})$ and $P_{\mathbf{v}N}(x|\text{spike})$ are estimated as histograms and will vary across different instantiations of neural noise, whereas probability distributions $P(\mathbf{s})$ and $P_{\mathbf{v}}(x)$ depend only on the input distribution. Our assumptions are that different stimuli elicit spikes independently. Given that a particular stimulus pattern $\mathbf{s}$ was encountered $N_{\mathbf{s}}$ times in the spike-conditional ensemble, and that there were N_{spike} spikes overall, we estimate the spike-conditional probability as $P_N(\mathbf{s}|\text{spike})=N_{\mathbf{s}}/N_{\text{spike}}$. On average, $\langle P_N(\mathbf{s}|\text{spike})\rangle=P(\mathbf{s}|\text{spike})$ and, similarly, $\langle P_{\mathbf{v}N}(x|\text{spike})\rangle=P_{\mathbf{v}}(x|\text{spike})$. Now we can use our primary assumption (1), $P(\mathbf{s}|\text{spike})=P(\mathbf{s})P(s_1|\text{spike})/P(s_1)$ (s_1 represents projections onto the true relevant dimension) to show that the expected value for the term in square brackets in (31) is zero. Note that the expected value for the cross-terms

$$\langle[\delta P_N(\mathbf{s}|\text{spike})-P(\mathbf{s})\delta P_{\mathbf{v}N}(x|\text{spike})/P_{\mathbf{v}}(x)]\delta P_{\mathbf{v}N}(x'|\text{spike})\rangle=0 \tag{32}$$

where $\delta P_N(\mathbf{s}|\text{spike})=P_N(\mathbf{s}|\text{spike})-P(\mathbf{s}|\text{spike})$ and $\delta P_{\mathbf{v}N}(x|\text{spike})=P_{\mathbf{v}N}(x|\text{spike})-P_{\mathbf{v}N}(x|\text{spike})$ describe deviations in the probability distributions due to finite sampling; the expected value in (32) is zero whether or not x' equals x. Putting it all together, we find that the expected value of the gradient ∇F is zero. In other words, there is no specific direction toward which the deviations $\delta\mathbf{v}$ are biased.

Next, let us compute the expected spread in the optimal dimensions around the true dimension $\hat{e}_1$. To achieve this, we need to evaluate $\langle\delta\mathbf{v}^2\rangle=\mathrm{Tr}[H^{-1}\langle\nabla F\nabla F^{\mathrm{T}}\rangle H^{-1}]$, where average is taken over different instantiations of neural noise [42]. In the expression $\langle\nabla F_i\nabla F_j\rangle$, the leading terms are $\sim 1/N_{\text{spike}}$. They arise from covariance of terms in square brackets of equation (31) for the two gradients. Because fluctuations in estimates of probability distribution for different stimuli are independent, we have

$$\langle\delta P_N(\mathbf{s}|\text{spike})\delta P_N(\mathbf{s}'|\text{spike})\rangle=P(\mathbf{s}|\text{spike})\delta(\mathbf{s}-\mathbf{s}')/N_{\text{spike}} \tag{33}$$

$$\langle\delta P_{\mathbf{v}N}(x|\text{spike})\delta P_{\mathbf{v}N}(x'|\text{spike})\rangle=P_{\mathbf{v}}(x|\text{spike})\delta(x-x')/N_{\text{spike}} \tag{34}$$

DOI: 10.1002/sim

Similarly, one can also find, based on (8) and (33), that

$$\langle \delta P_N(\mathbf{s}|\text{spike})\delta P_{\mathbf{v}N}(x'|\text{spike})\rangle = P(\mathbf{s}|\text{spike})\delta(x'-\mathbf{s}\cdot\mathbf{v})/N_{\text{spike}} \tag{35}$$

Using expressions (33)–(35), we find that $\langle \nabla F_i \nabla F_j\rangle = D_{ij}/N_{\text{spike}}$, where

$$D_{ij} = 4\int \mathrm{d}x\, P(x|\text{spike})C_{ij}(x)\left[\frac{\mathrm{d}}{\mathrm{d}x}\,\frac{P(x|\text{spike})}{P(x)}\right]^2 \tag{36}$$

Therefore, an expected error in the reconstruction of the optimal filter by maximizing variance is inversely proportional to the number of spikes:

$$\hat{v}\cdot\hat{e}_1 \approx 1-\frac{1}{2}\langle\delta\mathbf{v}^2\rangle = 1-\frac{\mathrm{Tr}'[H^{-1}DH^{-1}]}{2N_{\text{spike}}} \tag{37}$$

where Tr$'$ denotes the trace taken in the subspace orthogonal to the relevant dimension (this is because deviations along the relevant dimension have no meaning [22], which mathematically manifests itself in the fact that $\hat{e}_1$ corresponds to a zero eigenvalue of matrices H, D, and A below).

By comparison, the corresponding expected value of the projection between the reconstructed vector by maximizing information and the relevant direction $\hat{e}_1$ was shown to be [22]

$$\hat{v}\cdot\hat{e}_1 \approx 1-\frac{1}{2}\langle\delta\mathbf{v}^2\rangle = 1-\frac{\mathrm{Tr}'[A^{-1}]}{2N_{\text{spike}}} \tag{38}$$

where

$$A_{ij} = \int \mathrm{d}x\, P(x|\text{spike})C_{ij}(x)\left[\frac{\mathrm{d}}{\mathrm{d}x}\ln\frac{P(x|\text{spike})}{P(x)}\right]^2 \tag{39}$$

is the Hessian of information evaluated at the optimum.

