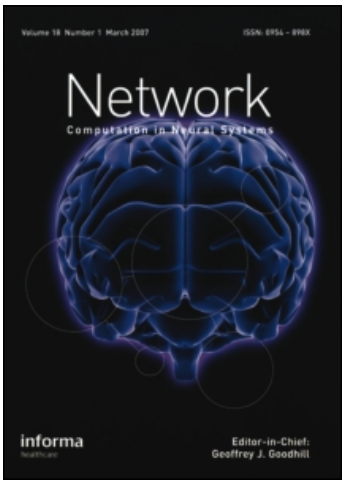

Network: Computation in Neural Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713663148

Estimating linear-nonlinear models using Rényi divergences

Minjoon Kouh [ab]; Tatyana O. Sharpee [ab]
[a] The Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA [b] The Center for Theoretical Biological Physics, University of California, San Diego, La Jolla, CA, USA

Online Publication Date: 01 June 2009

To cite this Article Kouh, Minjoon and Sharpee, Tatyana O.(2009)'Estimating linear-nonlinear models using Rényi divergences',Network: Computation in Neural Systems,20:2,49 — 68
To link to this Article: DOI: 10.1080/09548980902950891
URL: http://dx.doi.org/10.1080/09548980902950891

Network: Computation in Neural Systems
April–June 2009; 20(2): 49–68

informa
healthcare

Estimating linear–nonlinear models using Rényi divergences

MINJOON KOUH[1,2] & TATYANA O. SHARPEE[1,2]

[1]*The Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92093, and* [2]*The Center for Theoretical Biological Physics, University of California, San Diego, La Jolla, CA, USA*

(Received 12 February 2009; revised 6 April 2009; accepted 7 April 2009)

Abstract
This article compares a family of methods for characterizing neural feature selectivity using natural stimuli in the framework of the linear–nonlinear model. In this model, the spike probability depends in a nonlinear way on a small number of stimulus dimensions. The relevant stimulus dimensions can be found by optimizing a Rényi divergence that quantifies a change in the stimulus distribution associated with the arrival of single spikes. Generally, good reconstructions can be obtained based on optimization of Rényi divergence of any order, even in the limit of small numbers of spikes. However, the smallest error is obtained when the Rényi divergence of order 1 is optimized. This type of optimization is equivalent to information maximization, and is shown to saturate the Cramér–Rao bound describing the smallest error allowed for any unbiased method. We also discuss conditions under which information maximization provides a convenient way to perform maximum likelihood estimation of linear–nonlinear models from neural data.

Keywords: *information theory, natural scenes, single neuron computation, visual system*

Introduction

The application of system identification techniques to the study of sensory neural systems has a long history. One family of approaches employs the dimensionality reduction idea: while inputs are typically very high-dimensional, not all dimensions are equally important for eliciting a neural response (de Boer and Kuyper 1968; Hunter and Korenberg 1986; de Ruyter van Steveninck and Bialek 1988; Weisberg and Welsh 1994; Marmarelis 1997; Bialek and de Ruyter van

Correspondence: T. Sharpee, The Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, 92037 United State. E-mail: sharpee@salk.edu

ISSN 0954-898X print/ISSN 1361-6536 online/09/020049–68 © 2009 Informa Healthcare Ltd.
DOI: 10.1080/09548980902950891

Steveninck 2005). The aim is then to find a small set of K dimensions $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K\}$ in the D-dimensional stimulus space that are relevant for neural response, without imposing a particular functional dependence between the neural response and the stimulus components $\{s_1, s_2, \ldots, s_K\}$ along the relevant dimensions:

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

Here, the nonlinear gain function $g(s_1, s_2, \ldots, s_K)$ describes how the firing rate is modulated by stimulus components along K relevant dimensions compared to the average firing rate across all stimuli, which is given by $P(\text{spike})$. If the inputs are Gaussian, the last requirement is not important, because the relevant dimensions can be found without knowing a correct functional form for the nonlinear function g in Equation (1). However, for non-Gaussian inputs a wrong assumption for the form of the nonlinearity g will lead to systematic errors in the estimate of the relevant dimensions themselves (Hunter and Korenberg 1986; Ringach et al. 1997; Paninski 2003; Sharpee et al. 2004; Christianson et al. 2008). The larger the deviations of the stimulus distribution from a Gaussian, the larger will be the effect of errors in the presumed form of the nonlinearity function g on estimating the relevant dimensions. Because inputs derived from a natural environment, either visual or auditory, have been shown to be strongly non-Gaussian (Ruderman and Bialek 1994; Simoncelli and Olshausen 2001), we will concentrate here on system identification methods suitable for either Gaussian or non-Gaussian stimuli.

The problem of finding relevant dimensions for neural responses is illustrated in Figure 1. Here, a stimulus is represented as a point in a two-dimensional subspace of the high-dimensional input space. Each dimension may correspond, for example, to the luminance of a pixel of an image. In this example, the probability of a spike depends on stimulus components s_1, but not s_2. These dimensions can be distinguished by comparing the *a priori* probability distribution of all of the presented stimuli $P(s_1)$ or $P(s_2)$ along these two dimensions, with the conditional probability distribution of stimuli that elicited spikes: $P(s_1|\text{spike})$ and $P(s_2|\text{spike})$, respectively. When compared along the irrelevant stimulus dimension, the distributions $P(s_2|\text{spike})$ and $P(s_2)$ are similar, because the spikes would have occurred with equal probability for all values of s_2. On the other hand, the *a priori* and conditional probability distributions will be different along the relevant dimension s_1. Therefore, by finding the stimulus dimensions that give the most dissimilar distributions, one can find the relevant dimensions for neural responses. Mathematically, the dissimilarity between two probability distributions can be quantified by any one of the number of divergence measures. One family of such divergence measures are called Rényi divergences, and are defined for a pair of probability distributions P and Q as (Rényi 1961)

$$D^{(\alpha)}(P\|Q) = \frac{1}{\alpha - 1}\int \mathrm{d}x\, Q(x)\left[\frac{P(x)}{Q(x)}\right]^{\alpha}. \tag{2}$$

These divergences can be used as objective functions for an optimization search, and the result would yield the relevant stimulus dimensions.

In one of the earlier works of finding the relevant dimensions for neural responses probed with non–Gaussian inputs, Hunter and Korenberg (1986) proposed an iterative scheme where the relevant dimensions are first found by assuming that the

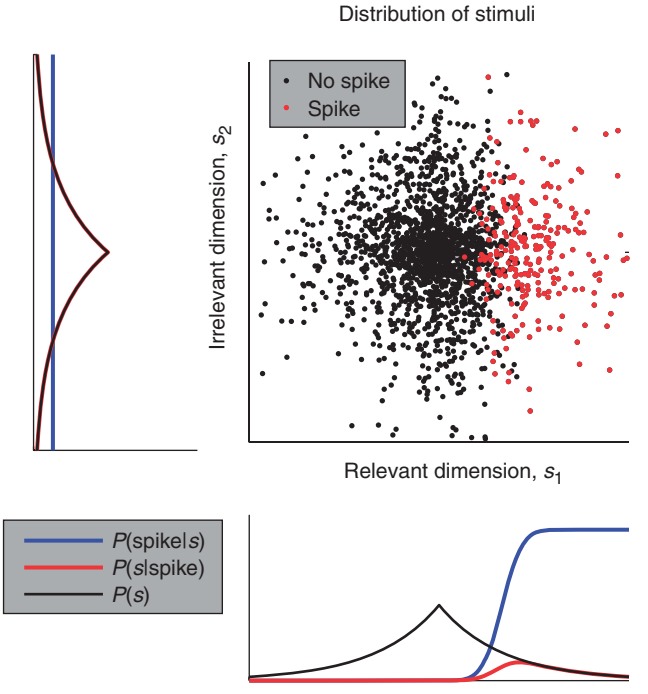

Figure 1. Each point in the plot represents a stimulus from a non-Gaussian distribution from a high-dimensional space (a two-dimensional space in this example). Some stimuli may elicit spikes (red), and others may not (black). Because the vertical dimension (s_2) is not relevant to, or correlated with, spikes, the probability distribution of stimuli $P(s_2)$ is similar to the distribution of stimuli given a spike $P(s_2|\text{spike})$. On the other hand, the horizontal dimension (s_1) can explain the spiking behavior because the spikes are observed whenever the stimulus component s_1 is large. Therefore, $P(s_1)$ and $P(s_1|\text{spike})$ are dissimilar. The ratio $P(s_1|\text{spike})/P(s_1)$ gives the nonlinearity $P(\text{spike}|s_1)$.

input–output function g is linear. The input–output function is subsequently updated based on the current estimate of the relevant dimensions. The inverse of g is in turn used to improve the estimate of the relevant dimensions. Optimization of Rényi divergences offers a way to improve the iterative procedure of Hunter and Korenberg by formulating optimization problem exclusively with respect to relevant dimensions. The improvement is due to the fact that the nonlinear function g is taken into account in the computation of a Rényi divergence. This eliminates the need for inverting a nonlinear function and for estimating both relevant dimensions and nonlinearity g iteratively.

