RESEARCH ARTICLE

The Equivalence of Information-Theoretic and Likelihood-Based Methods for Neural Dimensionality Reduction

Ross S. Williamson[1,2¤]*, Maneesh Sahani[1], Jonathan W. Pillow[3]*

1 Gatsby Computational Neuroscience Unit, University College London, London, UK, **2** Centre for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London, UK, **3** Princeton Neuroscience Institute, Department of Psychology, Princeton University, Princeton, New Jersey, USA

¤ Current Affiliation: Eaton-Peabody Laboratories, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA & Center for Computational Neuroscience and Neural Technology, Boston University, Boston, Massachusetts, USA

* ross_williamson@meei.harvard.edu (RSW); pillow@princeton.edu (JWP)

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Stimulus dimensionality-reduction methods in neuroscience seek to identify a low-dimensional space of stimulus features that affect a neuron's probability of spiking. One popular method, known as maximally informative dimensions (MID), uses an information-theoretic quantity known as "single-spike information" to identify this space. Here we examine MID from a model-based perspective. We show that MID is a maximum-likelihood estimator for the parameters of a linear-nonlinear-Poisson (LNP) model, and that the empirical single-spike information corresponds to the normalized log-likelihood under a Poisson model. This equivalence implies that MID does not necessarily find maximally informative stimulus dimensions when spiking is not well described as Poisson. We provide several examples to illustrate this shortcoming, and derive a lower bound on the information lost when spiking is Bernoulli in discrete time bins. To overcome this limitation, we introduce model-based dimensionality reduction methods for neurons with non-Poisson firing statistics, and show that they can be framed equivalently in likelihood-based or information-theoretic terms. Finally, we show how to overcome practical limitations on the number of stimulus dimensions that MID can estimate by constraining the form of the non-parametric nonlinearity in an LNP model. We illustrate these methods with simulations and data from primate visual cortex.

Author Summary

A popular approach to the neural coding problem is to identify a low-dimensional linear projection of the stimulus space that preserves the aspects of the stimulus that affect a neuron's probability of spiking. Previous work has focused on both information-theoretic and likelihood-based estimators for finding such projections. Here, we show that these two approaches are in fact equivalent. We show that maximally informative dimensions (MID), a

popular information-theoretic method for dimensionality reduction, is identical to the maximum-likelihood estimator for a particular linear-nonlinear encoding model with Poisson spiking. One implication of this equivalence is that MID may not find the information-theoretically optimal stimulus projection when spiking is non-Poisson, which we illustrate with a few simple examples. Using these insights, we propose novel dimensionality-reduction methods that incorporate non-Poisson spiking, and suggest new parametrizations that allow for tractable estimation of high-dimensional subspaces.

Introduction

The neural coding problem, an important topic in systems and computational neuroscience, concerns the probabilistic relationship between environmental stimuli and neural spike responses. Characterizing this relationship is difficult in general because of the high dimensionality of natural signals. A substantial literature therefore has focused on dimensionality reduction methods for identifying which stimuli affect a neuron's probability of firing. The basic idea is that many neurons compute their responses in a low dimensional subspace, spanned by a small number of stimulus features. By identifying this subspace, we can more easily characterize the nonlinear mapping from stimulus features to spike responses [1–5].

Neural dimensionality-reduction methods can be coarsely divided into three classes: (1) moment-based estimators, such as spike-triggered average (STA) and covariance (STC) [1, 5–8]; (2) model-based estimators, which rely on explicit forward encoding models [9–16]; and (3) information and divergence-based estimators, which seek to reduce dimensionality using an information-theoretic cost function [17–22]. For all such methods, the goal is to find a set of linear filters, specified by the columns of a matrix K, such that the probability of response r given a stimulus $\mathbf{s}$ depends only on the linear projection of $\mathbf{s}$ onto these filters, i.e., $p(r|\mathbf{s}) \approx p(r|K^\top \mathbf{s})$. Existing methods differ in computational complexity, modeling assumptions, and stimulus requirements. Typically, moment-based estimators have low computational cost but succeed only for restricted classes of stimulus distributions, whereas information-theoretic and likelihood-based estimators allow for arbitrary stimuli but have high computational cost. Previous work has established theoretical connections between moment-based and likelihood-based estimators [11, 14, 17, 19, 23], and between some classes of likelihood-based and information-theoretic estimators [14, 20, 21, 24].

Here we focus on maximally informative dimensions (MID), a well-known information-theoretic estimator introduced by Sharpee, Rust & Bialek [18]. We show that this estimator is formally identical to the maximum likelihood (ML) estimator for the parameters of a linear-nonlinear-Poisson (LNP) encoding model. Although previous work has demonstrated an asymptotic equivalence between these methods [20, 24, 25], we show that the correspondence is exact, regardless of time bin size or the amount of data. This equivalence follows from the fact that the plug-in estimate for the single-spike information [26], the quantity that MID optimizes, is equal to a normalized Poisson log-likelihood.

The connection between the MID estimator and the LNP model makes clear that MID does not incorporate information carried by non-Poisson statistics of the response. We illustrate this shortcoming by showing that MID can fail to find information-maximizing filters for simulated neurons with binary or other non-Poisson spike count distributions. To overcome this limitation, we introduce new dimensionality-reduction estimators based on non-Poisson noise models, and show that they can be framed equivalently in information-theoretic or likelihood-based terms.

Finally, we show that a model-based perspective leads to strategies for overcoming a limitation of traditional MID, that it cannot tractably estimate more than two or three filters. The difficulty arises from the intractability of using histograms to estimate densities in high-dimensional subspaces. However, the single-spike information depends only on the ratio of densities, which is proportional to the nonlinearity in the LNP model. We show that by restricting the parametrization of this nonlinearity so that the number of parameters does not grow exponentially with the number of dimensions, we can obtain flexible yet computationally tractable estimators for models with many filters or dimensions.

Results

Background

Linear-nonlinear-Poisson (LNP) encoding model.   Linear-nonlinear cascade models provide a useful framework for describing neural responses to high-dimensional stimuli. These models define the response in terms of a cascade of linear, nonlinear, and probabilistic spiking stages (see Fig. 1). The linear stage reduces the dimensionality by projecting the high-dimensional stimulus onto a set of linear filters, and a nonlinear function then converts the output of these filters to a non-negative spike rate.

Let $\theta = \{K, \alpha\}$ denote the parameters of the LNP model, where K is a (tall, skinny) matrix whose columns contain the stimulus filters (for cases with a single filter, we will denote the filter with a vector $\mathbf{k}$ instead of the matrix K), and α are parameters governing the nonlinear function f from feature space to instantaneous spike rate. Under this model, the probability of a spike response r given stimulus $\mathbf{s}$ is governed by a Poisson distribution:

$$
\begin{aligned}
\lambda &= f(K^\top \mathbf{s}) \\
p(r|\lambda) &= \tfrac{1}{r!}(\Delta\lambda)^r e^{-\Delta\lambda},
\end{aligned}
\tag{1}
$$
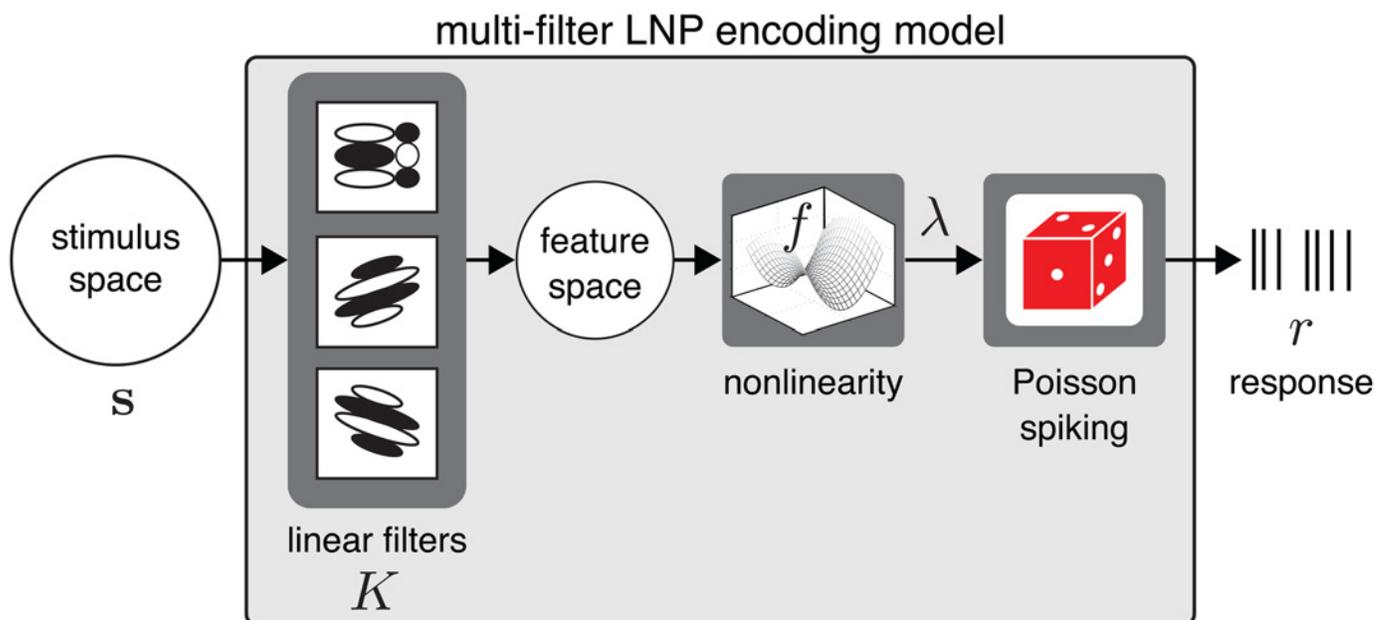
Fig 1. The linear-nonlinear-Poisson (LNP) encoding model formalizes the neural encoding process in terms of a cascade of three stages. First, the high-dimensional stimulus $\mathbf{s}$ projects onto bank of filters contained in the columns of a matrix K, resulting in a point in a low-dimensional neural feature space $K^\top \mathbf{s}$. Second, an instantaneous nonlinear function f maps the filtered stimulus to an instantaneous spike rate λ. Third, spikes r are generated according to an inhomogeneous Poisson process.

where λ denotes the stimulus-driven spike rate (or "conditional intensity") and Δ denotes a time bin size. The defining feature of a Poisson process is that responses in non-overlapping time bins are conditionally independent given the spike rate. In a discrete-time LNP model, the conditional probability of a dataset $D = \{(\mathbf{s}_t, r_t)\}$, consisting of stimulus-response pairs indexed by $t \in \{1, \ldots, N\}$, is the product of independent terms. The log-likelihood is therefore a sum over time bins:

$$\mathcal{L}_{lnp}(\theta; D) = \sum_{t=1}^{N} \log p(r_t | \mathbf{s}_t, \theta) = \sum_{t=1}^{N} \left(r_t \log(\Delta f(K^\top \mathbf{s}_t)) - \Delta f(K^\top \mathbf{s}_t) \right) - \left(\sum_{t=1}^{N} \log r_t! \right), \quad (2)$$

where $-(\sum \log r_t!)$ is a constant that does not depend on θ. The ML estimate for θ is simply the maximizer of the log-likelihood: $\hat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}_{lnp}(\theta; D)$.