With either optimization strategy, the error $\sim D/(2N_{\text{spike}})$ increases in proportion to the dimensionality D of inputs and decreases as more spikes are collected [22, 23]. Using a version of the Cauchy–Schwarz inequality, it can be shown that the average expected error for maximizing information (38) is less than or equal to the average expected error for maximizing variance (37). While complete derivation is provided in Appendix A.1, an intuition for the derivation can be obtained by considering the approximation where $C_{ij}(x)=C_{ij}f(x)$, where $f(x)$ is some positive function. Then, matrices A, H, and D can all be written as a product of C and averages of the combinations of functions $a(x)$ and $b(x)$, defined as

$$a^2(x) = f(x)\left[\frac{g'(x)}{g(x)}\right]^2, \quad b^2(x) = 4f(x)[g'(x)]^2$$

$$A_{ij} = C_{ij}\int \mathrm{d}x\, P(x|\text{spike})a^2(x), \quad D_{ij} = C_{ij}\int \mathrm{d}x\, P(x|\text{spike})b^2(x)$$

$$H_{ij} = C_{ij}\int \mathrm{d}x\, P(x|\text{spike})b(x)a(x)$$

Application of the Cauchy–Schwarz inequality, $\langle b^2\rangle/\langle ab\rangle^2 \geqslant 1/\langle a^2\rangle$, where averaging $\langle\ldots\rangle \equiv \int \mathrm{d}x\, P(x|\text{spike})$ is carried out with respect to the probability distribution $P(x|\text{spike})$, shows that

DOI: 10.1002/sim

the expected error for maximizing information (38) is less than or equal to that for maximizing variance (38), see Appendix A.1 for a full derivation. The lowest possible error that can be achieved by any unbiased method is derived in Appendix A.2 based on Fisher information [35].

Compared to the reverse correlation method, both of the expected errors given by equations (37) and (38) are of the same order as errors expected of the reverse correlation method when it is applied to Gaussian ensembles. The latter have been shown for correlated [22] and white noise inputs [6, 23] to be

$$\mathrm{Tr}'[C^{-1}]/[2N_{\mathrm{spike}}\langle g'^2(s_1)\rangle] \tag{40}$$

If the gain function $g(s_1)$ (which in this case depends only on a single relevant variable s_1) is a relatively sharp sigmoid, then integrals in (29, 36, 39) can be taken by the steepest descent method [43], and the error estimates associated with maximizing information or variance will be smaller than those of the reverse correlation method (40) by a factor of $g(t)$, where the value $s_1 = t$ corresponds to the maximum in the derivative of the gain function $g(s_1)$. This has been reported to be observed in numerical simulations [23]. While the errors expected for maximizing information or variance and those of the reverse correlation technique are similar when applied to Gaussian inputs, the reverse correlation will have larger errors if applied to the non-Gaussian ensemble.

Performance of the information maximization method was already explored in the relatively well-sampled regime, with $0.01 \lesssim D/N_{\mathrm{spike}} \lesssim 0.1$ [22]. For data sets of such size, the expected errors indeed decrease initially as $1/N_{\mathrm{spike}}$. At small spike numbers $D/N_{\mathrm{spike}} \gtrsim 0.03$, corrections of $\sim N_{\mathrm{spike}}^{-2}$ become important, but fortunately have a positive sign, so that the purely linear approximation underestimates the effectiveness of relatively small data sets [22].

The focus of numerical simulations in this paper is on the relatively under-sampled regime, $D \sim N_{\mathrm{spike}}$, where the asymptotic results (37) and (38) do not necessarily apply. The results of simulations for various numbers of trials, and therefore numbers of spikes, are shown in Plate 3 as a function of D/N_{spike} for four different model visual neurons. The model cells had one relevant dimension, shown in Plate 1(a). The input–output functions (20) were described by threshold $\theta = 2.0$ and noise standard deviation $\sigma = 1.5$, 1.0, 0.5, 0.25 for groups labeled A, B, C, and D, respectively. Across the four different model cells, simulations cover the range $0.1 \lesssim D/N_{\mathrm{spike}} \lesssim 3$. Identical numerical algorithms, including the binning procedure, were used for maximizing variance and information. The only differences between algorithms were, of course, in the expressions for objective functions and their gradients. Computations were performed with number of bins equal to 15, 21, 32, and 64 for cells A–D, respectively. These numbers of bins were selected such that, for each cell, the bin width would be $\lesssim \sigma/2$ and, therefore, adequate sampling of the input–output function [23] could be achieved. Generally, the bin size should be, on one hand, sufficiently small to adequately represent the input–output function, but not so small that some bins are not sampled at all by the ensemble of presented stimuli. This is because both neural noise and binning approximations (even without neural noise) contribute to errors in estimating the input–output function. Roughly, the noise contribution is $\sim [g(x)]^2/[P(x|\mathrm{spike})\Delta x]$, whereas the error due to binning approximation is $\sim (g'(x))^2 \Delta x$. The optimal bin size is then $\Delta x \sim g(x)/[g'(x)P(x|\mathrm{spike})]$. Typically, the smallest bin size should be in the center of the distribution. For uniform bin size and a threshold-like nonlinearity, this argument leads to $\Delta x \lesssim \sigma$, where σ is the width of the threshold transition. Therefore, more fine binning may be required to describe very sharp input–output functions. For example, analysis for a model cell D with $\sigma = 0.25$ was carried out with 64 bins. Given that stimulus projections typically cover the range ±5 when measured in standard deviations, the bin size with 64 bins was $\approx 0.16 < \sigma$. However, the improvement in projection values of relevant dimensions onto the true relevant dimensions from using 64 bins compared with 15 bins was marginal. Empirically, uniform binning with 15–20 bins covering all of the range of projection values encountered in the stimulus ensemble is adequate to describe the input–output functions of neurons in the primary visual cortex [37].