In the context of finding the most consistent linear–nonlinear model, optimizing certain Rényi divergences corresponds to some classical objective functions. For example, optimizing the Rényi divergence of order 2 corresponds to least square fitting of the linear–nonlinear model to the data (Sharpee 2007). The Kullback–Leibler divergence also belongs to this family, and represents the Rényi divergence of order 1. The Kullback–Leibler divergence between the *a priori* and the conditional probability distributions as described above corresponds to computing the mutual information between the neural response and the stimulus components

along the relevant dimension (Sharpee et al. 2004). The optimization scheme based on maximizing information has been previously proposed and implemented on model (Sharpee et al. 2004) and real cells from the primary visual cortex (Sharpee et al., 2006).

Here, we ask which optimization strategy is the best for discovering relevant dimensions for neural responses with natural stimuli. A comparison can be made by finding the errors between the true relevant dimensions and the estimates obtained with Rényi divergences of different orders. We analytically derive asymptotic errors and show that relevant dimensions found by maximizing Kullback–Leibler divergence have the smallest error in the limit of large spike numbers compared to maximizing any other Rényi divergence, including the one that implements the least squares method.

Maximizing Rényi divergence of order 2 as a way to minimize least squares

One way of selecting a low-dimensional model of neural response is to minimize a χ^2-difference between spike probabilities measured and predicted by the model after averaging across all inputs $\mathbf{s}$:

$$\chi^2[\mathbf{v}] = \int d\mathbf{s} P(\mathbf{s})[P(\text{spike}|\mathbf{s}) - P(\text{spike}|\mathbf{s}\cdot\mathbf{v})]^2/[P(\text{spike})]^2, \tag{3}$$

where dimension $\mathbf{v}$ is the relevant dimension of the model described in Equation 1 (multiple dimensions could also be used, see further). We have introduced a constant normalization by $[P(\text{spike})]^2$ in order to use the Bayes' rule, which after some re-arrangement of terms yields:

$$\begin{aligned}\chi^2[\mathbf{v}] = \int d\mathbf{s} P(\mathbf{s})\left[\frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})} - \frac{P(\mathbf{s}\cdot\mathbf{v}|\text{spike})}{P(\mathbf{s}\cdot\mathbf{v})}\right]^2 = \int d\mathbf{s}\frac{[P(\mathbf{s}|\text{spike})]^2}{P(\mathbf{s})} \\ - 2\int d\mathbf{s} P(\mathbf{s}|\text{spike})\frac{P(\mathbf{s}\cdot\mathbf{v}|\text{spike})}{P(\mathbf{s}\cdot\mathbf{v})} + \int d\mathbf{s} P(\mathbf{s})\left[\frac{P(\mathbf{s}\cdot\mathbf{v}|\text{spike})}{P(\mathbf{s}\cdot\mathbf{v})}\right]^2.\end{aligned} \tag{4}$$

In the last two integrals, we can carry out integration with respect to all stimulus components $\mathbf{s}_\perp$ that are orthogonal to the relevant direction $\mathbf{v}$. This partial integration yields (integration over $\mathbf{s}\cdot\mathbf{v}$ remains) $\int d\mathbf{s}_\perp P(\mathbf{s}|\text{spike}) = \int d\mathbf{s}_\perp P(\mathbf{s}_\perp, \mathbf{s}\cdot\mathbf{v}|\text{spike}) = P(\mathbf{s}\cdot\mathbf{v}|\text{spike})$, and similarly $\int d\mathbf{s}_\perp P(\mathbf{s}) = P(\mathbf{s}\cdot\mathbf{v})$. Combining these results with Equation (4), and introducing notation $x = \mathbf{s}\cdot\mathbf{v}$ for the remaining integration variable, we find that the χ^2-difference can be written as:

$$\chi^2[\mathbf{v}] = \int d\mathbf{s}\frac{[P(\mathbf{s}|\text{spike})]^2}{P(\mathbf{s})} - \int dx\,\frac{[P_\mathbf{v}(x|\text{spike})]^2}{P_\mathbf{v}(x)}. \tag{5}$$

Here, the subscript in $P_\mathbf{v}$ denotes the dimension $\mathbf{v}$ along which the probability distribution was computed. We note that this result is valid for an arbitrary stimulus distribution, and even for cells that cannot be perfectly described by the linear–nonlinear model. The probability distributions $P_\mathbf{v}(x)$ and $P_\mathbf{v}(x|\text{spike})$ are obtained by averaging across all presented stimuli and across all stimuli that lead to a spike, respectively.

If neural spikes are indeed based on one relevant dimension, then this dimension will explain all of the variance, leading to $\chi^2 = 0$. For all other dimensions $\mathbf{v}$, $\chi^2[\mathbf{v}] > 0$. Based on Equation (5), in order to minimize χ^2 we need to maximize

$$F[\mathbf{v}] = \int \mathrm{d}x\, P_{\mathbf{v}}(x)\left[\frac{P_{\mathbf{v}}(x|\mathrm{spike})}{P_{\mathbf{v}}(x)}\right]^2, \tag{6}$$

which is a Rényi divergence of order 2 between probability distribution $P_{\mathbf{v}}(x|\mathrm{spike})$ and $P_{\mathbf{v}}(x)$, and is part of a family of f-divergence measures that are based on a convex function of the ratio of two probability distributions (Ali and Silvey 1966; Csiszár 1967; Paninski 2003). For optimization strategy based on Rényi divergences of order α, the relevant dimensions are found by maximizing, (cf. Equation 2):

$$F^{(\alpha)}[\mathbf{v}] = \frac{1}{\alpha - 1}\int \mathrm{d}x\, P_{\mathbf{v}}(x)\left[\frac{P_{\mathbf{v}}(x|\mathrm{spike})}{P_{\mathbf{v}}(x)}\right]^{\alpha}. \tag{7}$$

By comparison, when the relevant dimension(s) are found by maximizing information (Sharpee et al. 2004), one would optimize the Kullback–Leibler divergence, which can be obtained by taking a formal limit $\alpha \to 1$ in Equation 7:

$$I[\mathbf{v}] = \int \mathrm{d}x\, P_{\mathbf{v}}(x|\mathrm{spike}) \ln \frac{P_{\mathbf{v}}(x|\mathrm{spike})}{P_{\mathbf{v}}(x)}. \tag{8}$$

Returning to the χ^2-minimization, the maximal value for $F[\mathbf{v}]$ that can be achieved by any dimension $\mathbf{v}$ is:

$$F_{\max} = \int \mathrm{d}\mathbf{s} \frac{[P(\mathbf{s}|\mathrm{spike})]^2}{P(\mathbf{s})}. \tag{9}$$

This quantity is closely related to the firing rate averaged across different inputs for the following reason. One notices that computation of the mutual information carried by individual spikes about the stimulus relies on similar integrals. Following the procedure outlined for computing mutual information (Brenner et al. 2000), one can use the Bayes' rule and the ergodic assumption to compute $F_{\max}$ as a time-average:

$$F_{\max} = \frac{1}{T}\int \mathrm{d}t \left[\frac{r(t)}{\bar{r}}\right]^2, \tag{10}$$

where the firing rate $r(t) = P(\mathrm{spike}|\mathbf{s})/\Delta t$ is measured in time bins of width Δt using multiple repetitions of the same stimulus sequence. The stimulus ensemble should be diverse enough to justify the ergodic assumption (this could be checked by computing $F_{\max}$ for increasing fractions of the overall dataset size). The average firing rate $\bar{r} = P(\mathrm{spike})/\Delta t$ is obtained by averaging $r(t)$ in time.