Maximally informative dimensions (MID). The maximally informative dimensions (MID) estimator seeks to find an informative low-dimensional projection of the stimulus by maximizing an information-theoretic quantity known as the single-spike information [26]. This quantity, which we denote I_{ss}, is the the average information that the time of a single spike (considered independently of other spikes) carries about the stimulus

Although first introduced as a quantity that can be computed from the peri-stimulus time histogram (PSTH) measured in response to a repeated stimulus, the single-spike information can also be expressed as the Kullback-Leibler (KL) divergence between two distributions over the stimulus (see [26], appendix B):

$$I_{ss} = \int p(\mathbf{s}|spike) \log \frac{p(\mathbf{s}|spike)}{p(\mathbf{s})} d\mathbf{s} = D_{KL}\Big(p(\mathbf{s}|spike) \,\|\, p(\mathbf{s}) \Big), \quad (3)$$

where $p(\mathbf{s})$ denotes the marginal or "raw" distribution over stimuli, and $p(\mathbf{s}|spike)$ is the distribution over stimuli conditioned on observing a spike, also known as the "spike-triggered" stimulus distribution. Note that $p(\mathbf{s}|spike)$ is *not* the same as $p(\mathbf{s}|r = 1)$, the distribution of stimuli conditioned on a spike count of $r = 1$, since a stimulus that elicits two spikes will contribute twice as much to the spike-triggered distribution as a stimulus that elicits only one spike.

The MID estimator [18] seeks to find the linear projection that preserves maximal single-spike information:

$$I_{ss}(K) = D_{KL}\Big(p(K^\top \mathbf{s}|spike) \,\|\, p(K^\top \mathbf{s}) \Big), \quad (4)$$

where $p(K^\top \mathbf{s})$ and $p(K^\top \mathbf{s}|spike)$ are the raw and spike-triggered stimulus distributions projected onto the subspace defined by the columns of K, respectively. In practice, the MID estimator maximizes an estimate of the projected single-spike information:

$$\hat{K}_{MID} = \arg\max_{K} \hat{I}_{ss}(K), \quad (5)$$

where $\hat{I}_{ss}(K)$ denotes an empirical estimate of $I_{ss}(K)$. The columns of $\hat{K}_{MID}$ can be conceived as "directions" or "axes" in stimulus space that are most informative about a neuron's probability of spiking, as quantified by single-spike information. Fig. 2 shows a simulated example illustrating the MID estimate for a single linear filter in a two-dimensional stimulus space.

Equivalence of MID and maximum-likelihood LNP

Previous work has shown that MID converges asymptotically to the maximum-likelihood (ML) estimator for an LNP model in the limit of small time bins [20, 24]. Here we present a stronger result, showing that the equivalence is not merely asymptotic. We show that standard
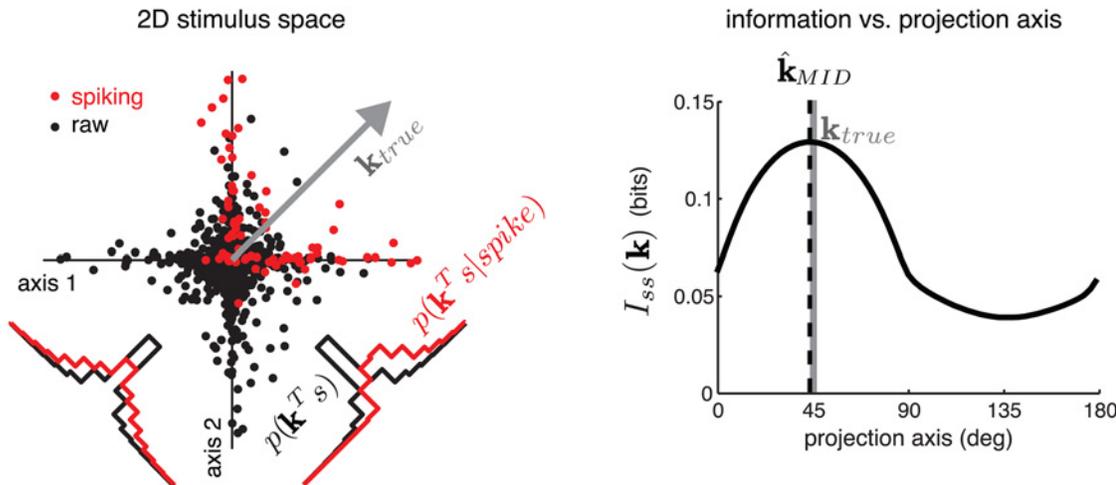
Fig 2. Geometric illustration of maximally-informative-dimensions (MID). Left: A two-dimensional stimulus space, with points indicating the location of raw stimuli (black) and spike-eliciting stimuli (red). For this simulated example, the probability of spiking depended only on the projection onto a filter $\mathbf{k}_{true}$, oriented at 45°. Histograms (inset) show the one-dimensional distributions of raw (black) and spike-triggered stimuli (red) projected onto $\mathbf{k}_{true}$ (lower right) and its orthogonal complement (lower left). **Right:** Estimated single-spike information captured by a 1D subspace, as a function of the axis of projection. The MID estimate $\hat{\mathbf{k}}_{MID}$ (dotted) corresponds to the axis maximizing single-spike information, which converges asymptotically to $\mathbf{k}_{true}$ with dataset size.

doi:10.1371/journal.pcbi.1004141.g002

MID, using histogram-based estimators for raw and spike-triggered stimulus densities $p(\mathbf{s})$ and $p(\mathbf{s}|spike)$, is exactly the ML estimator for the parameters of an LNP model, regardless of spike rate, the time bins used to count spikes, or the amount of data.

The standard implementation of MID [18, 20] uses histograms to estimate the projected stimulus densities $p(K^\top\mathbf{s})$ and $p(K^\top\mathbf{s}|spike)$. These density estimates are then used to compute $\hat{I}_{ss}(K)$, the plug-in estimate of single-spike information in a subspace defined by K (Equation 4). We will now unpack the details of this estimate in order to show its relationship to the LNP model log-likelihood.

Let $\{B_1, \ldots, B_m\}$ denote a group of sets ("histogram bins") that partition the range of the projected stimuli $K^\top\mathbf{s}$. In the one-dimensional case, we typically choose these sets to be intervals $B_i = [b_{i-1}, b_i)$, defined by bin edges $\{b_0, \ldots, b_m\}$, where $b_0 = -\infty$ and $b_m = +\infty$. Then let $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$ and $\hat{\mathbf{q}} = (\hat{q}_1, \ldots, \hat{q}_m)$ denote histogram-based estimates of $p(K^\top\mathbf{s})$ and $p(K^\top\mathbf{s}|spike)$, respectively, given by:

$$\hat{p}_i = \frac{\text{\# stimuli in } B_i}{\text{\# stimuli}} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{1}_{B_i}(\mathbf{x}_t)$$

$$\hat{q}_i = \frac{\text{\# stimuli in } B_i|spike}{\text{\# spikes}} = \frac{1}{n_{sp}}\sum_{t=1}^{N}\mathbf{1}_{B_i}(\mathbf{x}_t)r_t,$$

(6)

where $\mathbf{x}_t = K^\top\mathbf{s}_t$ denotes the linear projection of the stimulus $\mathbf{s}_t$, $n_{sp} = \sum_{t=1}^{N} r_t$ is the total number of spikes, and $\mathbf{1}_{B_i}(\cdot)$ is the indicator function for the set B_i, defined as:

$$\mathbf{1}_{B_i}(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in B_i \\ 0, & \mathbf{x} \notin B_i \end{cases}$$

(7)

The estimates $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are also known as the "plug-in" estimates, and correspond to maximum likelihood estimates for the densities in question. These estimates give us a plug-in

estimate for projected single-spike information:

$$\hat{I}_{ss} = \sum_{i=1}^{m} \hat{q}_i \log \frac{\hat{q}_i}{\hat{p}_i} = \frac{1}{n_{sp}} \sum_{i=1}^{m} \sum_{t=1}^{N} \mathbf{1}_{B_i}(\mathbf{x}_t) r_t \log \frac{\hat{q}_i}{\hat{p}_i} = \frac{1}{n_{sp}} \sum_{t=1}^{N} r_t \log \hat{g}(\mathbf{x}_t) \tag{8}$$

where the function $\hat{g}(\mathbf{x})$ denotes the ratio of density estimates:

$$\hat{g}(\mathbf{x}) \triangleq \sum_{i=1}^{m} \mathbf{1}_{B_i}(\mathbf{x}) \frac{\hat{q}_i}{\hat{p}_i}. \tag{9}$$

Note that $\hat{g}(\mathbf{x})$ is a piece-wise constant function that takes the value $\hat{q}_i/\hat{p}_i$ over the ith histogram bin B_i.

Now, consider an LNP model in which the nonlinearity f is parametrized as a piece-wise constant function, taking the value f_i over histogram bin B_i. Given a projection matrix K, the ML estimate for the parameter vector $\alpha = (f_1, \ldots, f_m)$ is the average number of spikes per stimulus in each histogram bin, divided by time bin width Δ, that is:

$$\hat{f}_i = \frac{1}{\Delta} \cdot \frac{\sum_{t=1}^{N} \mathbf{1}_{B_i}(\mathbf{x}_t) r_t}{\sum_{t=1}^{N} \mathbf{1}_{B_i}(\mathbf{x}_t)} = \left(\frac{n_{sp}}{N\Delta}\right) \frac{\hat{q}_i}{\hat{p}_i}. \tag{10}$$

Note that functions $\hat{f}$ and $\hat{g}$ are related by $\hat{f}(\mathbf{x}) = \left(\frac{n_{sp}}{N\Delta}\right) \hat{g}(\mathbf{x})$ and that the sum $\sum_{t=1}^{N} \hat{f}(\mathbf{x}_t)\Delta = n_{sp}$. We can therefore rewrite the LNP model log-likelihood (Equation 2):

$$
\begin{aligned}
\mathcal{L}_{lnp}(\theta; D) &= \sum_{t=1}^{N} r_t \log\left(\frac{n_{sp}}{N} \hat{g}(\mathbf{x}_t)\right) - n_{sp} - \sum_{t=1}^{N} \log r_t! \\
&= \sum_{t=1}^{N} r_t \log \hat{g}(\mathbf{x}_t) + n_{sp}\left(\log \frac{n_{sp}}{N} - 1\right) - \sum_{t=1}^{N} \log r_t!
\end{aligned}
\tag{11}
$$

This allows us to directly relate the empirical single-spike information (Equation 8) with the LNP model log-likelihood, normalized by the spike count as follows:

$$\hat{I}_{ss}(K) = \frac{1}{n_{sp}} \mathcal{L}_{lnp}(\theta; D) - \frac{1}{n_{sp}}\left[n_{sp} \log \frac{n_{sp}}{N} - n_{sp} - \sum \log r_t!\right] \tag{12}$$

$$= \frac{1}{n_{sp}} \mathcal{L}_{lnp}(\theta; D) - \frac{1}{n_{sp}} \mathcal{L}_{lnp}(\theta_0, D) \tag{13}$$

where $\mathcal{L}_{lnp}(\theta_0, D)$ denotes the Poisson log-likelihood under a "null" model in which spike rate does not depend on the stimulus, but takes constant rate $\lambda_0 = \frac{n_{sp}}{N\Delta}$ across the entire stimulus space. In fact, the quantity $-\mathcal{L}_{lnp}(\theta_0, D)$ can be considered an estimate for the marginal entropy of the response distribution, $H(r) = -\sum p(r) \log p(r)$, since it is the average log-probability of the response under a Poisson model, independent of the stimulus. This makes it clear that the single-spike information I_{ss} can be equally regarded as "LNP information".