The relevant dimension for each simulated spike train was obtained as an average of four jack-knife estimates computed by setting aside $\frac{1}{4}$ of the data as a test set. While maximizing information or variance on the training part of the data set, performance of the candidate relevant dimension was tested on the test data set after each line maximization. The dimension with the best performance on the test data set was selected as the relevant one. Results are shown after 1000 line optimizations ($D = 900$). In Plate 3, we also show results for relevant dimensions obtained by computing STA, decorrelated STA(dSTA), and regularized decorrelated STA(RdSTA). Regularization of the dSTA was also performed by setting aside $\frac{1}{4}$ of the data as a test data set, and choosing a cutoff on the eigenvalues of the input covariances matrix that would give the maximal information value on the test data set [7, 12, 39]. Selecting the dimension for the RdSTA based on variance, instead of information, did not produce significantly different results (data not shown).

As can be seen in Plate 3, a good reconstruction with projection values >0.7 can be obtained by maximizing either information or variance, even in the severely undersampled regime of $D < N_{\text{spike}}$. By comparison, the methods of STA and dSTA produce dimensions with projection values ~ 0.3 onto the relevant one. Although the regularized dSTA provides a remarkable improvement over the dSTA (panel B), it does not outperform full information and variance maximization, even for the lowest of the spike numbers tested and the lowest signal-to-noise ratio (cell A). In general, the less noisy the cell is, the more substantial the improvement that can be achieved by doing full information or variance optimization compared with computing the RdSTA. The same rule applies to increasing the number of spikes for the same cell. In this case, the most important drawback of the RdSTA is that it does not converge to the true relevant dimension [22, 40, 41], cf. right panels of Plates 2 and 3.

Comparison between information and variance optimization strategies reveals that both algorithms result in similar errors throughout the studied range of spike numbers and for different neural noise levels. Even though most of the error bars (showing standard deviations) overlap, projection values between the true relevant dimensions and those obtained by maximizing information are slightly larger in nearly all data points than those for dimensions obtained by maximizing the variance. This difference was quantified with a paired *t*-test to be highly significant ($p < 10^{-4}$) when simulations for all values of spike numbers and noise levels were included. Thus, the same inequality between errors of the two algorithms that was obtained in the asymptotic limit of large spike numbers (A11) are also empirically observed for numbers was spikes typical of physiological recordings.

DISCUSSION

The application of system identification techniques to the study of neural systems has a long history. While early on it was realized that, for Gaussian inputs [9, 18, 25, 26], the relevant dimensions can be found without precise knowledge of subsequent nonlinearities, the methods of characterizing neural feature selectivity with non-Gaussian inputs employed an iterative adjustment of features and corresponding nonlinearities [17, 32, 34, 44]. Such iterative adjustments often require

DOI: 10.1002/sim

inversion of nonlinearities, thus severely limiting the range of neurobiological systems that could be studied.

The main advantage of information and variance optimization schemes is that they allow one to bypass the steps of nonlinearity adjustment and its inversion in order to obtain a new estimate of the relevant dimension. Instead, information and variance optimization strategies consist only of feature optimization. Information or variance provides a measure to judge the best quality of a 'fit' that could possibly be obtained with a given dimension to the neural data at hand. Their computation according to equations (6) and (13) is based on an implicit calculation of the best possible nonlinearity for the candidate dimension and the neural data. In other words, instead of matching the recorded spike train directly to that predicted by a set of relevant features and guessed nonlinearities, one can create a measure (not necessarily a least-square one, e.g. Shannon information) to match the change in the probability distributions upon observation of a spike with inputs of reduced dimensionality. Ideally,

$$\frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})}=\frac{P(\mathbf{s}\cdot\mathbf{v}|\text{spike})}{P(\mathbf{s}\cdot\mathbf{v})}$$

could be achieved, with several, if not one, relevant dimensions. Matching probability distributions offers several advantages. First, the nonlinearity does not need to be inverted, and, in fact, nonlinearity could be of any form, as long as it is smooth enough to be binned at some reasonable resolution [23]. Second, there is no need to smooth spike trains in time in order for them to be compared. As a side bonus, the ratio

$$F[\mathbf{v}]/F_{\max} \tag{41}$$

measures the variance explained by the *nonlinear* model (1) of reduced dimensionality. It is designed for, and is based on, the binary nature of spike trains. Of course, the relevance of dimensions found by maximizing information can be double checked by computing the ratio (41) and, *vice versa*, relevance of dimensions found by maximizing variance can be quantified using the ratio $I[\mathbf{v}]/I_{\text{spike}}$.

The important component of both STA and covariance methods is to separate features that represent overall correlations present in the stimulus ensemble from those that are relevant to neural response. When inputs are truly white (without any correlations), this is not an issue. When inputs are described by a correlated Gaussian distribution, both dimensions that describe a change in the mean (spike-triggered average) [10–12] and the variance (eigenvectors of the change in variance distribution [1, 5, 7, 20, 21]) need to be 'decorrelated' by multiplying them with the inverse of the stimulus covariance matrix. This procedure is prone to noise amplification and typically requires regularization [7, 10–12, 22, 37–39]. In inverse problem theory, multiplication by an inverse of covariance matrix is often transformed into an optimization problem [45], just as has been done to study the feature selectivity of simple cells in the primary visual cortex [38]. Both information and variance optimization strategies allow one to correct for correlations in the inputs without an explicit inversion of the input covariance matrix. Intuitively, this is because the objective function in either case is based on the ratio of probability distributions $P_{\mathbf{v}}(x|\text{spike})/P_{\mathbf{v}}(x)$ evaluated along the dimension $\mathbf{v}$, cf. Table I. If the dimension $\mathbf{v}$ describes mostly input correlations, then the two probability distributions $P_{\mathbf{v}}(x|\text{spike})$ and $P_{\mathbf{v}}(x)$ will be similar, so that both objective functions of information or variance, which depend on their ratio, will be small.