Equation 10 demonstrates that the variance of the firing rate across different inputs can be simply obtained from $F_{\max}$ as $\bar{r}^2(F_{\max} - 1)$. In turn, $F[\mathbf{v}]$ can be used to compute the variance in the firing rate accounted for by the linear–nonlinear model based on the relevant stimulus dimension $\mathbf{v}$. This is because in computing $F[\mathbf{v}]$ we have grouped together all stimuli that have the same projection value on the dimension $\mathbf{v}$. Correspondingly, $F[\mathbf{v}] \leq F_{\max}$, which can be seen either from the fact that $\chi^2[\mathbf{v}] \geq 0$, or from the data processing inequality, which applies not only to

Kullback–Leibler divergence, but also to Rényi divergences (Ali and Silvey 1966; Csiszár 1967; Paninski 2003). Computing the ratio

$$(F[\mathbf{v}] - 1)/(F_{\max} - 1)$$

offers a simple way to find the percentage of variance that can be accounted for by a stimulus dimension $\mathbf{v}$ and an arbitrary nonlinear gain function.

Optimization schemes based on Rényi divergences of different orders have very similar structure. In particular, gradient could be evaluated in a similar way:

$$\nabla_{\mathbf{v}} F^{(\alpha)} = \frac{\alpha}{\alpha - 1} \int \mathrm{d}x\, P_{\mathbf{v}}(x|\mathrm{spike})[\langle \mathbf{s}|x, \mathrm{spike}\rangle - \langle \mathbf{s}|x\rangle] \cdot \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{P_{\mathbf{v}}(x|\mathrm{spike})}{P_{\mathbf{v}}(x)}\right]^{\alpha-1}, \qquad (11)$$

where $\langle \mathbf{s}|x,\ \mathrm{spike}\rangle = \int \mathrm{d}\mathbf{s}\, \mathbf{s}\delta(x - \mathbf{s}\cdot\mathbf{v})P(\mathbf{s}|\mathrm{spike})/P(x|\mathrm{spike})$, and similarly for $\langle \mathbf{s}|x\rangle$. The gradient is thus given by a weighted sum of spike-triggered averages (STA) $\langle \mathbf{s}|x, \mathrm{spike}\rangle - \langle \mathbf{s}|x\rangle$ that are conditional upon a particular projection value x of stimuli onto the dimension $\mathbf{v}$ for which the gradient is being evaluated. The similarity of the structure of both the objective functions and their gradients suggests that the same numerical algorithms can be used for optimizing the Rényi divergences of different orders. Examples of possible algorithms have been described (Paninski 2003; Sharpee et al. 2004, 2006) and include a combination of gradient ascent and simulated annealing.

Here are a few facts common to this family of optimization schemes. First, as proven in the case of information maximization based on the Kullback–Leibler divergence (Sharpee et al. 2004), the objective function $F^{(\alpha)}[\mathbf{v}]$ does not change with the length of the vector $\mathbf{v}$. Therefore $\mathbf{v} \cdot \nabla_{\mathbf{v}} F = 0$, which can also be seen directly from Equation 11, because $\mathbf{v} \cdot \langle \mathbf{s}|x, \mathrm{spike}\rangle = x$ and $\mathbf{v} \cdot \langle \mathbf{s}|x\rangle = x$. Second, the gradient is 0 when evaluated along the true receptive field. This is because projection onto the true relevant dimension completely determines the spike probability, so that $\langle \mathbf{s}|s_1, \mathrm{spike}\rangle = \langle \mathbf{s}|s_1\rangle$. The fact that the gradients are zero for the true receptive field $\hat{e}_1$ agrees with the earlier statement that $\mathbf{v} = \hat{e}_1$ maximizes variance $\bar{r}^2(F[\mathbf{v}] - 1)$, and more generally $F^{(\alpha)}[\mathbf{v}]$. Third, objective functions, including the variance explained and information, can be computed with respect to multiple dimensions by keeping track of stimulus projections on all the relevant dimensions when forming the probability distributions. For example, in the case of two relevant dimensions $\mathbf{v}_1$ and $\mathbf{v}_2$, one would use $P_{\mathbf{v}_1,\mathbf{v}_2}(x_1, x_2|\mathrm{spike})$ and $P_{\mathbf{v}_1,\mathbf{v}_2}(x_1, x_2)$ to compute the variance explained by the linear–nonlinear model as $\bar{r}^2(F[\mathbf{v}_1, \mathbf{v}_2] - 1)$, where

$$F[\mathbf{v}_1, \mathbf{v}_2] = \int \mathrm{d}x_1 \mathrm{d}x_2 [P(x_1, x_2|\mathrm{spike})]^2 / P(x_1, x_2). \qquad (12)$$

If multiple stimulus dimensions are relevant for eliciting the neural response, they can always be found (provided sufficient number of responses have been recorded) by optimizing the variance according to Equation 12 with the correct number of dimensions. In practice, this involves finding a single relevant dimension first, and then iteratively increasing the number of relevant dimensions considered while adjusting the previously found relevant dimensions. The amount by which relevant dimensions need to be adjusted is proportional to the contribution of subsequent relevant dimensions to neural spiking (the corresponding expression has the same functional form as that for relevant dimensions found by maximizing information, (cf. Appendix B; Sharpee et al. 2004)). If stimuli are Gaussian (either uncorrelated

or correlated), then the previously found dimensions do not need to be adjusted when additional dimensions are introduced (Sharpee 2007). In this case, all the relevant dimensions can be found one by one, by searching only for one relevant dimension at a time in the stimulus subspace orthogonal to the already recovered relevant dimensions.

Illustration for a model simple cell

Here we illustrate using numerical simulations how relevant dimensions can be found by maximizing Rényi divergences of various orders (1, 2 and 3), and compare these schemes with the one based on computing the STA (de Boer and Kuyper 1968; Rieke et al. 1997). Our goal is to reconstruct relevant dimensions of neurons probed with inputs of arbitrary statistics. We used stimuli derived from a natural visual environment (Sharpee et al. 2006) that are known to strongly deviate from a Gaussian distribution. The analyses have been carried out with respect to model neurons, for which the relevant dimensions are known. The example model neuron is taken to mimic properties of simple cells found in the primary visual cortex. It has a single relevant dimension shown in Figure 2(a), which is phase and orientation sensitive; we denote this relevant dimension as $\hat{e}_1$. In this model, 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: $P(\text{spike}|\mathbf{s})/P(\text{spike}) \equiv g(s_1) = \langle H(s_1 - \theta + \xi) \rangle$, 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.

Figure 2 shows that it is possible to obtain a good estimate of the relevant dimension $\hat{e}_1$ by maximizing either information, variance, or a higher-order Rényi divergence ($\alpha = 3$), as shown in panels (b)–(d), respectively. The final value of the projection depends on the size of the dataset, as will be discussed further. In the example shown in Figure 2 there were approximately 50,000 spikes with average probability of spike ≈ 0.05 per frame, and the reconstructed vector has a projection $\hat{v}_{max} \cdot \hat{e}_1 = 0.97$ when maximizing information, variance, or Rényi divergence of order 3.

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 one of the Rényi divergences, and taking their ratio. Because of the Bayes' rule, this yields the nonlinear input–output function g of Equation (1). In Figure 2(f), the spike probability of the reconstructed neuron $P(\text{spike}|\mathbf{s} \cdot \hat{v}_{\max})$ (crosses) is compared with the probability $P(\text{spike}|s_1)$ used in the model (solid line). A good match is obtained. For comparison, we show in panel (e) the reconstruction result obtained using the decorrelated STA method with regularization (for a review see David and Gallant (2005)). According to this method, one first computes the STA, and then multiplies the STA by the pseudo-inverse of the stimulus covariance matrix (only well sampled stimulus dimensions contribute to the pseudo-inverse; the cutoff between ''well-sampled'' and ''poorly-sampled'' dimensions is determined as one producing the best predictive power on a novel part of the data). In this case, regularization was performed by setting aside one-fourth of the data as a test dataset, and choosing