Empirical single-spike information is therefore equal to LNP model log-likelihood per spike, plus a constant that does not depend on model parameters. This equality holds independent of time bin size Δ, the number of samples N and the number of spikes n_{sp}. From this relationship, it is clear that the linear projection K that maximizes $\hat{I}_{ss}$ also maximizes the LNP log-likelihood $\mathcal{L}_{lnp}(\theta; D)$, meaning that the MID estimate is the same as an ML estimate for the

filters in an LNP model:

$$\hat{K}_{MID} = \hat{K}_{ML}. \tag{14}$$

Moreover, the histogram-based estimates of the raw and spike-triggered stimulus densities $\hat{p}$ and $\hat{q}$, which are used for computing the empirical single-spike information $\hat{I}_{ss}$, correspond to a particular parametrization of the LNP model nonlinearity f as a piece-wise constant function over histogram bins. The ratio of these plug-in estimates gives rise to the ML estimate for f. MID is thus formally equivalent to an ML estimator for both the linear filters and the nonlinearity of an LNP model.

Previous literature has not emphasized that the MID estimator implicitly provides an estimate of the LNP model nonlinearity, or that the number of histogram bins corresponds to the number of parameters governing the nonlinearity. Selecting the number of parameters for the nonlinearity is important both for accurately estimating single-spike information from finite data and for successfully finding the most informative filter or filters. Fig. 3 illustrates this point using data from a simulated neuron with a single filter in a two-dimensional stimulus
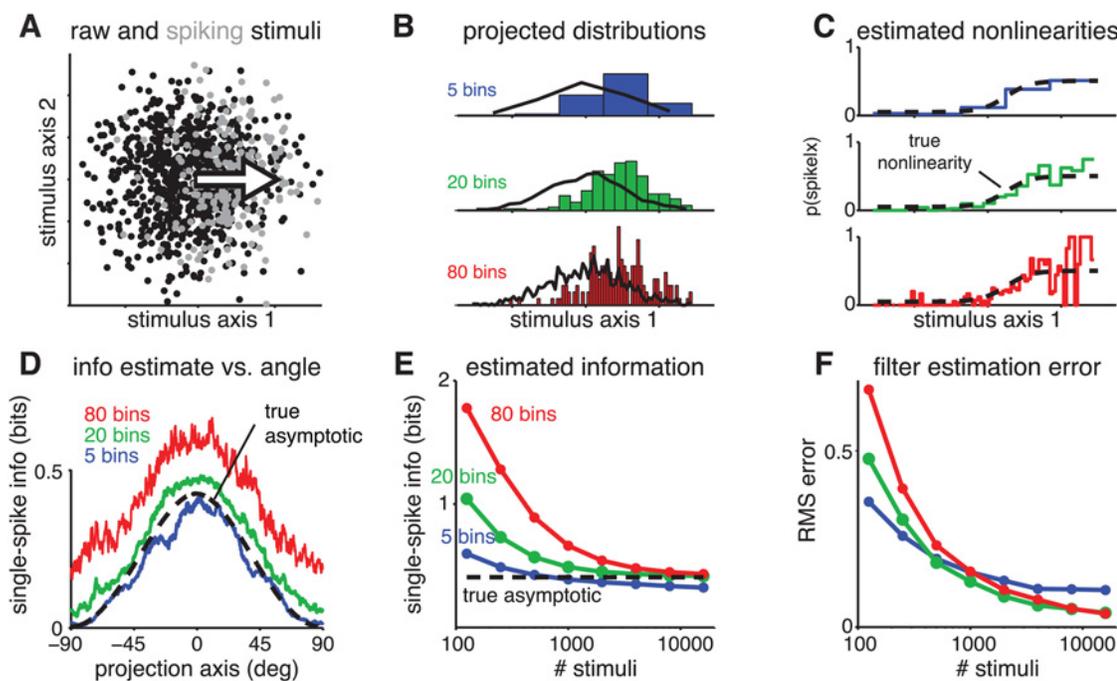
Fig 3. Effects of the number of histogram bins on empirical single-spike information and MID performance. (A) Scatter plot of raw stimuli (black) and spike-triggered stimuli (gray) from a simulated experiment using two-dimensional stimuli to drive a linear-nonlinear-Bernoulli neuron with sigmoidal nonlinearity. Arrow indicates the direction of the true filter $\mathbf{k}$. (B) Plug-In estimates of $p(\mathbf{k}^{\top}\mathbf{s}|spike)$, the spike-triggered stimulus distribution along the true filter axis, from 1000 stimuli and 200 spikes, using 5 (blue), 20 (green) or 80 (red) histogram bins. Black traces show estimates of raw distribution $p(\mathbf{k}^{\top}\mathbf{s})$ along the same axis. (C) True nonlinearity (black) and ML estimates of the nonlinearity (derived from the ratio of the density estimates shown in B). Roughness of the 80-bin estimate (red) arises from undersampling, or (equivalently) overfitting of the nonlinearity. (D) Empirical single-spike information vs. direction, calculated using 5, 20 or 80 histogram bins. Note that the 80-bin model overestimates the true asymptotic single-spike information at the peak by a factor of more than 1.5. (E) Convergence of empirical single-spike information along the true filter axis as a function of sample size. With small amounts of data, all three models overfit, leading to upward bias in estimated information. For large amounts of data, the 5-bin model underfits and therefore under-estimates information, since it lacks the smoothness to adequately describe the shape of the sigmoidal nonlinearity. (F) Filter error as a function of the number of stimuli, showing that the optimal number of histogram bins depends on the amount of data.

doi:10.1371/journal.pcbi.1004141.g003

space. For small datasets, the MID estimate computed with many histogram bins (i.e., many parameters for the nonlinearity) substantially overestimates the true I_{ss} and yields large errors in the filter estimate. Even with 1000 stimuli and 200 spikes, a 20-bin histogram gives substantial upward bias in the estimate of single-spike information (Fig. 3D). Parametrization of the nonlinearity is therefore an important problem that should be addressed explicitly when using MID, e.g., by cross-validation or other model selection methods.

Models with Bernoulli spiking

Under the discrete-time inhomogeneous Poisson model considered above, spikes are modeled as conditionally independent given the stimulus, and the spike count in a discrete time bin has a Poisson distribution. However, real spike trains may exhibit more or less variability than a Poisson process [27]. In particular, the Poisson assumption breaks down when the time bin in which the data are analyzed approaches the length of the refractory period, since in that case each bin can contain at most one spike. In that case, a Bernoulli model provides a more accurate description of neural data, since it allows only 0 or 1 spike per bin. The Bernoulli and discrete-time Poisson models approach the same limiting Poisson process as the bin size (and single-bin spike probability) approaches zero while the average spike rate remains constant. However, as long as single-bin spike probabilities are above zero, the two models differ.

Here we show that the standard "Poisson" MID estimator does not necessarily maximize information between stimulus and response when spiking is non-Poisson. That is, if the spike count r given stimulus $\mathbf{s}$ is not a Poisson random variable, then MID does not necessarily find the subspace preserving maximal information between stimulus and response. To show this, we derive the mutual information between the stimulus and a Bernoulli distributed spike count, and show that this quantity is closely related to the log-likelihood under a linear-nonlinear-Bernoulli encoding model.

Linear-nonlinear-Bernoulli (LNB) model. We can define the linear-nonlinear-Bernoulli (LNB) model by analogy to the LNP model, but with Bernoulli instead of Poisson spiking. The parameters $\theta = \{K, \alpha\}$ consist of a matrix K that determines a linear projection of the stimulus space, and a set of parameters α that govern the nonlinearity f. Here, the output of f is spike probability λ in the range $[0, 1]$. The probability of a spike response $r \in \{0,1\}$ given stimulus $\mathbf{s}$ is governed by a Bernoulli distribution. We can express this model as

$$\lambda = f(K^{\top}\mathbf{s}) \tag{15}$$

$$p(r|\lambda) = \lambda^r (1 - \lambda)^{1-r}, \tag{16}$$

and the log-likelihood for a dataset $D = \{(\mathbf{s}_t, r_t)\}$ is

$$\mathcal{L}_{lnb}(\theta; D) = \sum_{t=1}^{N} \Big(r_t \log\left(f(K^{\top}\mathbf{s}_t)\right) - (1 - r_t) \log\left(1 - f(K^{\top}\mathbf{s}_t)\right) \Big). \tag{17}$$

If K has a single filter and the nonlinearity is restricted to be a logistic function, $f(x) = 1/(1+\exp(-x))$, this reduces to the logistic regression model. Note that the spike probability λ is analogous to the single-bin Poisson rate $\lambda\Delta$ from the LNP model (Equation 1), and the two models become identical in the small-bin limit where the probability of spiking $p(r = 1)$ goes to zero [24, 26].

Bernoulli information. We can derive an equivalent dimensionality-reduction estimator in information-theoretic terms. The mutual information between the projected stimulus

$\mathbf{x} = K^{\top}\mathbf{s}$ and a Bernoulli spike response $r \in \{0,1\}$ is given by:

$$
\begin{aligned}
I(\mathbf{x}, r) &= H(\mathbf{x}) - H(\mathbf{x}|r) \\
&= -\int d\mathbf{x}\, p(\mathbf{x}) \log p(\mathbf{x}) + \sum_{j\in\{0,1\}} p(r=j)\int d\mathbf{x}\, p(\mathbf{x}|r=j) \log p(\mathbf{x}|r=j) \\
&= \sum_{j\in\{0,1\}} p(r=j)\int d\mathbf{x}\, p(\mathbf{x}|r=j) \log \frac{p(\mathbf{x}|r=j)}{p(\mathbf{x})} \\
&= \sum_{j\in\{0,1\}} p(r=j) D_{KL}\Big(p(\mathbf{x}|r=j) \,\|\, p(\mathbf{x})\Big).
\end{aligned}
\tag{18}
$$

If we normalize by the probability of observing a spike, we obtain a quantity with units of bits-per-spike that can be directly compared to single-spike information. We refer to this as the Bernoulli information:

$$
I_{Ber} = \frac{1}{p(r=1)} I(\mathbf{x}, r) = I_0 + I_{ss}
\tag{19}
$$

where $I_0 = \frac{p(r=0)}{p(r=1)} D_{KL}\Big(p(\mathbf{x}|r=0) \,\|\, p(\mathbf{x})\Big)$ is the information (per spike) carried by silences, and I_{ss} is the single-spike information (Equation 4). Thus, where single-spike information quantifies the information conveyed by each spike alone (no matter how many spikes might co-occur in the same time bin) but neglects the information conveyed by silences, the Bernoulli information reflects the information conveyed by both spikes and silences.