As with the methods based on STA, information and variance optimizations are also prone to overfitting and regularization can make a big difference in predicting power. All of the optimization

DOI: 10.1002/sim

Table I. Comparison of information and variance optimization strategies.

	Information	Variance
Objective function	$I[\mathbf{v}]=\int \mathrm{d}x\, P_{\mathbf{v}}(x\|\text{spike})\log_2 \frac{P_{\mathbf{v}}(x\|\text{spike})}{P_{\mathbf{v}}(x)}$ Kullback–Leibler divergence	$F[\mathbf{v}]=\int \mathrm{d}x \frac{[P_{\mathbf{v}}(x\|\text{spike})]^2}{P_{\mathbf{v}}(x)}$ Rényi divergence ($\alpha=2$)
Gradient	$\int \mathrm{d}x\, P_{\mathbf{v}}(x\|\text{spike})[\langle\mathbf{s}\|x,\text{spike}\rangle-\langle\mathbf{s}\|x\rangle]$ $\times\left[\frac{\mathrm{d}}{\mathrm{d}x}\log_2\frac{P_{\mathbf{v}}(x\|\text{spike})}{P_{\mathbf{v}}(x)}\right]$	$2\int \mathrm{d}x\, P_{\mathbf{v}}(x\|\text{spike})[\langle\mathbf{s}\|x,\text{spike}\rangle-\langle\mathbf{s}\|x\rangle]$ $\times\left[\frac{\mathrm{d}}{\mathrm{d}x}\frac{P_{\mathbf{v}}(x\|\text{spike})}{P_{\mathbf{v}}(x)}\right]$
Maximal value	$I_{\text{spike}}=\frac{1}{T}\int \mathrm{d}t\frac{r(t)}{\bar{r}}\log_2\frac{r(t)}{\bar{r}}$	$F_{\max}=\frac{1}{T}\int \mathrm{d}t\left[\frac{r(t)}{\bar{r}}\right]^2$
Multiple dimensions	$I[\mathbf{v}_1,\mathbf{v}_2]=\int \mathrm{d}x\, P_{\mathbf{v}_1,\mathbf{v}_2}(x_1,x_2\|\text{spike})$ $\times\log_2\frac{P_{\mathbf{v}_1,\mathbf{v}_2}(x_1,x_2\|\text{spike})}{P_{\mathbf{v}_1,\mathbf{v}_2}(x_1,x_2)}$	$F[\mathbf{v}_1,\mathbf{v}_2]=\int \mathrm{d}x\frac{[P_{\mathbf{v}_1,\mathbf{v}_2}(x_1,x_2\|\text{spike})]^2}{P_{\mathbf{v}_1,\mathbf{v}_2}(x_1,x_2)}$

parameters, i.e. coordinates of the relevant dimension, can be controlled by regularization (*via* performance on the test data set) to minimize the effects of overfitting. In contrast, regularization of the decorrelated STA adjusts only its overall smoothness [38, 46] or, very similarly, the value for the cutoff in the stimulus covariance matrix separating the well-sampled stimulus dimensions from the poorly sampled ones [7, 10–12, 39]. Because receptive fields of the linear model (decorrelated STA) computed with natural stimuli do not principally describe linear filters of the linear–nonlinear model, their regularization can offset effects due to insufficient data, but also can lead to artifacts in physiologically important quantities. For example, smoothness constraints directly affect estimation of frequency tuning. Therefore, it is not clear whether smoothness constraints can be used, for example, in studies designed to measure stimulus-dependent changes in frequency tuning. As Plate 2 illustrates, one-parameter regularization of decorrelated STA computed from naturalistic stimuli can lead to artifacts in frequency tuning. At the same time, numerical results shown in Plate 3 demonstrate that information and variance optimization outperform the one-parameter regularization of the decorrelated STA, even for the lowest number of spikes.

The above arguments describing the advantages of information and variance maximization strategies also apply to optimization strategies that are based on other divergence measures [23]. However, not all divergences might be equally good. For example, gradient-based algorithms could not be used to optimize Kolmogorov's measure of 'variation distance' $\int \mathrm{d}x|P(x|\text{spike})-P(x)|$, which belongs to the class of f-divergence measures [29, 30]. We have also seen that, in the asymptotic regime, the expected error in relevant dimensions found by maximizing information is less than or equal to that expected for variance maximization. Both information and variance maximization produce good matches between the true relevant dimensions and the reconstruction for a wide range of spike numbers, even for very small number of spikes, cf. Plate 3 and Figure 1. Even though the difference in performance between information and variance maximization strategies can be quite small, it was significant when simulation results for different number of spikes and neural noise level were combined ($p<10^{-4}$, paired t-test). One potential reason for this is the non-Gaussian character of stimulus correlations present in the natural ensemble used in these simulations. With non-Gaussian inputs, accounting for a certain percentage of information present in neural responses could be more appropriate than accounting for a certain percentage