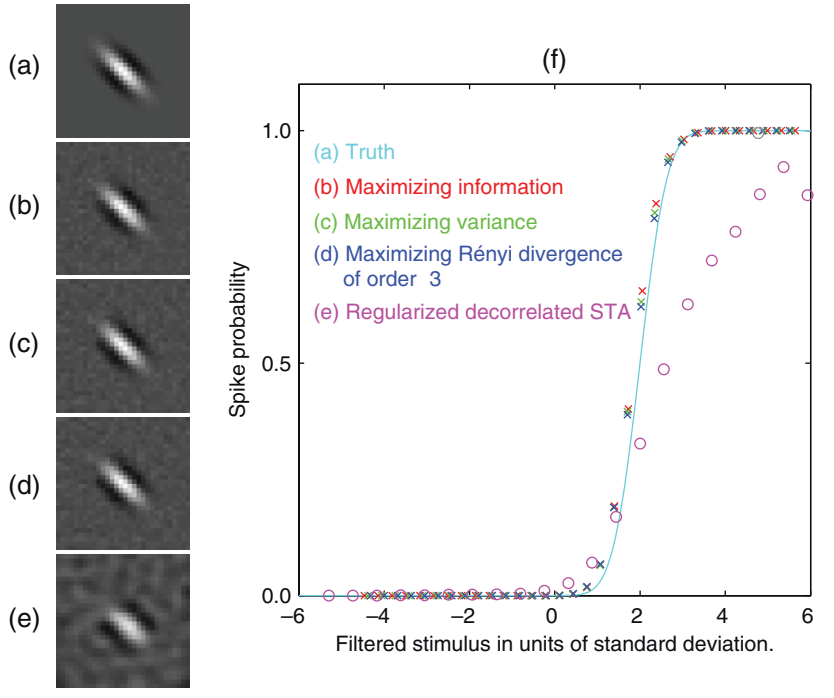

Figure 2. Analysis of a model visual neuron with one relevant dimension shown in (a). Panels (b), (c), and (d) show normalized vectors $\hat{v}_{\max}$ found by maximizing information, variance, and the Rényi divergence of order 3, respectively. Panel (e) shows the relevant dimension computed according to the reverse correlation method as regularized decorrelated STA (see text for details). In (f), the probability of a spike $P(\text{spike}|\mathbf{s}\cdot\hat{v}_{max})$ (red, green, and blue crosses for maximizing information, variance, and Rényi divergence of order 3, respectively) is compared to $P(\text{spike}|s_1)$ used in generating spikes (cyan solid line). The spike probabilities for regularized decorrelated STA are also shown (magenta circles). Parameters of the model are $\sigma = 0.5$ and $\theta = 2$, both given in units of standard deviation of s_1, which is also the unit for the x-axis in (f).

a cutoff on the eigenvalues of the input covariances matrix that would give the maximal information value on the test dataset (Theunissen et al. 2001; Felsen et al. 2005). This method also yielded a good reconstruction with a projection value of 0.8. In some cases, however, the regularized decorrelated STA can have systematic deviations from the true relevant dimensions (Ringach et al. 1997, 2002; Sharpee et al. 2004; Sharpee et al. 2006).

Asymptotic errors for large numbers of spikes

In the limit of infinite data the relevant dimensions can be found by maximizing variance, information, higher-order Rényi divergences, or in principle other objective functions (Paninski 2003). In a real experiment, with a dataset of finite size, the optimal vector found by maximizing information or variance $\hat{v}$ will deviate from the true relevant dimension $\hat{e}_1$. In this section, we compare the rates of

convergence of optimization strategies based on Rényi divergences of various orders as the dataset size increases and/or neural noise decreases.

The reconstruction errors resulting from differences between the true relevant dimension $\hat{e}_1$ and the dimension $\hat{v}$ that maximizes the objective function for a given (finite) dataset size arise because the probability distributions $P_{\mathbf{v}}(x)$ and $P_{\mathbf{v}}(x|\text{spike})$ are estimated from experimental histograms and differ from the distributions found in the limit of infinite data size. The effect of noise on the reconstruction of the relevant dimension $\hat{v}$ can be characterized by taking the dot product between the relevant dimension and the optimal vector for a particular data sample. Indeed, defining the deviation of the optimal vector from the true relevant dimension as $\delta\mathbf{v} = \hat{v} - \hat{e}_1$, and taking into account that both $\hat{v}$ and $\hat{e}_1$ are normalized, we have $\delta\mathbf{v}^2 = 2(1 - \hat{v} \cdot \hat{e}_1)$. We are interested in finding the expected value $\langle\, \delta\mathbf{v}^2 \,\rangle$, averaged over different instantiations of neural noise. For any given dataset, the deviation $\delta\mathbf{v}$ is not zero, but averaging across different datasets, $\langle\, \delta\mathbf{v} \,\rangle$ approaches zero, because on average empirical probability distributions compiled from experimental histograms are equal to their zero-noise values (in other words, the method is unbiased). The variance of values in the empirical probability distributions, however, is not zero, but decreases as $1/N_{\text{samples}}$ within each bin. Therefore, in general we expect that $\langle \delta\mathbf{v}^2 \rangle \sim 1/N_{\text{spike}}$, and will be small when the number of collected spikes N_{spike} is large. The smallness of $\langle\, \delta\mathbf{v}^2 \rangle$ allows us to use the quadratic approximation of the objective function near its maximum, and neglect differences between the Hessians computed for vectors $\hat{v}$ and $\hat{e}_1$. In what follows, the Hessian will be computed at the true vector $\hat{e}_1$. Note that, in the presence of neural noise, the gradient of objective function is not zero when computed at the true vector $\hat{e}_1$. Because objective function reaches a maximum at the vector $\hat{v}$, we find (using a quadratic approximation) that the deviation $\delta\mathbf{v} = -[H^{(\alpha)}]^{-1}\nabla F^{(\alpha)}$, where $H^{(\alpha)}$ is the Hessian of Rényi divergence of order α [both the gradients $\nabla F^{(\alpha)}$ and $H^{(\alpha)}$ are evaluated along the optimal dimension $\hat{e}_1$]. Similarly to the case of optimizing information (Sharpee et al. 2004), the Hessian of Rényi divergence of order α is given by

$$H_{ij}^{(\alpha)} = \alpha \int \mathrm{d}x\, P(x|\text{spike}) C_{ij}(x) \left[\frac{P(x|\text{spike})}{P(x)}\right]^{\alpha-3} \left[\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{P(x|\text{spike})}{P(x)}\right)\right]^2, \qquad (13)$$

where matrix $C_{ij}(x) = (\langle s_i s_j|x\rangle - \langle s_i|x\rangle\langle s_j|x\rangle)$ is a covariance matrix of stimuli having projection x along the optimal dimension $\hat{e}_1$. Therefore the Hessian has a structure which is similar to the covariance matrix across all stimuli, with the difference that the covariance matrices $C_{ij}(x)$ are weighted more heavily for those projection values x where the gain function is changing more rapidly with x.

In order to measure the expected spread of optimal dimensions around the true relevant dimension $\hat{e}_1$, we need to evaluate $\langle \delta\mathbf{v}^2 \rangle = \mathrm{Tr}[[H^{(\alpha)}]^{-1}\langle \nabla F^{(\alpha)} \nabla F^{(\alpha)T} \rangle [H^{(\alpha)}]^{-1}]$. Therefore, we need to know the variance of the gradient of F averaged across different equivalent datasets. Assuming that the probability of generating a spike is independent for different bins, we find that $\langle \nabla F_i^{(\alpha)} \nabla F_j^{(\alpha)} \rangle = B_{ij}^{(\alpha)} / N_{\text{spike}} + O(N_{\text{spike}}^{-2})$ where

$$B_{ij}^{(\alpha)} = \alpha^2 \int \mathrm{d}x\, P(x|\text{spike}) C_{ij}(x) \left[\frac{P(x|\text{spike})}{P(x)}\right]^{2\alpha-4} \left[\frac{\mathrm{d}}{\mathrm{d}x}\frac{P(x|\text{spike})}{P(x)}\right]^2. \qquad (14)$$

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

$$\hat{v} \cdot \hat{e}_1 \approx 1 - \frac{1}{2}\langle \delta \mathbf{v}^2 \rangle = 1 - \frac{\mathrm{Tr}'[H^{-1}BH^{-1}]}{2N_{\mathrm{spike}}} + O(N_{\mathrm{spike}}^{-2}), \tag{15}$$

where we omitted superscripts $^{(\alpha)}$ for clarity. Tr′ denotes the trace taken in the subspace orthogonal to the relevant dimension [deviations along the relevant dimension do not lead to reconstruction errors (Sharpee et al. 2004); this mathematically manifests itself in the fact that the dimension $\hat{e}_1$ is an eigenvector of matrices H and B with the zero eigenvalue]. Note that when $\alpha = 1$, which corresponds to Kullback–Leibler divergence and information maximization,

$$A \equiv B^{\alpha=1} = H^{\alpha=1} = \int \mathrm{d}x\, P(x|\mathrm{spike}) C_{ij}(x) \left[\frac{\mathrm{d}}{\mathrm{d}x} \ln \left(\frac{P(x|\mathrm{spike})}{P(x)} \right) \right]^2. \tag{16}$$

The asymptotic errors in this case are determined by the trace of the Hessian of information, $\langle \delta \mathbf{v}^2 \rangle \propto \mathrm{Tr}'[A^{-1}]$, reproducing the previous result for maximally informative dimensions (Sharpee et al. 2004). Qualitatively, the expected error for optimization of Rényi divergence of any order behaves as $\sim D/(2N_{\mathrm{spike}})$. This dependence is in common with expected errors of relevant dimensions found by maximizing information (Sharpee et al. 2004), as well as methods based on computing the STA both for white noise (Paninski 2003; Rust et al. 2005; Schwartz et al. 2006) and correlated Gaussian inputs (Sharpee et al. 2004).