Let $\hat{I}_{Ber} = \hat{I}_0 + \hat{I}_{ss}$ denote the empirical or plug-in estimate of the Bernoulli information, where $\hat{I}_{ss}$ is the empirical single-spike information (Equation 8), and $\hat{I}_0$ is a plug-in estimate of the KL divergence between $p(\mathbf{x}|r=0)$ and $p(\mathbf{x})$, weighted by $(N-n_{sp})/n_{sp}$, the ratio of the number of silences to the number of spikes. It is straightforward to show that empirical Bernoulli information equals the LNB model log-likelihood per spike plus a constant:

$$
\hat{I}_{Ber} = \frac{1}{n_{sp}}\mathcal{L}_{lnb} + \frac{1}{\bar{r}}\hat{H}[r]
\tag{20}
$$

where $\bar{r} = \frac{n_{sp}}{N}$ denotes mean spike count per bin and $\hat{H}[r] = -\frac{n_{sp}}{N}\log\frac{n_{sp}}{N} - \frac{N-n_{sp}}{N}\log\frac{N-n_{sp}}{N}$ is the plug-in estimate for the marginal response entropy. Because the second term is independent of θ, the maximum of the empirical Bernoulli information is identical to the maximum of the LNB model likelihood, meaning that once again, we have an exact equivalence between likelihood-based and information-based estimators.

Failure modes for MID under Bernoulli spiking. The empirical Bernoulli information is strictly greater than the estimated single-spike (or "Poisson") information for a binary spike train that is not all zeros or ones, since $\hat{I}_0 > 0$ and these spike absences are neglected by the single-spike information measure. Only in the limit of infinitesimal time bins, where $p(r=1) \to 0$, does $\hat{I}_{Ber}$ converge to $\hat{I}_{ss}$[24, 26]. As a result, standard MID can fail to identify the most informative subspace when applied to a neuron with Bernoulli spiking. We illustrate this phenomenon with two (admittedly toy) simulated examples. For both examples, we compute the standard MID estimate $\hat{\mathbf{k}}_{MID}$ by maximizing $\hat{I}_{ss}$, and the LNB filter estimate $\hat{\mathbf{k}}_{Ber}$ which maximizes the LNB likelihood.

The first example (Fig. 4) uses stimuli uniformly distributed on the right half of the unit circle. The Bernoulli spike probability λ increases linearly as a function of stimulus angle: $\lambda = (\mathbf{s}-\pi/2)/\pi$, for $\mathbf{s} \in (-\pi/2,\pi/2)$. For this neuron, the most informative 1D axis is the vertical
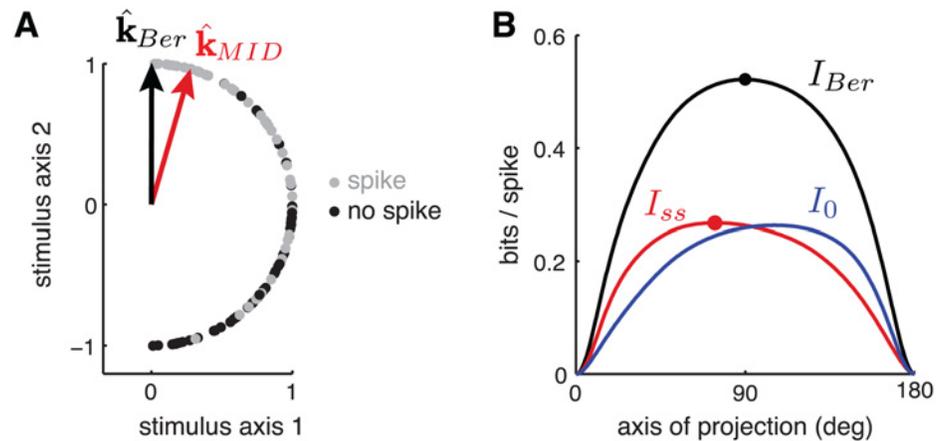
Fig 4. Illustration of MID failure mode due to non-Poisson spiking. (A) Stimuli were drawn uniformly on the unit half-circle, $\theta \sim$ Unif$(-\pi/2, \pi/2)$. The simulated neuron had Bernoulli (i.e., binary) spiking, where the probability of a spike increased linearly from 0 to 1 as θ varied from $-\pi/2$ to $\pi/2$, that is: $p(spike|\theta) = \theta/\pi + 1/2$. Stimuli eliciting "spike" and "no-spike" are indicated by gray and black circles, respectively. For this neuron, the most informative one-dimensional linear projection corresponds to the vertical axis ($\hat{\mathbf{k}}_{Ber}$), but the MID estimator ($\hat{\mathbf{k}}_{MID}$) exhibits a $16°$ clockwise bias. **(B)** Information from spikes (black), silences (gray), and both (red), as a function of projection angle. The peak of the Bernoulli information (which defines $\hat{\mathbf{k}}_{Ber}$) lies close to $\pi/2$, while the peak of single-spike information (which defines $\hat{\mathbf{k}}_{MID}$) exhibits the clockwise bias shown in A. Note that $\hat{\mathbf{k}}_{MID}$ does not converge to the optimal direction even in the limit of infinite data, due to its lack of sensitivity to information from silences. Although this figure is framed in an information-theoretic sense, equations (19) and (20) detail the equivalence between I_{Ber} and $\mathcal{L}_{Inb}$, so that this figure can be viewed from either an information-theoretic or likelihood-based perspective.

doi:10.1371/journal.pcbi.1004141.g004

axis, which is closely matched by the estimate $\hat{\mathbf{k}}_{Ber}$. By contrast, $\hat{\mathbf{k}}_{MID}$ exhibits a substantial clockwise bias, resulting from its failure to take into account the information from silences (which are more informative when spike rate is high). [Fig. 4B](#) shows the breakdown of total Bernoulli information into $\hat{I}_{ss}$ (spikes) and $\hat{I}_0$ (silences) as a function of projection angle, which illustrates the relative biases of the two quantities.

A second example ([Fig. 5](#)) uses stimuli drawn from a standard bivariate Gaussian (0 mean and identity covariance), in which standard MID makes a 90° error in identifying the most informative one-dimensional subspace. The neuron's nonlinearity ([Fig. 5A](#)) is excitatory in stimulus axis s_1 and suppressive in stimulus axis s_2 (indicating that a large projection onto s_1 increases spike probability, while a large projection onto s_2 decreases spike probability). For this neuron, both stimulus axes are clearly informative, but the (suppressive, vertical) axis s_2 carries 13% more information than the (excitatory, horizontal) axis s_1. However, the standard MID estimator identifies s_1 as the most informative axis ([Fig. 5C](#)), due once again to the failure to account for the information carried by silences.

These artificial examples were designed to emphasize the information carried by missing spikes, and we do not expect such stark differences between Bernoulli and Poisson estimators to arise in typical neural data. However, it is clear that the assumption of Poisson firing can lead the standard MID estimator to make mistakes when spiking is actually Bernoulli. In general, we suggest that the question of which estimator performs better is an empirical one, and depends on which model describes the true spiking process more accurately.

Quantifying MID information loss for binary spike trains. In the limit of infinitesimal time bins, the information carried by silences goes to zero, and the plug-in estimates for
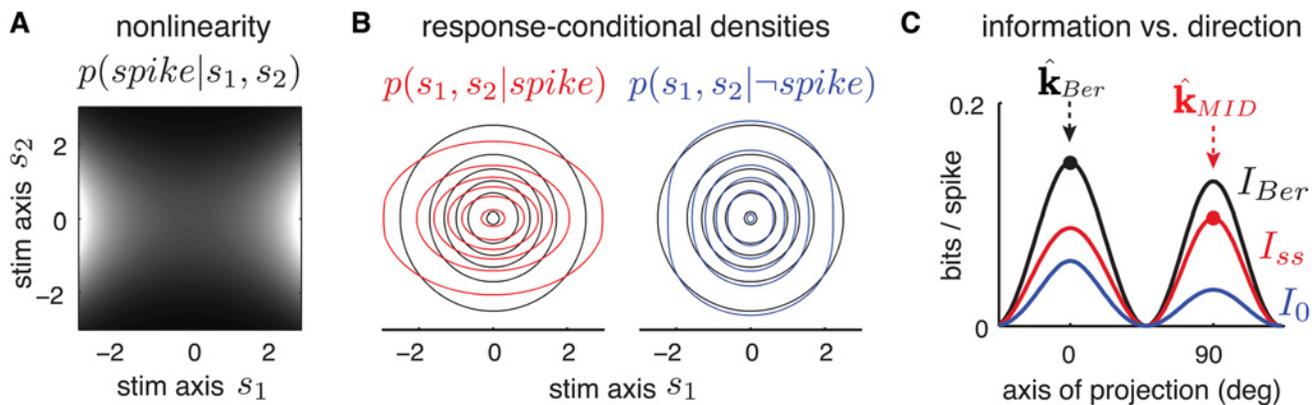
Fig 5. A second example Bernoulli neuron for which $\hat{\mathbf{k}}_{MID}$ fails to identify the most-informative one-dimensional subspace. The stimulus space has two dimensions, denoted s_1 and s_2, and stimuli were drawn *iid* from a standard Gaussian $\mathcal{N}(0,1)$. **(A)** The nonlinearity $f(s_1,s_2) = p(spike|s_1,s_2)$ is excitatory in s_1 and suppressive in s_2; brighter intensity indicates higher spike probability. **(B)** Contour plot of the stimulus-conditional densities given the two possible responses: "spike" (red) or "no-spike" (blue), along with the raw stimulus distribution (black). **(C)** Information carried by silences (I_0), single spikes (I_{ss}), and total Bernoulli information ($I_{Ber} = I_0 + I_{ss}$) as a function of subspace orientation. The MID estimate $\hat{\mathbf{k}}_{MID} = 90°$ is the maximum of I_{ss}, but the total Bernoulli information is in fact 13% higher at $\hat{\mathbf{k}}_{Ber} = 0°$ due to the incorporation of no-spike information. Although both stimulus axes are clearly relevant to the neuron, MID identifies the less informative one. As with the previous figure, equations (19) and (20) detail the equivalence between I_{Ber} and $\mathcal{L}_{lnb}$, so that this figure can be viewed from either an information-theoretic or likelihood-based perspective.