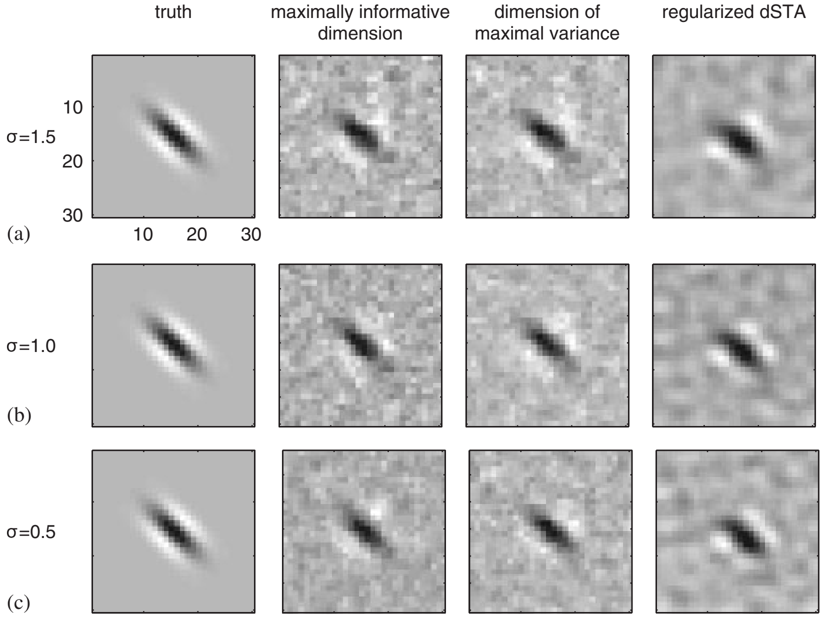

Figure 1. Comparison of performance for small numbers of spikes. Results are shown for three model cells with the same relevant dimension (leftmost column) and three different noise levels. From top to bottom, $\sigma = 1.5$, 1.0, and 0.5, measured in units of standard deviation of stimulus projections along the relevant dimension. Receptive fields shown correspond to the leftmost points of the curves labeled A, B, and C in Plate 3, and were computed with approximately 2100, 1200, and 500 spikes, respectively.

of variance. This makes the problem of finding relevant dimensions one of the few examples where information-theoretic measures are no more data limited than the variance-based measures. The key feature of the problem that makes this possible is its low-dimensional character: during optimization, information (variance) needs to be estimated only with respect to one dimension, or perhaps several dimensions, at a time.

SUMMARY

It is shown here that the least-square fitting of a linear–nonlinear model to a spike train can be transformed into an optimization problem for finding the relevant dimensions as those that account for most variance in the neural response. The resulting optimization strategy was proposed in [23], and has a structure similar to the optimization scheme of searching for maximally informative dimensions [22]. Both methods do not assume a particular form of the input–output function. Multiple dimensions can be found by iteratively adding relevant dimensions to the model and simultaneously optimizing the relevant dimensions. Either of the two methods can be used with stimulus ensembles of arbitrary statistical properties, as long as stimuli span the space of possible relevant dimensions. Thus, information and variance optimization can be used even with those ensembles that are strongly non-Gaussian, such as in the case of natural signals. The performance

DOI: 10.1002/sim

of both methods has been demonstrated here on model neurons responding to natural movies. With finite data, the two algorithms have similar convergence rates in the limit of large numbers of spikes (studied analytically) and at medium-to-small numbers of spikes (studied numerically). In the limit of large spike numbers, the expected error for relevant dimensions found by maximizing information was found analytically to be less than or equal to that with variance maximization. Slightly, but significantly, smaller errors associated with information maximization, compared variance optimization were also observed in numerical simulations with medium-to-small numbers of spikes. Both information and variance optimization strategies outperformed methods based on STA for nearly all values of spike numbers, with similar performance for the lowest spike numbers.

APPENDIX

A.1. Comparison of asymptotic errors for information and variance optimization

The error expected for information maximization from equations (38) and (39) is

$$\langle \delta \mathbf{v}^2 \rangle_{\text{information}} = \frac{\mathrm{Tr}'[A^{-1}]}{2N_{\text{spike}}}, \quad A_{ij} = \int \mathrm{d}x\, P(x|\text{spike}) C_{ij}(x) \left[\frac{g'(x)}{g(x)}\right]^2 \tag{A1}$$

and the gain function $g(x) = P(x|\text{spike})/P(x)$. Because covariance matrices $C_{ij}(x)$ are symmetric and positive definite, they can be represented as $C_{ij}(x) = \gamma_{ik}(x)\gamma_{jk}(x)$, where sum over k is implied (the exact expression of matrices γ will not be important). Then, matrix A can be written as an average over the probability distribution $P(x|\text{spike})$ of the product of matrix a and its transpose:

$$A = \int \mathrm{d}x\, P(x|\text{spike}) a(x) a^{\mathrm{T}}(x), \quad a_{ij}(x) = \gamma_{ij}(x) \frac{g'(x)}{g(x)} \tag{A2}$$

In the case of variance optimization, the expected error from equations (29)–(37) is

$$\langle \delta \mathbf{v}^2 \rangle_{\text{variance}} = \frac{\mathrm{Tr}'[DH^{-2}]}{2N_{\text{spike}}} \tag{A3}$$

where

$$D_{ij} = 4 \int \mathrm{d}x\, P(x|\text{spike}) C_{ij}(x) [g'(x)]^2, \quad H_{ij} = 2 \int \mathrm{d}x\, P(x|\text{spike}) C_{ij}(x) \frac{[g'(x)]^2}{g(x)} \tag{A4}$$

Similar to the matrix A in (A2), matrices H and D can also be written as an average over the same probability distribution $P(x|\text{spike})$:

$$D = \int \mathrm{d}x\, P(x|\text{spike}) b(x) b^{\mathrm{T}}(x), \quad H = \int \mathrm{d}x\, P(x|\text{spike}) a(x) b^{\mathrm{T}}(x), \quad b_{ij}(x) = 2\gamma_{ij}(x) g'(x) \tag{A5}$$

To summarize, the errors associated with maximizing information and variance are determined by three matrices, each of which can be written as an average:

$$A = \langle aa^{\mathrm{T}} \rangle_{P(x|\text{spike})}, \quad D = \langle bb^{\mathrm{T}} \rangle_{P(x|\text{spike})}, \quad H = \langle ab^{\mathrm{T}} \rangle_{P(x|\text{spike})} = \langle ba^{\mathrm{T}} \rangle_{P(x|\text{spike})} \tag{A6}$$

DOI: 10.1002/sim

In the rest of this section, $\langle\ldots\rangle$ will mean averaging with respect to the probability distribution $P(x|\text{spike})$, and we will omit the explicit reference to it.