Next, we examine which of the optimization strategies based on different Rényi divergences provides the smallest asymptotic error (15). We will denote the gain function as $g(x) = P(x|\mathrm{spike})/P(x)$ and write the covariance matrix as $C_{ij}(x) = \sum_k \gamma_{ik}(x)\gamma_{jk}(x)$ (exact expression for matrices γ will not be needed). With these notations, we can represent the Hessian matrix H and matrix B (covariance matrix for the gradient F) as averages with respect to probability distribution $P(x|\mathrm{spike})$:

$$B = \int \mathrm{d}x\, P(x|\mathrm{spike}) b(x) b^T(x), \quad H = \int \mathrm{d}x\, P(x|\mathrm{spike}) a(x) b^T(x), \tag{17}$$

where $b_{ij}(x) = \alpha \gamma_{ij}(x) g'(x) [g(x)]^{\alpha-2}$ and $a_{ij}(x) = \gamma_{ij}(x) g'(x)/g(x)$. The Cauchy–Schwarz inequality for scalar quantities states that, $\langle b^2 \rangle / \langle ab \rangle^2 \geq 1/\langle a^2 \rangle$, where the averages are taken with respect to some probability distribution. A similar result can also be proven for matrices under a Tr operation, as in Equation 15 [see appendix of Sharpee (2007) for this derivation]. Applying the Cauchy–Schwarz inequality to Equation 15, we find that the smallest error is obtained when

$$\mathrm{Tr}'[H^{-1}BH^{-1}] = \mathrm{Tr}'[A^{-1}], \quad \text{with} \quad A = \int \mathrm{d}x\, P(x|\mathrm{spike}) a(x) a^T(x), \tag{18}$$

Matrix A corresponds to the Hessian of the merit function for $\alpha = 1$: $A = H^{\alpha=1}$. Thus, among the various optimization strategies based on Rényi divergences, Kullback–Leibler divergence ($\alpha = 1$) has the smallest asymptotic errors. This is a somewhat surprising result, because the least square fitting corresponds to optimization based on Rényi divergence with $\alpha = 2$, whereas Kullback–Leibler divergence implements information maximization. At the same time,

Kullback–Leibler divergence has also been shown to be the optimal objective function for the problem of lossy compression (Harremoës and Tishby 2007), suggesting that robustness of information maximization may be a common theme in diverse dimensionality reduction problems.

Maximizing information as a way to perform likelihood maximization

Maximum likelihood is a popular general method for estimating parameters of a model that are most consistent with a series of observations (Daniels 1961; MacKay 2003). In a number of dimensionality reduction problems, including independent component analysis (Bell and Sejnwoski 1995), the information-bottleneck method (Tishby et al. 1999; Dimitrov et al. 2003), or learning with Boltzmann machines (Hinton and Sejnowski 1986), the likelihood may be expressed as a Kullback–Leibler divergence between a pair of probability distributions whose particular expression depends on the problem at hand (Cardoso 1997; Dayan and Abbott 2001; Kinney et al. 2007; Sahani 2008). Here we show that maximizing information as given by the Kullback–Leibler divergence between probability distributions $P(x|\text{spike})$ and $P(x)$ represents a convenient way to perform likelihood maximization, provided spikes are ''rare'', i.e., $P(\text{spike})$ to observe a spike at time window Δt is small.

The log-likelihood that a given sequence of neural responses were elicited based on stimulus components along $\mathbf{v}$ is given by:

$$L[\mathbf{v}] = N \int \mathrm{d}\mathbf{s} \sum_{\text{response}} P(\text{response}, \mathbf{s}) \ln P(\text{response}, \mathbf{s}|\mathbf{v}).$$

Here we assume that responses to different stimuli $\mathbf{s}$ represent independent observations. According to the assumptions of the linear–nonlinear model

$$P(\text{response}, \mathbf{s}|\mathbf{v}) = P(\text{response}|\mathbf{s}\cdot\mathbf{v})P(\mathbf{s}|\mathbf{v}) = P(\text{response}|\mathbf{s}\cdot\mathbf{v})P(\mathbf{s}), \qquad (19)$$

where we took into account that the stimulus distribution is independent of which stimulus dimension $\mathbf{v}$ is presumed to be relevant. Therefore, the log-likelihood can be written as:

$$\begin{aligned} L[\mathbf{v}] &= N \int \mathrm{d}\mathbf{s} \sum_{\text{response}} P(\mathbf{s})P(\text{response}|\mathbf{s}\cdot\mathbf{v}) \ln P(\text{response}|\mathbf{s}\cdot\mathbf{v}) \\ &\quad + N \int \mathrm{d}\mathbf{s} \sum_{\text{response}} P(\mathbf{s})P(\text{response}|\mathbf{s}\cdot\mathbf{v}) \ln P(\mathbf{s}) \\ &= N \int \mathrm{d}\mathbf{s} \sum_{\text{response}} P(\mathbf{s})P(\text{response}|\mathbf{s}\cdot\mathbf{v}) \ln P(\text{response}|\mathbf{s}\cdot\mathbf{v}) - NH[\mathbf{s}], \end{aligned}$$

where the last term represents the entropy of the stimulus distribution $H[\mathbf{s}]$ and does not depend on $\mathbf{v}$. One can carryout the integration with respect to all stimulus components that are orthogonal to dimension $\mathbf{v}$. Then, the log-likelihood can be written as:

$$L[\mathbf{v}] = N \int \mathrm{d}(\mathbf{s}\cdot\mathbf{v}) \sum_{\text{response}} P(\text{response}, \mathbf{s}\cdot\mathbf{v}) \ln P(\text{response}|\mathbf{s}\cdot\mathbf{v}) - NH(\mathbf{s}).$$

This expression can be further separated into model-dependent and model-independent terms, which includes the entropy of neural response, $H[\text{response}] = \sum_{\text{response}} P(\text{response}) \ln P(\text{response})$:

$$L[\mathbf{v}] = N \int \mathrm{d}(\mathbf{s}\cdot\mathbf{v}) \sum_{\text{response}} P(\text{response}, \mathbf{s}\cdot\mathbf{v}) \ln \frac{P(\text{response}|\mathbf{s}\cdot\mathbf{v})}{P(\text{response})} - NH[\mathbf{s}] - NH[\text{response}]$$
$$= NI[\text{response}, \mathbf{s}\cdot\mathbf{v}] - NH[\mathbf{s}] - NH[\text{response}], \tag{20}$$

where $I[\text{response}, \mathbf{s}\cdot\mathbf{v}]$ is the information between the neural response (either "spike" or "no spike") and stimulus components along the dimension $\mathbf{v}$. Finding relevant dimensions by maximizing information $I[\mathbf{v}]$ (8) or, more generally Rényi divergences (7), is carried out assuming that spikes are "rare", i.e., that $P(\text{spike}) = \bar{r}\Delta t \ll 1$ (Δt is the width of time bins used to discretize spike trains and stimulus sequences). This condition can always be satisfied by choosing a sufficiently fine discretization time scale Δt. When $P(\text{spike}) \ll 1$, both $P(\text{no spike}|\mathbf{s}\cdot\mathbf{v})$ and $P(\text{no spike})$ tend to 1, which makes their contribution to $I[\text{response}, \mathbf{s}\cdot\mathbf{v}]$ negligible because of the logarithm (Brenner et al. 2000), so that information is dominated by "spike" events:

$$I[\text{response}, \mathbf{s}\cdot\mathbf{v}] \approx P(\text{spike}) I[\mathbf{v}] + O\big(P(\text{spike})^2\big).$$

Here, $I[\mathbf{v}]$ is the mutual information (8) between spikes and stimulus components along the dimension $\mathbf{v}$. Therefore, omitting terms that are independent of $\mathbf{v}$, maximizing log-likelihood is equivalent, in the limit of low average spike probability, to maximizing information:

$$L[\mathbf{v}] = \text{const} + N_{\text{spike}} I[\mathbf{v}].$$

We note that in other situations, for example where the nonlinear gain function is not estimated empirically, but is instead chosen according to some prior assumptions, information maximization is not precisely equivalent to maximum likelihood (Kinney et al. 2007). However, for a large class of priors, including the uniform prior on the shape of the nonlinear gain function, the difference between information maximization and maximum likelihood vanishes in the limit of infinite data (Kinney et al. 2007).