Bernoulli and single-spike information converge: $I_0 \rightarrow 0$ and $\hat{I}_{Ber} \rightarrow \hat{I}_{ss}$. However, for finite time bins, the Bernoulli information can substantially exceed single-spike information.

In the previous section, we showed that this mismatch can lead to errors in subspace identification. Here we derive a lower bound on the information lost due to the neglect of I_0, the information (per spike) carried by silences, as a function of marginal probability of a spike, $p(r = 1)$.

In the limit of rare spiking, $p(r = 1) \rightarrow 0$, we find that:

$$\frac{I_0}{I_{Ber}} = \frac{I_0}{I_0 + I_{ss}} \geq \frac{p(r = 1)}{2}. \tag{21}$$

The fraction of lost information is at least half the marginal spike probability. Thus, for example, if 20% of the bins in a binary spike train contain a spike, the standard MID estimator will necessarily neglect at least 10% of the total mutual information. We show this bound holds in the asymptotic limit of small $p(r = 1)$ (see Methods for details), but conjecture that it holds for all $p(r = 1)$. The bound is tight in the Poisson limit, $p(r = 1) \rightarrow 0$, but is substantially loose in the limit where spiking is common $p(r = 1) \rightarrow 1$, in which all information is carried by silences. Fig 6 shows our bound compared to the actual (numerical) lower bound for an example with a binary stimulus.

Models with arbitrary spike count distributions

For neural responses binned at the stimulus refresh rate (e.g., 100 Hz), it is not uncommon to observe multiple spikes in a single bin. For the general case, then, we must consider an arbitrary distribution over counts conditioned on a stimulus. As we will see, maximizing the mutual information based on histogram estimators is once again equivalent to maximizing the likelihood of an LN model with piece-wise constant mappings from the linear stimulus projection to count probabilities.

Linear-nonlinear-count (LNC) model. Suppose that a neuron responds to a stimulus **s** with a spike count $r \in \{0, \ldots, r_{max}\}$, where r_{max} is the maximum possible number of spikes

minimum information loss for binary spiking
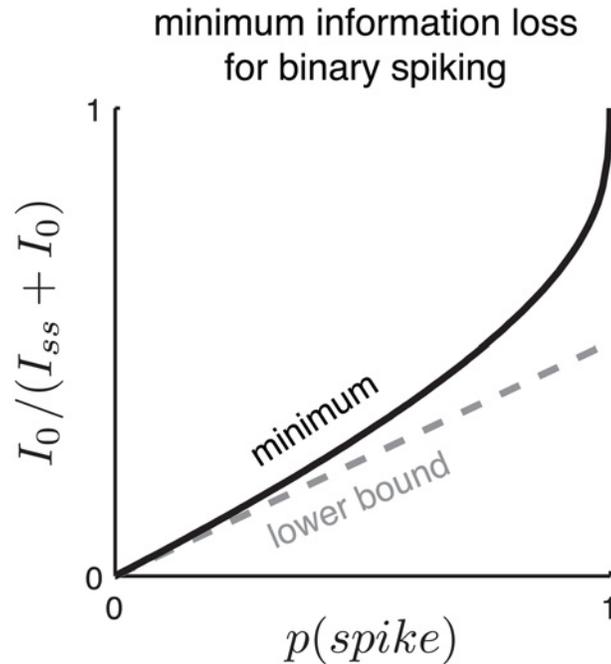
Fig 6. Lower bound on the fraction of total information neglected by MID for a Bernoulli neuron, as a function of the marginal spike probability $p(spike) = p(r = 1)$, for the special case of a binary stimulus. Information loss is quantified as the ratio $I_0/(I_0+I_{ss})$, the information due to no-spike events, I_0, divided by the total information due to spikes and silences, I_0+I_{ss}. The dashed gray line shows the lower bound derived in the limit $p(spike) \to 0$. The solid black line shows the actual minimum achieved for binary stimuli $s \in \{0,1\}$ with $p(s = 1) = q$, computed via a numerical search over the parameter $q \in [0, 1]$ for each value of $p(spike)$. The lower bound is substantially loose for $p(spike) > 0$, since as $p(spike) \to 1$, the fraction of information due to silences goes to 1.

within the time bin (constrained by the refractory period or firing rate saturation). The linear-nonlinear-count (LNC) model, which includes LNB as a special case, is defined by a linear dimensionality reduction matrix K and a set of nonlinear functions $\{f^{(0)}\ldots, f^{(r_{max})}\}$ that map the projected stimulus to the probability of observing $\{0, \ldots, r_{max}\}$ spikes, respectively. We can write the probability of a spike response r given projected stimulus $\mathbf{x} = K^\top \mathbf{s}$ as:

$$
\begin{aligned}
\lambda^{(j)} &= f^{(j)}(\mathbf{x}), \quad \text{for } j = 0, \ldots, r_{max} \\
p(r = j|\lambda^{(j)}) &= \lambda^{(j)}.
\end{aligned}
\tag{22}
$$

Note that there is a constraint on the functions f requiring that $\sum_j f^{(j)}(\mathbf{x}) = 1, \forall \mathbf{x}$, since the probabilities over possible counts must add to 1 for each stimulus.

The LNC model log-likelihood for parameters $\theta = (K, \alpha^{(0)}, \ldots \alpha^{(r_{max})})$ given data $D = \{(\mathbf{s}_t, r_t)\}$ can be written:

$$
\mathcal{L}_{lnc}(\theta; D) = \sum_{t=1}^{N} \sum_{j=0}^{r_{max}} \left(\mathbf{1}_j(r_t) \log \left(f^{(j)}(K^\top \mathbf{s}_t) \right) \right),
\tag{23}
$$

where $\mathbf{1}_j(r_t)$ is an indicator function selecting time bins t in which the spike count is j. As before, we consider the case where $f^{(j)}$ takes a constant value in each of m histogram bins $\{B_i\}$, so that the parameters are just those constant values: $\alpha^{(j)} = (f_0^{(j)}, \ldots, f_m^{(j)})$. The maximum-

likelihood estimates for the values can be given in terms of the histogram probabilities:

$$\hat{f}_i^{(j)} \;=\; \frac{n_i^{(j)}}{n_i} \;=\; \frac{\hat{q}_i^{(j)}}{\hat{p}_i} \frac{N^{(j)}}{N} \,. \tag{24}$$

where n_i is the number of stimuli in bin B_i, $n_i^{(j)}$ is the number of stimuli in bin B_i that elicited j spikes, $N^{(j)}$ is the number of stimuli in all bins that elicited j spikes, and N is the total number of stimuli. The histogram fractions of the projected raw spike counts $\hat{p}_i$ are defined as in Equation 6, with the j-spike conditioned histograms defined analogously:

$$\hat{q}_i^{(j)} \;=\; \frac{1}{N^{(j)}} \sum_t \mathbf{1}_{B_i}(\mathbf{x}_t)\, \mathbf{1}_j(r_t) \;=\; \frac{n_i^{(j)}}{N^{(j)}} \,, \tag{25}$$

Thus, the log-likelihood for projection matrix K, having already maximized with respect to the nonlinearities by using their plug-in estimates, is

$$\mathcal{L}_{lnc}(K; D) \;=\; \sum_{t=1}^{N} \sum_{j=0}^{r_{\max}} \left(\mathbf{1}_j(r_t)\, \log\left(\hat{f}^{(j)}(K^\top \mathbf{s}_t) \right) \right) \tag{26}$$

$$=\; \sum_{t=1}^{N} \sum_{j=0}^{r_{\max}} \sum_{i=1}^{m} \left(\mathbf{1}_j(r_t)\, \mathbf{1}_{B_i}(K^\top \mathbf{s}_t)\, \log\left(\hat{f}_i^{(j)} \right) \right) \tag{27}$$

$$=\; \sum_{j=0}^{r_{\max}} \sum_{i=1}^{m} \left(n_i^{(j)} \log \frac{n_i^{(j)}}{n_i} \right) \tag{28}$$

$$=\; \sum_{j=0}^{r_{\max}} \sum_{i=1}^{m} \left(N^{(j)} \hat{q}_i^{(j)} \log \left(\frac{\hat{q}_i^{(j)}}{\hat{p}_i} \frac{N^{(j)}}{N} \right) \right) \tag{29}$$

$$=\; \sum_{j=0}^{r_{\max}} \sum_{i=1}^{m} \left(N^{(j)} \hat{q}_i^{(j)} \log \left(\frac{\hat{q}_i^{(j)}}{\hat{p}_i} \right) \right) + \sum_{j=0}^{r_{\max}} \left(N^{(j)} \log \left(\frac{N^{(j)}}{N} \right) \right) . \tag{30}$$

Information in spike counts. If the binned spike-counts r_t measured in response to stimuli $\mathbf{s}_t$ are not Poisson distributed, the projection matrix K which maximizes the mutual information between $K^\top \mathbf{s}$ and r can be found as follows. Recalling that $r_{\max}$ is the maximal spike count possible in the time bin and writing $\mathbf{x} = K^\top \mathbf{s}$, we have:

$$I(\mathbf{x}, r) \;=\; H(\mathbf{x}) - H(\mathbf{x}|r) \tag{31}$$

$$=\; -\int d\mathbf{x}\, p(\mathbf{x}) \log p(\mathbf{x}) + \sum_{j=0}^{r_{\max}} p(r=j) \int d\mathbf{x}\, p(\mathbf{x}|r=j) \log p(\mathbf{x}|r=j) \tag{32}$$

$$=\; \sum_{j=0}^{r_{\max}} p(r=j) \int d\mathbf{x}\, p(\mathbf{x}|r=j) \log \frac{p(\mathbf{x}|r=j)}{p(\mathbf{x})} \tag{33}$$

$$=\; \sum_{j=0}^{r_{\max}} p(r=j) D_{KL}\Big(p(\mathbf{x}|r=j) \,\|\, p(\mathbf{x}) \Big) . \tag{34}$$

To ease comparison with the single-spike information, which is measured in bits per spike, we normalize the mutual information by the mean spike count to obtain:

$$I_{count} = \frac{1}{\bar{r}} I(\mathbf{x}, r) = I_0 + I_1 + \cdots + I_{r_{max}} \tag{35}$$

where $\bar{r} = \sum_t r_t/N$ is the mean spike count, and $I_j = \frac{p(r=j)}{\bar{r}} D_{KL}(p(\mathbf{x}|r=j) \| p(\mathbf{x}))$ is the normalized information carried by the j-spike responses. Note that I_1, the information carried by single-spike responses, is *not* the same as the single-spike information I_{ss}, since the latter combines information from all responses with 1 or more spikes, by assuming that each spike is conditionally independent of all other spikes.