The derivation will closely follow the logic of deriving the Cauchy–Schwarz inequality (as in Reference [47, p. 120]). We consider the following matrix:

$$M=\langle(\langle bb^{\mathrm{T}}\rangle a-\langle ab^{\mathrm{T}}\rangle b)(a^{\mathrm{T}}\langle bb^{\mathrm{T}}\rangle-b^{\mathrm{T}}\langle ba^{\mathrm{T}}\rangle)\rangle \tag{A7}$$

This matrix is, by construction, positive definite. Using expressions (A6), it is equal to

$$M=DAD-HHD-DHH-HDH \tag{A8}$$

To connect with expressions (A1) and (A3) for expected standard deviations in estimates of the relevant dimension, we consider a trace of the matrix $H^{-1}A^{-1}MD^{-1}$. This matrix is also positive definite, being a product of the positive-definite matrices:

$$\begin{aligned}\mathrm{Tr}(H^{-1}A^{-1}MD^{-1}H^{-1})&=\mathrm{Tr}(H^{-1}DH^{-1})-\mathrm{Tr}(A^{-1}D^{-1}H^{2}DH^{-2})\\&\quad-\mathrm{Tr}(A^{-1})+\mathrm{Tr}(A^{-1}D^{-1}HDH^{-1})\end{aligned} \tag{A9}$$

To verify that $\mathrm{Tr}(A^{-1}D^{-1}HDH^{-1})=\mathrm{Tr}(A^{-1})$, we introduce matrix $K=HD^{-1}$. Then, $\mathrm{Tr}(A^{-1}D^{-1}HDH^{-1})=\mathrm{Tr}(A^{-1}K^{\mathrm{T}}K^{-1})$, where we have used the fact that matrices H and D are symmetric. Because trace can be computed in any basis, we evaluate it in the basis where matrix K is diagonal. In this basis, K^{T} and K^{-1} will be diagonal as well, so that $\mathrm{Tr}(A^{-1}D^{-1}HDH^{-1})=\mathrm{Tr}(A^{-1})$. An identical argument can be used to show that $\mathrm{Tr}(A^{-1}D^{-1}H^{2}DH^{-2})=\mathrm{Tr}(A^{-1})$ (in this case $K=H^{2}D^{-1}$). Expression (A9) is therefore given by

$$\mathrm{Tr}(H^{-1}A^{-1}MD^{-1}H^{-1})=\mathrm{Tr}(H^{-1}DH^{-1})-\mathrm{Tr}(A^{-1})\geqslant 0 \tag{A10}$$

where we have used the fact that the trace of a positive-definite matrix $\mathrm{Tr}(H^{-1}A^{-1}MD^{-1}))\geqslant 0$. As a final note, we point out that the two traces in equation (A10) should be taken with respect to all directions in the stimulus space, except for the relevant dimension, so that $\mathrm{Tr}'(H^{-1}DH^{-1})\geqslant\mathrm{Tr}'(A^{-1})$. This is because we are computing deviations between the true relevant dimension and its estimates, which by definition can only be orthogonal to it. Mathematically, this is manifested by the zero eigenvalue that all three matrices A, D, and H exhibit when applied to the relevant dimension vector (all three matrices are linearly related to $C_{ij}(x)$, which is zero along the relevant dimension). Thus, the inverse of each of these three matrices is not even well defined in all of the stimulus space, only in the subspace orthogonal to the relevant dimension. This is precisely the subspace we are interested in when considering reconstruction errors for the relevant dimension. To summarize, in the asymptotic regime of large spike numbers, the expected error for relevant dimensions found by maximizing information is less than or equal to the expected error in relevant dimension found by maximizing the variance

$$\langle\delta\mathbf{v}^{2}\rangle_{\text{information}}\leqslant\langle\delta\mathbf{v}^{2}\rangle_{\text{variance}} \tag{A11}$$

A.2. Fisher information for estimating a single relevant dimension

Here, we compute the lower bound on the reconstruction error for a single relevant dimension. According to the Cramér–Rao inequality [35], the lowest variance in estimating relevant dimensions that can be achieved by any unbiased method is equal to a trace of the inverse of the Fisher information matrix:

$$\langle \delta \mathbf{v}^2 \rangle = \mathrm{Tr}'[I_F^{-1}] \tag{A12}$$

where the Fisher information matrix I_{F} depends on the likelihood $P(\text{spike}, \mathbf{s}|\mathbf{v})$ that the response—'spike'—co-occurred with a particular stimulus $\mathbf{s}$, given our model for the data. In this paper, a model for the data is specified by the relevant dimension $\mathbf{v}$:

$$I_{\mathrm{F}ij} = N \int \mathrm{d}\mathbf{s} P(\text{spike}, \mathbf{s}|\mathbf{v}) \partial_{v_i} [\ln P(\text{spike}, \mathbf{s}|\mathbf{v})] \partial_{v_j} [\ln P(\text{spike}, \mathbf{s}|\mathbf{v})] \tag{A13}$$

Here, N is the total number of stimulus presentations. This factor arises because Fisher information is additive for independent measurements [35]. To transform expression for the likelihood, we will first use the fact that the stimulus likelihood does not depend on relevant dimension: $P(\text{spike}, \mathbf{s}|\mathbf{v}) = P(\mathbf{s}) P(\text{spike}|\mathbf{s}, \mathbf{v})$. According to our models (1) and (2), the spike probability for a given stimulus depends only on its projection onto the relevant dimension. Therefore,

$$P(\text{spike}, \mathbf{s}|\mathbf{v}) = P(\mathbf{s}) P(\text{spike}|\mathbf{s} \cdot \mathbf{v}) = P(\mathbf{s}) P(\text{spike}) g(\mathbf{s} \cdot \mathbf{v}) \tag{A14}$$