Fisher information

Maximum likelihood methods of estimating model parameters are known to achieve the lowest variance among unbiased methods under quite general assumptions (Daniels 1961). The Cramér–Rao inequality (Cover and Thomas 1991) states that the smallest achievable variance is equal to a trace of the inverse of the Fisher information matrix $\langle \delta \mathbf{v}^2 \rangle = \mathrm{Tr}'[I_F^{-1}]$. Here we demonstrate that the Fisher information matrix $I_F = N_{\text{spike}} A$, where matrix A from Equation 16 determines the reconstruction error for finding relevant stimulus dimensions by maximizing information. According to Equation 15, proving this statement would mean that information does indeed saturate the Cramér–Rao bound. To demonstrate this, we note that the Fisher information matrix along a stimulus dimension $\mathbf{v}$ is given by:

$$I_{Fij} = N \int \mathrm{d}\mathbf{s} P(\mathbf{s}) P(\text{spike}, \mathbf{s}|\mathbf{v}) \partial_{v_i} [\ln P(\text{spike}, \mathbf{s}|\mathbf{v})] \partial_{v_j} [\ln P(\text{spike}, \mathbf{s}|\mathbf{v})] + NO\big((\bar{r}\Delta t)^2\big),$$

where likelihood from data points with a ''no spike'' response contributes only terms of order $O((\bar{r}\Delta t)^2)$. Using the above arguments (19), together with the definition of the nonlinear gain function (1), the Fisher information matrix can be re-written as:

$$I_{Fij} = N_{\text{spike}} \int \mathrm{d}\mathbf{s} P(\mathbf{s}) g(\mathbf{s}\cdot\mathbf{v}) \partial_{v_i}[\ln g(\mathbf{s}\cdot\mathbf{v})] \partial_{v_j}[\ln g(\mathbf{s}\cdot\mathbf{v})]. \tag{21}$$

Two factors contribute to the derivative $\partial_{v_i}[\ln g(\mathbf{s}\cdot\mathbf{v})]$. The first contribution comes from changes in the functional form of $g(x)$ following a change in $\mathbf{v}$. The second contribution is due to changes in argument $\mathbf{s}\cdot\mathbf{v}$ for an unchanged function $g(x)$:

$$\partial_{v_i}[\ln g(\mathbf{s}\cdot\mathbf{v})] = s_i \frac{\mathrm{d}\ln g(x)}{\mathrm{d}x} + \frac{\partial_{v_i} P_{\mathbf{v}}(x|\text{spike})}{P_{\mathbf{v}}(x|\text{spike})} - \frac{\partial_{v_i} P_{\mathbf{v}}(x)}{P_{\mathbf{v}}(x)}, \quad x = \mathbf{s}\cdot\mathbf{v}. \tag{22}$$

The last two terms in Equation (22) represent derivatives with respect to changes in the functional form of the probability distributions $P_{\mathbf{v}}(x|\text{spike})$ and $P_{\mathbf{v}}(x)$ when the dimension $\mathbf{v}$ changes. We note that $P_{\mathbf{v}}(x)$ and $P_{\mathbf{v}}(x|\text{spike})$ can be expressed as

$$P_{\mathbf{v}}(x) = \int \mathrm{d}\mathbf{s} P(\mathbf{s}) \delta(x - \mathbf{s}\cdot\mathbf{v}), \quad P_{\mathbf{v}}(x|\text{spike}) = \int \mathrm{d}\mathbf{s}\, P(\mathbf{s}|\text{spike}) \delta(x - \mathbf{s}\cdot\mathbf{v}), \tag{23}$$

where $\delta(x)$ is a delta-function. Using these expressions, we find that

$$\partial_{v_i} P_{\mathbf{v}}(x|\text{spike}) = -\frac{\mathrm{d}}{\mathrm{d}x}[\langle s_i|x, \text{spike}\rangle P(x|\text{spike})], \quad \partial_{v_i} P_{\mathbf{v}}(x) = -\frac{\mathrm{d}}{\mathrm{d}x}[\langle s_i|x\rangle P(x)].$$

The Fisher information matrix is evaluated at the maximum of the likelihood function attained at the true relevant dimension $\mathbf{v} = \hat{e}_1$. Stimulus projections along the true relevant dimensions completely determine the spike probability, 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,

$$\partial_{v_i}[\ln g(x)] = (s_i - \langle s_i|x\rangle)\left[\frac{\mathrm{d}}{\mathrm{d}x}\ln g(x)\right], \quad x = \mathbf{s}\cdot\hat{e}_1.$$

Expression (21) for the Fisher information matrix can now be written as:

$$I_{Fij} = N_{\text{spike}} \int \mathrm{d}\mathbf{s}_\perp \mathrm{d}s_1 P(\mathbf{s}_\perp, s_1) g(s_1)(s_i - \langle s_i|s_1\rangle)\big(s_j - \langle s_j|s_1\rangle\big)\left[\frac{\mathrm{d}\ln g(s_1)}{\mathrm{d}s_1}\right]^2,$$

where $\mathbf{s}_\perp$ represents all stimulus components that are orthogonal to the true dimension $\hat{e}_1$. Integrating over these stimulus components $\mathbf{s}_\perp$, we get the final answer:

$$\begin{aligned} I_{Fij} &= N_{\text{spike}} \int \mathrm{d}s_1 P(s_1) g(s_1)\big(\langle s_i s_j|s_1\rangle - \langle s_i|s_1\rangle\langle s_j|s_1\rangle\big)\left[\frac{\mathrm{d}\ln g(s_1)}{\mathrm{d}s_1}\right]^2 \\ &= N_{\text{spike}} A_{ij}. \end{aligned}$$

Comparison with Equations 15–18 demonstrates that the variance in the relevant stimulus dimensions obtained by maximizing information

$$\langle \delta\mathbf{v}^2 \rangle = \frac{\mathrm{Tr}'[A^{-1}]}{N_{\text{spike}}} = \mathrm{Tr}'\big[I_F^{-1}\big]$$

saturates the Cramér–Rao bound. This indicates that information maximization, which in the case of estimating linear–nonlinear models is equivalent to maximum likelihood, provides the smallest possible reconstruction error.

Performance in the regime of small numbers of spikes

We now use numerical simulations on model cells to study the performance of maximizing Rényi divergences in the regime of relatively small number of spikes. We are interested in the range $0.1 \lesssim D/N_{\text{spike}} \lesssim 1$ (i.e., the number of spike is small with respect to stimulus dimensionality D), where the asymptotic results do not necessarily apply. We compare performance of maximizing information ($\alpha = 1$), variance ($\alpha = 2$), and Rényi divergence of order 3 ($\alpha = 3$). The results of simulations for various numbers of spikes and four different neural noise levels are shown in Figure 3 as a function of D/N_{spike}. The four model cells had the same relevant dimension shown in Figure 2(a). The nonlinear gain functions had a threshold value of $\theta = 2.0$ and noise standard deviation $\sigma = 1.5$, 1.0, 0.5, 0.25 for groups labeled A–D, respectively. Noise standard deviation σ was taken as a proxy for neural noise level. Relevant dimensions for each of the four model cells were found by optimizing

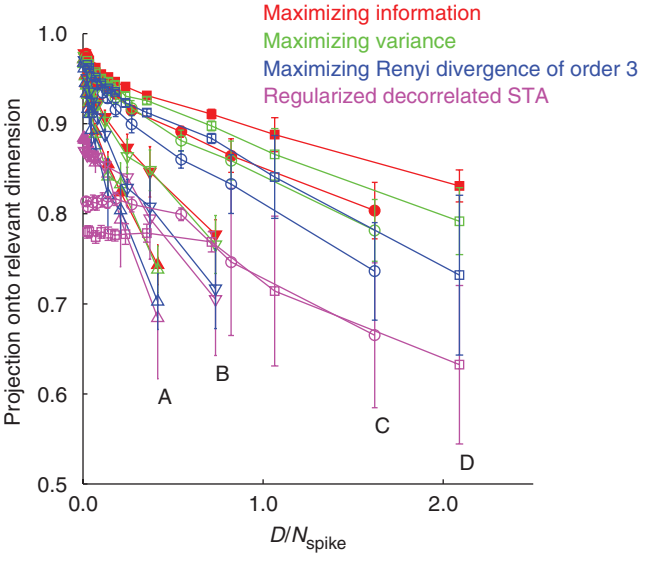

Figure 3. Reconstruction quality in the regime of medium to small numbers of recorded spikes. Projection of vector $\hat{v}_{\max}$ obtained by maximizing information (red filled symbols), variance (green open symbols), Rényi divergence of order 3 (blue open symbols), or regularized decorrelated STA (magenta 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 Figure 2(a) and the input–output function 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. All error bars show standard deviations across 10 different model neurons.