We can estimate the mutual information from data using a histogram based plug-in estimator:

$$\hat{I}_{count} = \sum_{j=0}^{r_{max}} \frac{1}{\bar{r}} \frac{N^{(j)}}{N} \sum_{i=1}^{m} \hat{q}_i^{(j)} \log \frac{\hat{q}_i^{(j)}}{\hat{p}_i} . \tag{36}$$

Comparison with the LNC model log-likelihood (Equation 30) reveals that:

$$\hat{I}_{count} = \frac{1}{n_{sp}} \mathcal{L}_{lnc} + \frac{1}{\bar{r}} \hat{H}[r] \tag{37}$$

where $\hat{H}[r] = -\sum_{j=0}^{r_{max}} \frac{N^{(j)}}{N} \log \frac{N^{(j)}}{N}$ is the plug-in estimate for the marginal entropy of the observed spike counts. Note that this reduces to the relationship between Bernoulli information and LNB model log-likelihood (Equation 20) in the special case where $r_{max} = 1$.

Thus, even in the most general case of an arbitrary distribution over spike counts given a stimulus, the subspace projection K that maximizes the histogram-based estimate of mutual information is identical to the maximum-likelihood K for an LN model with a corresponding piece-wise constant parametrization of the nonlinearities.

Failures of MID under non-Poisson count distributions. We formulate two simple examples to illustrate the sub-optimality of standard MID for neurons whose stimulus-conditioned count distributions are not Poisson. For both examples, the neuron was sensitive to a one-dimensional projection along the horizontal axis and emitted either 0, 1, or 2 spikes in response to a stimulus.

Both are illustrated in Fig. 7. The first example (A) involves a deterministic neuron, where spike count is 0, 1, or 2 according to a piece-wise constant nonlinear function of the projected stimulus. Here, MID does not use the information from zero or two-spike bins optimally; it ignores information from zero-spike responses entirely, and treats stimuli eliciting two spikes as two independent samples from $p(\mathbf{x}|spike)$. The I_{count} estimator, by contrast, is sensitive to the non-Poisson statistics of the response, and combines information from all spike counts (Equation 35), yielding both higher information and faster convergence to the true filter.

Our second example (Fig. 7B) involves a model neuron in which a sigmoidal nonlinearity determines the probability that it fires exactly 1 spike (high at negative stimulus projections) or stochastically emits either 0 or 2 spikes, each with probability 0.5 at positive stimulus projections). Thus, the nonlinearity does not change the mean spike rate, but strongly affects its variance. Because the probability of observing a single spike is not affected by the stimulus, single-spike information is zero for all projections, and the MID estimate does not converge to the true filter even with infinite data. However, the full count information $\hat{I}_{count}$ correctly weights the information carried by different spike counts and provides a consistent estimator for K.
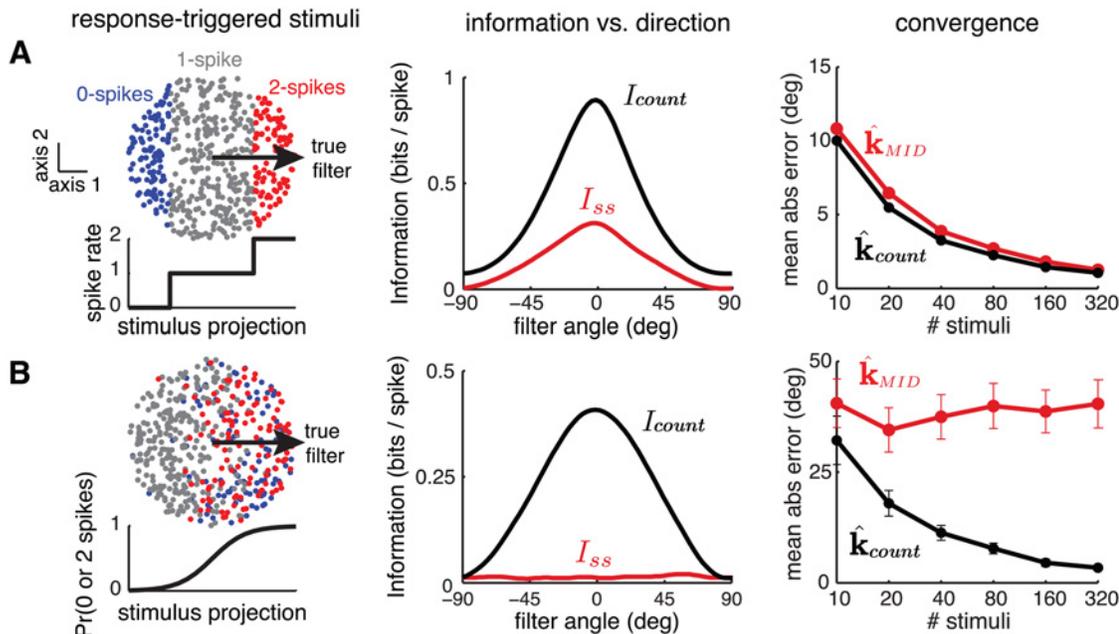
Fig 7. Two examples illustrating sub-optimality of MID under discrete (non-Poisson) spiking. In both cases, stimuli were uniformly distributed within the unit circle and the simulated neuron's response depended on a 1D projection of the stimulus onto the horizontal axis ($\theta = 0$). Each stimulus evoked 0, 1, or 2 spikes. **(A)** Deterministic neuron. *Left*: Scatter plot of stimuli labelled by number of spikes evoked, and the piece-wise constant nonlinearity governing the response (below). The nonlinearity sets the response count deterministically, thus dramatically violating Poisson expectations. *Middle*: information vs. axis of projection. The total information I_{count} reflects the information from 0-, 1-, and 2-spike responses (treated as distinct symbols), while the single-spike information I_{ss} ignores silences and treats 2-spike responses as two samples from $p(s|spike)$. *Right*: Average absolute error in $\hat{\mathbf{k}}_{MID}$ and $\hat{\mathbf{k}}_{count}$ as a function of sample size; the latter achieves 18% lower error due to its sensitivity to the non-Poisson structure of the response. **(B)** Stochastic neuron with sigmoidal nonlinearity controlling the stochasticity of responses. The neuron transitions from almost always emitting 1 spike for large negative stimulus projections, to generating either 0 or 2 spikes with equal probability at large positive projections. Here, the nonlinearity does not modulate the mean spike rate, so $\hat{I}_{ss}$ is approximately zero for all stimulus projections (middle) and the MID estimator does not converge (right). However, the $\hat{\mathbf{k}}_{count}$ estimate converges because the LNC model is sensitive to the change in conditional response distribution. Equation (37) details the relationship between I_{count} and $\mathcal{L}_{lnc}$, so that this figure can be interpreted from either an information-theoretic or likelihood-based perspective.

doi:10.1371/journal.pcbi.1004141.g007

Identifying high-dimensional subspaces

A significant drawback to standard MID is that it does not scale tractably to high-dimensional subspaces; that is, to the simultaneous estimation of many filters. MID has usually been limited to estimation of only one or two filters, and we are unaware of a practical setting in which it has been used to recover more than three. This stands in contrast to methods like spike-triggered covariance (STC) [1, 7], information-theoretic spike-triggered average and covariance (iSTAC) [19], projection-pursuit regression [28], Bayesian spike-triggered covariance [14], and quadratic variants of MID [21, 22], all of which can tractably estimate ten or more filters. This capability may be important, given that V1 neurons exhibit sensitivity to as many as 15 dimensions [29], and many canonical neural computations (e.g., motion estimation) require a large number of stimulus dimensions [22, 30].

Before we continue, it is helpful to consider *why* MID is impractical for high-dimensional feature spaces. The problem isn't the number of filter parameters: these scale linearly with dimensionality, since a p-filter model with D-dimensional stimuli requires only Dp parameters, or indeed only $(D-1)p - \frac{1}{2}p(p-1)$ parameters to specify the subspace after accounting for degeneracies. The problem is instead the number of parameters needed to specify the densities $p(\mathbf{x})$ and $p(\mathbf{x}|spike)$. For histogram-based density estimators, the number of parameters grows

exponentially with dimension: a histogram with m bins along each of p filter axes requires m^p parameters, a phenomenon sometimes called the "curse of dimensionality".

Density vs. nonlinearity estimation.   A key benefit of the LNP model likelihood framework is that it shifts the focus of estimation away from the separate densities $p(\mathbf{x}|spike)$ and $p(\mathbf{x})$ to a single nonlinear function f. This change in focus makes it easier to scale the likelihood approach to high dimensions for a few different reasons. First, direct estimation of a single nonlinearity in place of two densities immediately halves the number of parameters required to achieve a similarly detailed picture of the neuron's response to the filtered stimulus. Second, the dependence of the MID cost function on the logarithm of the ratio $p(\mathbf{x}|spike)/p(\mathbf{x})$ makes it very sensitive to noise in the estimated value of the denominator $p(\mathbf{x})$ when that value is near 0. Unfortunately, as $p(\mathbf{x})$ is also the probability with which samples are generated, these low-value regions are precisely where the fewest samples are available. This is a common difficulty in the empirical estimation of information-theoretic quantities, and others working in more general machine-learning settings have suggested direct estimation of the ratio rather than its parts [31–33]. In LN neural modeling such direct estimation of the ratio is equivalent to direct estimation of the nonlinearity.

As the nonlinearity is a property of the neuron rather than the stimulus, it may be more straightforward to construct a smoothed or structured parametrization for f (or to regularize its estimate based on prior beliefs about neuronal properties) regularizing the estimates of stimulus densities. For example, consider an experiment using natural visual images. While natural images presumably form a smooth manifold within the space of all possible pixel patterns, the structure of this manifold is neither simple nor known. The natural distribution of images does not factor over disjoint sets of pixels, nor over linear projections of pixel values. A small random perturbation in all pixels makes a natural image appear unnaturally noisy, violating the underlying presumption of kernel density estimators that local perturbations do not alter the density much. Indeed the question of how best to model the distribution of natural stimuli is a matter of active research. By contrast, we might expect to be able to develop better parametric forms to describe the non-linearities employed by neural systems. For instance, we might expect the neural nonlinearity to vary smoothly in the space of photoreceptor activation, and thus of filter outputs. Thus, locally kernel-smoothed estimates of the non-linear mapping—or even parametric choices of function class, such as low-order polynomials—might be valid, even if the stimulus density changes abruptly. Alternatively, subunits within the neural receptive field might lead to additively or multiplicatively separable components of the nonlinearity that act on the outputs of different filters. In this case, it would be possible to factor f between two subsets of filter outputs, say to give $f(\mathbf{x}) = f_1(\mathbf{x}_1)f_2(\mathbf{x}_2)$, even though there is no reason for the stimulus distribution to factor: $p(\mathbf{x}) \neq p(\mathbf{x}_1)p(\mathbf{x}_2)$. This reduction of f to two (or more) lower-dimensional functions would avoid the exponential parameter explosion implied by the curse of dimensionality.

Indeed, such strategies for parametrization of the nonlinear mapping are already implicit in likelihood-based estimators inspired by the spike-triggered average and covariance. In many such cases, f is parametrized by a quadratic form embedded in a 1D nonlinearity [14], so that the number of parameters scales only quadratically with the number of filters. A similar approach has been formulated in information-theoretic terms using a quadratic logistic Bernoulli model [21, 22, 24]. Another method, known as extended projection pursuit regression (ePPR) [28], parametrizes f as a sum of one-dimensional nonlinearities, in which case the number of parameters grows only linearly with the number of filters.