We can now substitute this expression for the likelihood into the expression (A13) for the Fisher information matrix. Taking into account that the first two factors in (A14) do not depend on $\mathbf{v}$ and that $NP(\text{spike}) = N_{\text{spike}}$, we get

$$I_{\mathrm{F}ij} = N_{\text{spike}} \int \mathrm{d}\mathbf{s} P(\mathbf{s}) g(\mathbf{s} \cdot \mathbf{v}) \partial_{v_i} [\ln g(\mathbf{s} \cdot \mathbf{v})] \partial_{v_j} [\ln g(\mathbf{s} \cdot \mathbf{v})] \tag{A15}$$

Integrating over all stimulus components other than the component $x = \mathbf{s} \cdot \mathbf{v}$ along the relevant dimension, we get

$$I_{\mathrm{F}ij} = N_{\text{spike}} \int \mathrm{d}x \, P_{\mathbf{v}}(x|\text{spike}) \partial_{v_i} [\ln g(x)] \partial_{v_j} [\ln g(x)] \tag{A16}$$

Using expressions (7) and (8) for the probability distributions $P_{\mathbf{v}}(x|\text{spike})$ and $P_{\mathbf{v}}(x)$, we find that the gradient of $\ln g(x) = \ln[P_{\mathbf{v}}(x|\text{spike})/P_{\mathbf{v}}(x)]$ is given by

$$\begin{aligned} \partial_{v_i}[\ln g(x)] &= \frac{\partial_{v_i} P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x|\text{spike})} - \frac{\partial_{v_i} P_{\mathbf{v}}(x)}{P_{\mathbf{v}}(x)} \\ &= \frac{\frac{\mathrm{d}}{\mathrm{d}x}[\langle s_i|x, \text{spike}\rangle P(x|\text{spike})]}{P_{\mathbf{v}}(x|\text{spike})} - \frac{\frac{\mathrm{d}}{\mathrm{d}x}[\langle s_i|x\rangle P(x)]}{P_{\mathbf{v}}(x)} \end{aligned} \tag{A17}$$

Fisher information matrix is evaluated at the maximum of the likelihood function at the true relevant dimension $\mathbf{v} = \hat{e}_1$. Stimulus projections along the true relevant dimensions provide a sufficient statistic for the 'spike' response, so that $\langle s_i|s_1, \text{spike}\rangle = \langle s_i|s_1\rangle$ (this can also be verified directly, using $P(\mathbf{s}|\text{spike})/P(\mathbf{s}) = P(s_1|\text{spike})/P(s_1)$). Therefore,

$$\partial_{v_i}[\ln g(x)] = \langle s_i|x\rangle \left[\frac{\mathrm{d}}{\mathrm{d}x} \ln \frac{P(x|\text{spike})}{P(x)}\right], \quad x \equiv s_1 \tag{A18}$$

DOI: 10.1002/sim

Putting it all together, we find that the Fisher information matrix is given by

$$I_{\mathrm{F}ij} = N_{\mathrm{spike}} \int \mathrm{d}x\, P(x|\mathrm{spike}) \langle s_i|x\rangle \langle s_j|x\rangle \left[\frac{\mathrm{d}}{\mathrm{d}x} \ln \frac{P(x|\mathrm{spike})}{P(x)}\right]^2 \tag{A19}$$

By comparison, the variance of relevant dimensions computed by maximizing information from (39) is inversely proportional to a trace of matrix A:

$$A_{ij} = N_{\mathrm{spike}} \int \mathrm{d}x\, P(x|\mathrm{spike}) (\langle s_i s_j|x\rangle - \langle s_i|x\rangle \langle s_j|x\rangle) \left[\frac{\mathrm{d}}{\mathrm{d}x} \ln \frac{P(x|\mathrm{spike})}{P(x)}\right]^2 \tag{A20}$$

According to the Cramér–Rao bound, $\mathrm{Tr}'[I_F^{-1}] \leqslant \mathrm{Tr}'[A^{-1}]$. Matrices I_F and A have very similar structures, and it might be possible that the two traces, computed in the subspace orthogonal to the relevant direction, are equal to each other for some types of stimulus distributions and input–output nonlinearities. In those situations, maximizing information provides the lowest possible reconstruction error. It remains to be determined whether maximizing objective functions based on other f-divergences [23] could saturate the Cramér–Rao bound.

ACKNOWLEDGEMENTS

The author is indebted to W. Bialek, A. J. Doupe, K. D. Miller, and M. P. Stryker for many useful discussions. This work was supported by an NIMH mentored quantitative research career award 5K25MH068094. Numerical calculations were carried out using Teragrid supercomputer resources at the San Diego, Texas, Pittsburgh, and University of Chicago Supercomputing Centers.

REFERENCES

1. 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. *Proceedings of the Royal Society of London Series B-Biological Sciences* 1988; **265**:259–265.
2. Brenner N, Bialek W, de Ruyter van Steveninck RR. Adaptive rescaling maximizes information transmission. *Neuron* 2000; **26**:695–702.
3. Schwartz O, Chichilnisky EJ, Simoncelli E. Characterizing neural gain control using spike-triggered covariance. In *Advances in Neural Information Processing*, vol. 14, Dietterich TG, Becker S, Ghahramani Z (eds). MIT Press, 2002.
4. Touryan J, Lau B, Dan Y. Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience* 2002; **22**:10811–10818.
5. Bialek W, de Ruyter van Steveninck RR. Features and dimensions: motion estimation in fly vision. q-bio/0505003, 2005.
6. Rust NC, Schwartz O, Movshon JA, Simoncelli EP. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* 2005; **46**:945–956.
7. Felsen G, Touryan J, Han F, Dan Y. Cortical sensitivity to visual features in natural scenes. *PLoS Biology* 2005; **3**:1819–1828.
8. Fairhall AL, Burlingame CA, Narasimhan R, Harris RA, Puchalla JL, Berry MJ 2nd. *Journal of Neurophysiology* 2006; **96**(5):2724–2738.
9. de Boer E, Kuyper P. Triggered correlation. *IEEE Transactions on Biomedical Engineering* 1968; **15**:169–179.
10. Theunissen FE, Sen K, Doupe AJ. Spectral–temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience* 2000; **20**:2315–2331.
11. Sen K, Theunissen FE, Doupe AJ. Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of Neurophysiology* 2001; **86**:1445–1458.
12. 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.