Rényi divergences of orders 1–3. Identical numerical algorithms were used for all three cost functions. Computations were performed with number of bins equal to 15, 21, 32, and 64 for cells A–D, respectively (larger number of bins were needed to characterize sharper nonlinear gain functions for smaller neural noise levels). While the maximum number of iterations was limited to 1000 line optimizations (stimulus dimensionality was $D = 900$), we have also allowed for the possibility of an early stopping in the optimization. To implement early stopping, the data for each model cell was split into a training part (three-fourth size of the overall dataset) and a test part (one-fourth of the overall dataset). Stimulus dimensions were searched to maximize performance on the training dataset, and then applied to the test dataset after each line optimization. Finally, relevant stimulus dimension was selected as the one that yielded the best performance on the test dataset. To ensure that we did not miss the actual maximum, optimization continued until performance on the test dataset fell below 75% of its maximum (The numerical code is available at http://cnl-t.salk.edu.). For each model cell, four such relevant dimensions were computed (corresponding to four different ways of selecting the test dataset). The average of these four dimensions was used for all subsequent comparisons between methods. Generally good reconstructions with projection values $\gtrsim 0.7$ can be obtained by maximizing either information, variance, or Rényi divergence of order 3, even in the severely undersampled regime $D > N_{\text{spike}}$.

Although the reconstruction errors are numerically very similar across the three optimization strategies, information maximization achieved consistently smaller errors throughout all simulation conditions (Figure 4) in agreement with the analytical derivations for large numbers of spikes. The performance of using the Rényi divergence of order 1 (information maximization) was slightly, but significantly, superior to the Rényi divergence of order 2 (least square fitting), cf. Figure 4(a). Optimization of Rényi divergence of order 2 was in turn superior to the Rényi divergence of order 3, cf. Figure 4(b) [$p < 10^{-4}$, paired t-tests between Rényi divergences of order 1, 2, and 3 combining the results across different noise levels and spike numbers]. The standard deviation between reconstructed vectors for different instantiations of the same model neurons was also less for information maximization than for variance maximization, cf. Figure 4(c), which in turn is less than for maximizing Rényi divergence of order 3, cf. Figure 4(d). The improvements in reconstruction quality from optimizing Rényi divergences of smaller orders were greater for less noisy neurons (with smaller $\sigma = 0.25$). These numerical results suggest that reconstruction errors in relevant dimensions obtained by maximizing Rényi divergences systematically improve with decreasing the order of the Rényi divergence. This is perhaps due to the fact that deviations between probability distributions $P_{\mathbf{v}}(x)$ and $P_{\mathbf{v}}(x|\text{spike})$ are compared in the expression for the Rényi divergence of order α after being raised to power α, cf. Equation 7. Higher powers are known to make fitting algorithms more susceptible to outliers (Press et al., 1992). In fact, robust fitting often favors minimizing mean absolute value of deviations, as in the case where errors are distributed according to a Laplace distribution. Our analysis demonstrates that the best results, both in terms of the projection value onto the true relevant dimension and in terms of variance across different instantiations of the same neural noise model, are obtained by optimizing the Rényi divergence of order 1, also known as the Kullback–Leibler divergence.

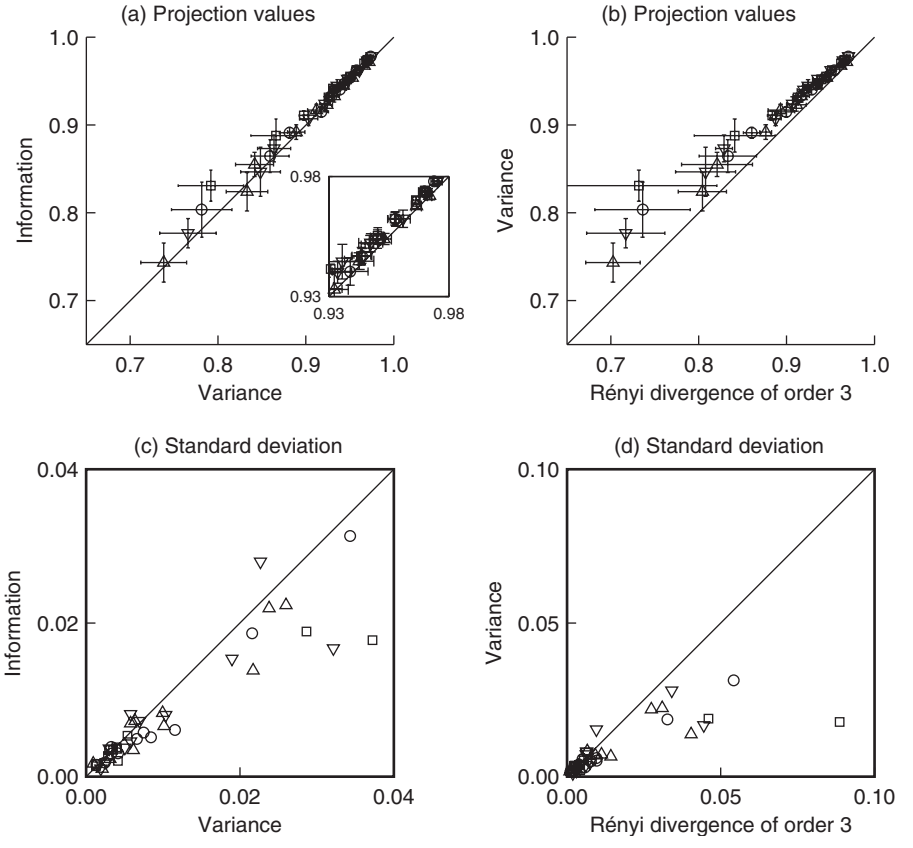

Figure 4. Reconstruction errors increase with the order of the maximized Rényi divergence. Projection values of the reconstructed dimension $\hat{v}_{\max}$ onto the true relevant dimension $\hat{e}_1$ are compared for maximizing information and variance in (a), and for maximizing variance and Rényi divergence of order 3 in (b). Most data points lie above the diagonal line, indicating that maximizing information performed better than maximizing variance, which in turn performed better than maximizing Rényi divergence of order 3. Panels (c) and (d) show the standard deviation of the projection values. Majority of the data points lie below the diagonal line, indicating that the procedure of maximizing information performed with lower variance than maximizing Rényi divergences of order 2 or 3. The model neurons with the smallest noise ($\sigma = 0.25$, □ symbols) tend to lie farther from the diagonal lines. This suggests that the smaller the neural noise, the greater is the improvement in performance obtained by maximizing Rényi divergences of smaller orders. Notations are as in Figure 3, with data for $\sigma = 1.5$, 1.0, 0.5, 0.25 plotted using △, ▽, ○, and □ symbols, respectively.

Summary and discussion

We have compared the accuracy of a family of optimization strategies for analyzing neural responses based on the linear–nonlinear model using Rényi divergences. The advantage of this approach over the standard least squares fitting procedure is that it does not require their nonlinear gain function to be invertible. Finding relevant dimensions by maximizing one of the objective functions, the Rényi divergence of order 2, corresponds to minimizing the χ^2 difference between data and model predictions. In the asymptotic regime of large spike numbers, an analytical derivation of the reconstruction errors expected for maximizing

Rényi divergences of arbitrary order demonstrated that the smallest errors are achieved by information maximization (via Kullback–Leibler divergence or the Rényi divergence of order 1). We showed that, in the case of the linear–nonlinear model, finding relevant stimulus features by maximizing information is equivalent to using the maximum likelihood method (Daniels 1961). Correspondingly, the variance of relevant dimensions found by maximizing information achieves the smallest value allowed by the Cramér–Rao bound for any unbiased method. This method also avoids problems due to incorrect prior assumptions about the functional form of the nonlinear gain function that might invalidate relevant stimulus features found by maximum likelihood (Kinney et al. 2007). Although the maximum likelihood estimation of relevant dimensions is a convex optimization problem for Gaussian inputs and exponential nonlinearities (Paninski 2004), the optimization problem is no longer convex in the case of more complex, non-Gaussian natural stimuli. Nevertheless, as we demonstrate here, this non convex optimization problem can be successfully solved by information maximization.