Parametrizing the many-filter LNP model.   Here we provide a general formulation that encompasses both standard MID and constrained methods that scale to high-dimensional

subspaces. We can rewrite the LNP model ([Equation 1](#)) as follows:

$$\mathbf{x} = K^\top \mathbf{s} \qquad \text{(dimensionality reduction)} \tag{38}$$

$$\lambda = f(\mathbf{x}) = g\left(\sum_{i=1}^{n_\phi} \alpha_i \phi_i(\mathbf{x})\right) \qquad \text{(nonlinearity)} \tag{39}$$

$$r|\lambda \sim \text{Poiss}(\lambda\Delta) \qquad \text{(spiking).} \tag{40}$$

The nonlinearity f is parametrized using basis functions $\{\varphi_i(\cdot)\}$, $i = 1, \ldots, n_\varphi$, which are linearly combined with weights α_i and then passed through a scalar nonlinearity g. We refer to g as the output nonlinearity; its primary role is to ensure the spike rate λ is positive regardless of weights α_i. This can also be considered a special case of an LNLN model [15, 34, 35].

If we fix g and the basis functions $\{\varphi_i\}$ in advance, fitting the nonlinearity simply involves estimating the parameters α_i from the projected stimuli and associated spike counts. If g is convex and log-concave, then the log-likelihood is concave in $\{\alpha_i\}$ given K, meaning the parameters governing f can be fit without getting stuck in non-global maxima [11].

Standard MID can be seen as a special case of this general framework: it sets g to the identity function and the basis functions φ_i to histogram-bin indicator functions (denoted $\mathbf{1}_{B_i}(\cdot)$ in [Equation 7](#)). The maximum-likelihood weights $\{\hat{\alpha}_i\}$ are proportional to the ratio between the number of spikes and number of stimuli in the i'th histogram bin ([Equation 10](#)). As discussed above, the number of basis functions n_φ scales exponentially with the number of filters, making this parametrization impractical for high-dimensional feature spaces.

Another special case of this framework corresponds to Bayesian spike-triggered covariance analysis [14], in which the basis functions φ_i are taken to be linear and quadratic functions of the projected stimulus. If the stimulus is Gaussian, then standard STC and iSTAC provide an asymptotically optimal fit to this model under the assumption that g is exponential [14, 19].

In principle, we can select any set of basis functions. Other reasonable choices include polynomials, sigmoids, sinusoids (i.e., Fourier components), cubic splines, radial basis functions, or any mixture of these bases. Alternatively, we could use non-parametric models such as Gaussian processes, which have been used to model low-dimensional tuning curves and firing rate maps [36, 37]. Theoretical convergence for arbitrary high-dimensional nonlinearities requires a scheme for increasing the complexity of the basis or non-parametric model as we increase the amount of data recorded [38–41]. We do not examine such theoretical details here, focusing instead on the problem of choosing a particular basis that is well suited to the dataset at hand. Below, we introduce basis functions $\{\varphi_i\}$ that provide a reasonable tradeoff between flexibility and tractability for parametrizing high-dimensional nonlinear functions.

Cylindrical basis functions for the LNP nonlinearity. We propose to parametrize the nonlinearity for many-filter LNP models using cylindrical basis functions (CBFs), which we introduce by analogy to radial basis functions (RBFs). These functions are restricted in some directions of the feature space (like RBFs), but are constant along other dimensions. They are therefore the function-domain analogues to the probability "experts" used in product-of-experts models [42] in that they constrain a high-dimensional function along only a small number of dimensions, while imposing no structure on the others.

We define a "first-order" CBF as a Gaussian bump in one direction of the feature space, parametrized by center location μ and a characteristic width σ:

$$\phi_{1st}(\mathbf{x}) = \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right), \tag{41}$$

which affects the function along vector component x_i and is invariant along $x_{j \neq i}$. Parametrizing f with first-order CBFs is tantamount to assuming f can be parametrized as the sum of 1D functions along each filter axis, that is $f(\mathbf{x}) = g(f_1(x_1) + \ldots f_m(x_m))$, where each function f_i is parametrized with a linear combination of "bump" functions. This setup resembles the parametrization used in the extended projection-pursuit regression (ePPR) model [28], although the nonlinear transformation g confers some added flexibility. For example, we can have multiplicative combination when $g(\cdot) = \exp(\cdot)$, resulting in a separable f, or rectified additive combination when $g(\cdot) = \max(\cdot, 0)$, which is closer to ePPR. If we use d basis functions along each filter axis, the resulting nonlinearity requires kd parameters for a k-filter LNP model.

We can define second-order CBFs as functions that have Gaussian dependence on two dimensions of the input space and that are insensitive to all others:

$$\phi_{2nd}(\mathbf{x}) = \exp\left(\frac{-(x_i - \mu_i)^2 - (x_j - \mu_j)^2}{2\sigma^2}\right) \qquad (42)$$

where μ_i and μ_j determine the center of the basis function in the (x_i, x_j) plane. A second-order basis represents f as a (transformed) sum of these bivariate functions, giving $\binom{k}{2}d^2$ parameters if we use d^2 basis functions for each of the $\binom{k}{2}$ possible pairs of k filter outputs, or merely $\frac{k}{2}d^2$ if we instead partition the k filters into disjoint pairs. Higher-order CBFs can be defined analogously: k'th order CBFs are Gaussian RBFs in a k'-dimensional subspace while remaining constant in the remaining $k-k'$ dimensions. Of course, there is no need to represent the entire nonlinearity using CBFs of the same order. It might make sense, for example, to represent the nonlinear combination of the first two filter responses with second order CBFs (which is comparable to standard MID with a 2D histogram representation of the nonlinearity), and then use first order CBFs to represent the contributions of additional (less-informative) filter outputs.

To illustrate the feasibility of this approach, we applied dimensionality reduction methods to a previously published dataset from macaque V1 [29]. This dataset contains extracellular single unit recordings of simple and complex cells driven by an oriented 1D binary white noise stimulus sequence (i.e., "flickering bars"). For each neuron, we fit an LNP model using: (1) the information-theoretic spike-triggered average and covariance (iSTAC) estimator [19]; and (2) the maximum likelihood estimator for an LNP model with nonlinearity parametrized by first-order CBFs. The iSTAC estimator, which combines information from the STA and STC, returns a list of filters ordered by informativeness about the neural response. It models the nonlinearity as an exponentiated quadratic function (an instance of a generalized quadratic model [23]), and yields asymptotically optimal performance under the condition that stimuli are Gaussian. For comparison, we also implemented a model with a less-constrained nonlinearity, using Gaussian RBFs sensitive to all filter outputs (rbf-LNP). This approach is comparable to to "classic" MID, although it exploits the LNP formulation to allow local smoothing of the nonlinearity (rather than square histogram bins). Even so, the number of parameters in the nonlinearity still grows exponentially with the number of filters so, computational concerns prevented us from recovering more than four filters with this method.

Fig. 8 compares the performance of these estimators on neural data, and illustrates the ability to tractably recover high-dimensional feature spaces using maximum likelihood methods, provided that the nonlinearity is parametrized appropriately. We used 3 CBFs per filter output for the cbf-LNP model (resulting in $3p$ parameters for the nonlinearity), and a grid with 3 RBFs per dimension for the rbf-LNP model (3^p parameters). By contrast, the exponentiated-quadratic nonlinearity underlying the iSTAC estimator requires $O(p^2)$ parameters.
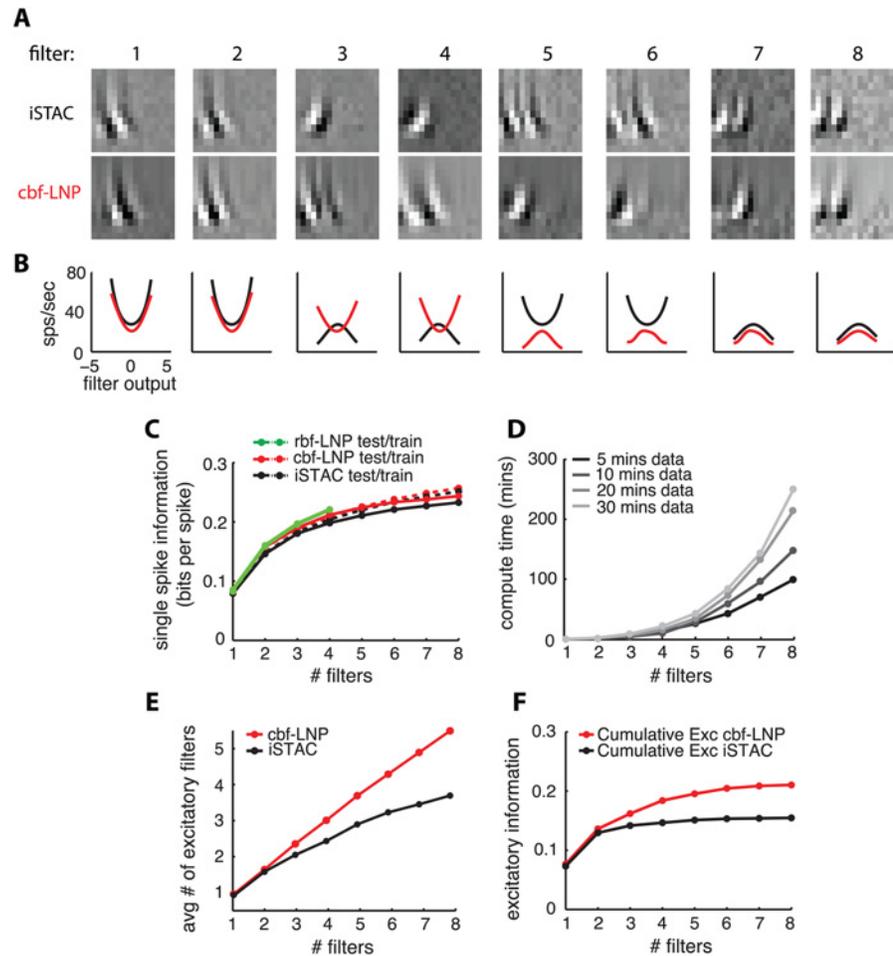
Fig 8. Estimation of high-dimensional subspaces using a nonlinearity parametrized with cylindrical basis functions (CBFs). (A) Eight most informative filters for an example complex cell, estimated with iSTAC (*top row*) and cbf-LNP (*bottom row*). For the cbf-LNP model, the nonlinearity was parametrized with three first-order CBFs for the output of each filter (see Methods). **(B)** Estimated 1D nonlinearity along each filter axis, for the filters shown in (A). Note that third and fourth iSTAC filters are suppressive while third and fourth cbf-LNP filter are excitatory. **(C)** Cross-validated single-spike information for iSTAC, cbf-LNP, and rbf-LNP, as a function of the number of filters, averaged over a population of 16 neurons (selected from [29] for having $\geq$ 8 informative filters). The cbf-LNP estimate outperformed iSTAC in all cases, while rbf-LNP yielded a slight further increase for the first four dimensions. **(D)** Computation time for the numerical optimization of the cbf-LNP likelihood for up to 8 filters. Even for 30 minutes of data and 8 filters, optimisation took about 4 hours. **(E)** Average number of excitatory filters as a function of total number of filters, for each method. **(F)** Information gain from excitatory filters, for each method, averaged across neurons. Each point represents the average amount of information gained from adding an excitatory filter, as a function of the number of filters.

doi:10.1371/journal.pcbi.1004141.g008

To compare performance, we analyzed the growth in empirical single-spike information (computed on a "test" dataset) as a function of the number of filters. Note that this is equivalent to computing test log-likelihood under the LNP model. For a subset of neurons determined to have 8 or more informative filters (16/59 cells), the cbf-LNP filters captured more information than the iSTAC filters (Fig. 8C). This indicates that the CBF nonlinearity captures the nonlinear mapping from filter outputs to spike rate more accurately than an exponentiated quadratic, and that this flexibility confers advantages in identifying the most informative stimulus dimensions. The first four filters estimated under the rbf-LNP model captured slightly more

information again than the cbf-LNP filters, indicating that first-order CBFs provide slightly too restrictive a parametrization for these neurons. Due to computational considerations, we did not attempt to fit the rbf-LNP model with > 4 filters, but note that the cbf-LNP model scaled easily to 8 filters (Fig. 8D).