13. Escabi MA, Schreiner CE. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *Journal of Neuroscience* 2002; **22**:4114–4131.
14. Slee SJ, Higgs MH, Fairhall AL, Spain WJ. Two-dimensional time coding in the auditory brainstem. *Journal of Neuroscience* 2005; **25**:9978–9988.
15. Kearney RE, Hunter IW. System identification of human joint dynamics. *Critical Review in Biomedical Engineering* 1990; **18**:55–87.
16. Paninski L, Shoham S, Fellows MR, Hatsopoulos NG, Donoghue JP. Superlinear population encoding of dynamic hand trajectory in primar motor cortex. *Journal of Neuroscience* 2004; **24**:8551–8561.
17. Hunter IW, Korenberg MJ. The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. *Biological Cybernetics* 1986; **55**:135–144.
18. Rieke F, Warland D, de Ruyter van Steveninck RR, Bialek W. *Spikes: Exploring the Neural Code*. MIT Press: Cambridge, MA, 1997.
19. Chichilnisky EJ. A simple white noise analysis of neuronal light responses. *Network—Computation in Neural Systems* 2001; **12**:199–213.
20. Agüera y Arcas B, Fairhall A, Bialek W. Computation in a single neuron: Hodgkin and Huxley revisited. *Neural Computation* 2003; **15**:1715–1749; See also physics/0212113.
21. 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.
22. Sharpee T, Rust NC, Bialek W. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Computation* 2004; **16**:223–250; See also physics/0212110, and a preliminary account in *Advances in Neural Information Processing*, vol. 15, Becker S, Thrun S, Obermayer K (eds). MIT Press: Cambridge, MA, 2003; 261–268.
23. Paninski L. Convergence properties of three spike-triggered average techniques. *Network—Computation in Neural Systems* 2003; **14**:437–464.
24. Schwartz O, Pillow JW, Rust NC, Simoncelli EP. Spike-triggered neural characterization. *Journal of Vision* 2006; **6**(4):484–507.
25. Bussgang JJ. Crosscorrelation functions of amplitude distorted Gaussian signals. In *Research Laboratory of Electronics, M.I.T. Technical Report*, vol. 216, 1952.
26. Price R. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory* 1958; **4**:69–72.
27. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. *Annual Review of Neuroscience* 2001; **24**:1193–1216.
28. Escabi MA, Miller LM, Read HL, Schreiner CE. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *Journal of Neuroscience* 2003; **23**:11489–11504.
29. Ali SM, Silvey SD. A general class of coefficient of divergence of one distribution from another. *Journal of Royal Statistical Society Series B* 1966; **28**:131–142.
30. Csiszár I. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* 1967; **2**:299–318.
31. Rényi R. On measures of entropy and information. In *Proceedings of Fourth Berkeley Symposium Mathematical Statistics Probability*, vol. 1, Berkeley, California, Neyman J (ed.), 1961; 547–561.
32. Weisberg S, Welsh AH. Adapting for the missing link. *Annals of Statistics* 1994; **22**:1674–1700.
33. Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer: New York, 1998.
34. Korenberg MJ, Hunter IW. Two methods for identifying Wiener cascades having noninvertible static nonlinearities. *Annals of Biomedical Engineering* 1999; **27**:793–804.
35. Cover TM, Thomas JA. *Information Theory*. Wiley: New York, 1991.
36. Adelman TL, Bialek W, Olberg RM. The information content of receptive fields. *Neuron* 2003; **40**:823–833.
37. 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.
38. Smyth D, Willmore B, Baker GE, Thompson ID, Tolhurst DJ. The receptive fields organization of simple cells in the primary visual cortex of ferrets under natural scene stimulation. *Journal of Neuroscience* 2003; **23**:4746–4759.
39. Woolley SMN, Gill PR, Theunissen FE. Stimulus-dependent auditory tuning results in synchronous population coding of vocalization in the songbird midbrain. *Journal of Neuroscience* 2006; **26**:2499–2512.
40. Ringach DL, Sapiro G, Shapley R. A subspace reverse-correlation technique for the study of visual neurons. *Vision Research* 1997; **37**:2455–2464.

41. 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.
42. van der Vaart AW. *Asymptotic Statistics*. Cambridge University Press: Cambridge, 1998.
43. Arfken GB, Weber HJ. *Mathematical Methods for Physicists*. Elsevier: Amsterdam, 1970.
44. Marmarelis VZ. Modeling methodology for nonlinear physiological systems. *Annals of Biomedical Engineering* 1997; **25**:239–251.
45. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press: Cambridge, 1992.
46. Sahani M, Linden JF. Evidence optimization techniques for estimating stimulus-response functions. In *Advances in Neural Information Processing Systems*, vol. 15, Thrun S, Becker S, Obermayer K (eds). MIT Press: Cambridge, MA, 2003; 301–308.
47. Dayan P, Abbott LF. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press: Cambridge, MA, 2001.

DOI: 10.1002/sim