In the limit of large spike numbers, a near perfect reconstruction can be obtained by maximizing either information, variance, or a Rényi divergence of the third order (cf. Figure 2). However, analysis of real neural data can be complicated in part by input nonlinearities (Ahrens et al. 2008), jitter in spike timing or spike-history effects that can cause the spike probability for a given stimulus to deviate from a Poisson distribution. One potential way of capturing the response properties of the real neurons is to find the transformation of the stimuli that effectively linearizes the input dimensions and to apply linear–nonlinear model assumption in this transformed space. Such a strategy has been used in the analysis of V4 neurons (David et al. 2006). Another potential way is to incorporate other nonlinear operations directly into the model and optimization routines. The spike-history effects can be taken into account by expanding the linear–nonlinear model so that spike trains can influence future spike probabilities in either the same neuron (Pillow et al. 2005) or interconnected pairs (Pillow et al. 2008). With respect to spike-time jitter, recent studies (Aldworth et al. 2005; Dimitrov and Gedeon, 2006; Gollisch 2006; Dimitrov et al. 2009) demonstrate that the estimation of the relevant stimulus dimensions can be improved significantly in the case of Gaussian stimuli by taking spike-time jitter into account. The dimensionality reduction techniques discussed here in the case of natural stimuli can presumably be also improved by incorporating the spike-timing jitter into account. We hope to undertake the development of the methods addressing these issues in the future.

Numerical analysis in the regime of medium-to-small numbers of spikes demonstrated that the optimization of Rényi divergences leads to reliable reconstructions of relevant stimulus features for model cells, even when the number of spikes is less than the stimulus dimensionality. Estimation of linear–nonlinear models using information maximization had significantly smaller errors than those derived using either least square fitting or maximizing Rényi divergence of higher orders. This makes the problem of finding relevant dimensions one of the examples where information-theoretic measures are no more data-limited than variance-based measures. Our findings complement recent analyses of a related dimensionality reduction problem showing that Kullback–Leibler divergence is also an optimal measure for performing lossy compression

(Harremoës and Tishby 2007). This suggests that robustness of information maximization might be a common theme in diverse dimensionality reduction problems.

Acknowledgments

We would like to thank William Bialek, Naftali Tishby, Surya Ganguli, and Yuan Liu for helpful discussions. The relationship between the maximum likelihood and information maximization methods, which is the topic of the Section ''maximizing information as a way to perform likelihood maximization'', was suggested to us by William Bialek. This research was supported by grant K25MH068904 from the National Institute of Mental Health, Searle Scholarship, The Ray Thomas Edwards Career Development Award in Biomedical Sciences and Alfred P. Sloan Research Fellowship. 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).

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Ahrens MB, Paninski L, Sahani M. 2008. Inferring input nonlinearities in neural encoding models. Network: Computation in Neural Systems 19:35–67.

Aldworth ZN, Miller JP, Gedeon T, Cummins GI, Dimitrov AG. 2005. Dejittered spike-conditioned stimulus waveforms yield improved estimates of neuronal feature selectivity and spike-timing precision of sensory interneurons. Journal of Neuroscience 25:5323–5332.

Ali SM, Silvey SD. 1966. A general class of coefficient of divergence of one distribution from another. Journal of Royal Statistical Society B 28:131–142.

Bell AJ, Sejnwoski TJ. 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7:1129–1159.

Bialek W, de Ruyter van Steveninck RR. 2005, Features and dimensions: Motion estimation in fly vision. arXiv:q-bio/0505003v1.

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

Cardoso JF. 1997. Infomax and maximum likelihood for blind source separation. Signal Processing Letters, IEEE 4:112–114.

Christianson GB, Sahani M, Linden JF. 2008. The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. Journal of Neuroscience 28:446–455.

Cover TM, Thomas JA. 1991. Information theory. New York: John Wiley & Sons, INC.

Csiszár I. 1967. Information-type measures of difference of probability distributions and indirect observations. Studia Scientarium Mathematicarum Hungararica 2:299–318.

Daniels HE. 1961. The asymptotic efficiency of a maximum likelihood estimator. Proceeding of the Fourth Symposium on Mathematical Statistics and Problems 1:151–163.

David SV, Gallant JL. 2005. Predicting neuronal responses during natural vision. Network 22:239–260.

David SV, Hayden BY, Gallant JL. 2006. Spectral receptive field properties explain shape selectivity in area V4. Journal of Neurophysics 96:3492–3505.

Dayan P, Abbott LF. 2001. Theoretical neuroscience: Computational and mathematical modeling of neural systems Cambridge: MIT Press.

de Boer E, Kuyper P. 1968. Triggered correlation. IEEE Transactions Biomedical Engineering 15:169–179.

de Ruyter van Steveninck RR, Bialek W. 1988. Real-time performance of a movement-sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. Proceeding of the Royal Society London B 265:259–265.

Dimitrov AG, Gedeon T. 2006. Effects of stimulus transformations on estimates of sensory neuron selectivity. Journal of Computational Neuroscience 20:265–283.

Dimitrov AG, Miller JP, Gedeon T, Aldworth Z, Parker AE. 2003. Analysis of neural coding through quantization with an information-based distortion measure. Network: Computation Neural Systems 14:151–176.

Dimitrov AG, Sheiko MA, Baker J, Yen SC. 2009. Spatial and temporal jitter distort estimated functional properties of visual sensory neurons. DOI: 10.1007/S10827-009-0144-8. Journal of Computational Neuroscience [Epub ahead of print].

Felsen G, Touryan J, Han F, Dan Y. 2005. Cortical sensitivity to visual features in natural scenes. Public Library of Science Biology 3:1819–1828.

Gollisch T. 2006. Estimating receptive fields in the presence of spike-time jitter. Network: Computation in Neural Systems 17:103–129.

Harremoës P, Tishby N. 2007. The information bottleneck revisited or how to choose a good distortion measure. Proceeding of the IEEE International Symposium on Information Theory (ISIT) .

Hinton GE, Sejnowski TJ. 1986. Learning and relearning in Boltzmann machines. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge: MIT Press. pp 282–317.

Hunter IW, Korenberg MJ. 1986. The identification of nonlinear biological systems: Wiener and hammerstein cascade models. Biological Cybernetics 55:135–144.

Kinney JB, Tkaćik G, Callan CG. 2007. Precise physical models of protein-DNA interaction from high-throughput data. Proceeding of the National Academy of Science 104:501–506.

MacKay D. 2003. Information theory, inference, and learning algorithms Cambridge: Cambridge University Press.

Marmarelis VZ. 1997. Modeling methodology for nonlinear physiological systems. Annals of Biomedical Engineering 25:239–251.

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

Paninski L. 2004. Maximum likelihood estimation of cascade point-process neural encoding models. Network: Computation in Neural Systems 15:243–262.

Pillow JW, Paninski L, Uzzell VJ, Simoncelli EP, Chichilnisky EJ. 2005. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. Journal of Neuroscience 25:11003–11013.

Pillow JW, Shlens J, Paninski L, Shar A, Litke AM, Chichilnisky EJ, Simoncelli EP. 2008. Spatio-temporal correlations and visual signaling in a complete neural population. Nature 454:995–999.

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

Rényi, A. 1961. On measures of entropy and information. In: Neyman, J, editor. Proceeding of the Fourth Berkeley Symposium Mathematical Statistics Problems. Vol. 1: pp 547–561.

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

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

Ringach DL, Sapiro G, Shapley R. 1997. A subspace reverse-correlation technique for the study of visual neurons. Vision Research 37:2455–2464.

Ruderman DL, Bialek W. 1994. Statistics of natural images: Scaling in the woods. Physical Review Letters 73:814–817.

Rust NC, Schwartz O, Movshon JA, Simoncelli EP. 2005. Spatiotemporal elements of macaque v1 receptive fields. Neuron 46:945–956.

Sahani M. 2008, Neural coding. [Published 2008]. Available from: http://www.gatsby.ucl.ac.uk/~maneesh/

Schwartz O, Pillow JW, Rust NC, Simoncelli EP. 2006. Spike-triggered neural characterization. Journal of Vision 176:484–507.

Sharpee T, Rust NC, Bialek W. 2004. Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Computation* 16:223–250, 2004. [See also physics/0212110 and a preliminary account in Advances in Neural Information Processing 15 edited by Becker S, Thrun S, Obermayer K. 2003. pp. 261–268.]

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

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

Simoncelli EP, Olshausen BA. 2001. Natural image statistics and neural representation. Annual Review of Neuroscience 24:1193–1216.

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

Tishby N, Pereira FC, Bialek W. 1999. The information bottleneck method. In Hajek B and Sreenivas RS, editors. Proceedings of the 37th Allerton Conference on Communication, Control and Computing. pp. 368–377, University of Illinois. [See also physics/0004057.]

Weisberg S, Welsh AH. 1994. Adapting for the missing link. Annals of Statistics 22:1674–1700.