In addition to its quantitative performance, the cbf-LNP estimate exhibited a qualitative difference from iSTAC with regard to the ordering of filters by informativeness. In particular, the cbf-LNP fit reveals that excitatory filters provide more information than iSTAC attributes to them, and that excitatory filters should come earlier relative to suppressive filters when ordering by informativeness. Fig. 8A-B, which shows the first 8 filters and associated marginal one-dimensional nonlinearities for an example V1 complex cell, provides an illustration of this discrepancy. Under the iSTAC estimate (Fig. 8A, top row), the first two most informative filters are excitatory but the third and fourth are suppressive (see nonlinearities in Fig. 8B). However, the cbf-LNP estimate (and rbf-LNP estimate, not shown) indicates that the four most informative filters are all excitatory. This tendency holds across the population of neurons. We can quantify it in terms of the number of excitatory filters within the first n filters identified (Fig. 8E) or the total amount of information (i.e., log-likelihood) contributed by excitatory filters (Fig. 8F). This shows that iSTAC, which nevertheless provides a computationally inexpensive initialization for the cbf-LNP estimate, does not accurately quantify the information contributed by excitatory filters. Most likely, this reflects the fact that an exponentiated-quadratic does not provide as accurate a description of the nonlinearity along excitatory stimulus dimensions as can be obtained with a non-parametric estimator.

Relationship to previous work

A variety of neural dimensionality reduction methods have been proposed previously. Here, we consider the relationship of the methods described in this study to these earlier approaches. Rapela *et al* [28] introduced a technique known as extended Projection Pursuit Regression (ePPR), where the high-dimensional estimation problem is reduced to a sequence of simpler low-dimensional ones. The approach is iterative. A one-dimensional model is found first, and the dimensionality is then progressively increased to optimize a cost function, but with the search for filters restricted to dimensions orthogonal to all the filters already identified. From a theoretical perspective this assumes that the spiking probability can be defined as a sum of functions of the different stimulus components; that is,

$$p(spike|\mathbf{s}) = g_1(\mathbf{k}_1^\top \mathbf{s}) + g_2(\mathbf{k}_2^\top \mathbf{s}) + \cdots g_N(\mathbf{k}_N^\top \mathbf{s}). \tag{43}$$

Rowekamp *et al* [43] compared such an approach to the joint optimization more common in MID analysis (as in [18]), and derived the bias that results from sequential optimization and its implicit additivity. By contrast, we have focused here on parametrization rather than sequential optimization. In all cases, we optimized the log-likelihood simultaneously over all filter dimensions. For high-dimensional models, we advocate parametrization of the nonlinearity so as to avoid the curse of dimensionality. However, the CBF form we have introduced is more flexible than that of ePPR, both in that two- or more dimensional components are easily included, and in that the outputs of the components can be combined non-linearly.

Other proposals can be seen as assuming specific quadratic-based parametrizations for the nonlinearity, that are more restrictive than the CBF form. The iSTAC estimator, introduced by Pillow & Simoncelli [19], is based on maximization of the KL divergence between Gaussian approximations to the spike-triggered and stimulus ensembles—thus finding the feature space that maximizes the single-spike information under a Gaussian model of both the spike-triggered and stimulus ensembles. Park & Pillow [44] showed its relationship to an LNP model

with an exponentiated quadratic spike rate, which takes the form:

$$p(spike|\mathbf{s}) = \exp\left(a + K^\top \mathbf{s} + \mathbf{s}^\top C \mathbf{s}\right). \tag{44}$$

Such a nonlinearity readily yields maximum likelihood estimators based on the STA and STC. Moreover, they proposed a new model, known as "elliptical LNP", which allowed estimation of a non-parametric nonlinearity around the quadratic function (instead of assuming an exponential form). Rajan et al. [24] considered a similar model within an information-theoretic framework and proposed extending it to nonlinear combinations of outputs from multiple quadratic functions. In a similar vein, Sharpee *et al*[45, 46] used

$$p(spike|\mathbf{s}) = \frac{1}{1 + \exp\left(a + K\mathbf{s} + \mathbf{s}^\top C\mathbf{s}\right)}. \tag{45}$$

This model corresponds to quadratic logistic regression, and thus assumes Bernoulli output noise (and a binary response). The authors also proposed a "nonlinear MID" in which the standard MID estimator is extended by setting the firing rate to be a quadratic function of the form $f(\mathbf{k}^\top \mathbf{s} + \mathbf{s}^\top C \mathbf{s})$. This method is one-dimensional in a quadratic stimulus space (unlike multidimensional linear MID) and therefore avoids the curse of dimensionality. Other work has used independent component analysis to find directions in stimulus space in which the spike-triggered distribution has maximal deviations from Gaussianity [8].

Discussion

Distributional assumptions implicit in MID

We have studied the estimator known as maximally informative dimensions (MID) [18], a popular approach for estimating informative dimensions of stimulus space from spike-train data. Although the MID estimator was originally described in information-theoretic language, we have shown that, when used with plug-in estimators for information-theoretic quantities, it is mathematically identical to the maximum likelihood estimator for a linear-nonlinear-Poisson (LNP) encoding model. This equivalence holds irrespective of spike rate, the amount of data, or the size of time bins used to count spikes. We have shown that this follows from the fact that the plug-in estimate for single-spike information is equal (up to an additive constant) to the log-likelihood per spike of the data under an LNP model.

Estimators defined by the optima of information-theoretic functionals have attractive theoretical properties, including that they provide well-defined and (theoretically) distribution-agnostic characterizations of data. In practice, however, such agnosticism can be difficult to achieve, as the need to estimate information-theoretic quantities from data requires the choice of a particular estimator. MID has the virtue of using a non-parametric estimator for raw and spike-triggered stimulus densities, meaning that the number of parameters (i.e., the number of histogram bins) can grow flexibly with the amount of data. This allows it to converge for arbitrary densities, in the limit of infinite data. However, for a finite dataset, the choice of number of bins is critical for obtaining an accurate estimate. As we show in Fig. 3, a poor choice can lead to a systematic under- or over-estimate of the single-spike information, and in turn, a poor estimate of the most informative stimulus dimensions. Determining the number of histogram bins should therefore be considered a model selection problem, validated with a statistical procedure such as cross-validation.

A second kind of distributional assumption arises from MID's reliance on single-spike information, which is tantamount to an assumption of Poisson spiking. To be clear, the single-spike information represents a valid information-theoretic quantity that does not explicitly assume any model. As noted in [26], it is simply the information carried by a single spike time,

independent of all other spike times. However, conditionally independent spiking is also the fundamental assumption underlying the Poisson model and, as we have shown, the standard MID estimator (based on the KL-divergence between histograms) is mathematically identical to the maximum likelihood estimator for an LNP model with piece-wise constant nonlinearity. Thus, MID achieves no more and no less than a maximum likelihood estimator for a Poisson response model. As we illustrate in Fig. 4, MID does not maximize the mutual information between the projected stimulus and the spike response when the distribution of spikes conditioned on stimuli is not Poisson; it is an inconsistent estimator for the relevant stimulus subspace in such cases.

The distributional-dependence of MID should therefore be considered when interpreting its estimates of filters and nonlinearities. MID makes different, but not necessarily fewer, assumptions when compared to other LN estimators. For instance, although the maximum-likelihood estimator for a generalized linear model assumes a less-flexible model for the neural nonlinearity than does MID, it readily permits estimation of certain forms of spike-interdependence that MID neglects. In particular, MID-derived estimates are subject to concerns regarding model mismatch that arise whenever the true generative family is unknown [47].

In light of the danger that these distributional assumptions be obscured by the information-theoretic framing of MID, our belief is that the safer approach is to specify a model explicitly and adopt a likelihood-based estimation framework. Where the information theoretic and likelihood-based estimators are identical, nothing is lost by this approach. However, besides making assumptions explicit, the likelihood-based framework also readily facilitates the introduction of suitable priors for regularization, or hierarchical models [48, 49], or of more structured models the type discussed here.

Generalizations

Having clarified the relationship between MID and the LNP model, we introduced two generalizations designed to recover a maximally informative stimulus projection when neural response variability is not well described as Poisson. From a model-based perspective, the generalizations correspond to maximum likelihood estimators for a linear-nonlinear-Bernoulli (LNB) model (for binary spike counts), and the linear-nonlinear-Count (LNC) model (for arbitrary discrete spike counts). For both models, we obtained an equivalent relationship between log-likelihood and an estimate of mutual information between stimulus and response. This correspondence extends previous work that showed only approximate or asymptotic relationships between between information-theoretic and maximum-likelihood estimators [20, 24, 25]. The LNC model is the most general of the models we have considered. It requires the fewest assumptions, since it allows for arbitrary distributions over spike count given the stimulus. It includes both LNB and LNP as special cases (i.e., when the count distribution is Bernoulli or Poisson, respectively).

We could analogously define arbitrary "LNX" models, where X stands in for any probability distribution over the neural response (analog or discrete), and perform dimensionality reduction by maximizing likelihood for the filter parameters under this model. The log-likelihood under any such model can be associated with an information-theoretic quantity, analogous to single-spike, Bernoulli, and count information, using the difference of log-likelihoods (see also [35]):

$$I_{lnx} \triangleq \sum_{r,\mathbf{s}} p(\mathbf{s}) p_x(r|\mathbf{s}, \theta) \log p_x(r|\mathbf{s}, \theta) \; - \; \sum_r p_x(r|\theta_0) \log p_x(r|\theta_0), \qquad (46)$$

where $p_x(r|\mathbf{s},\theta)$ denotes the conditional response distribution associated with the LNX model with parameters θ, and $p_x(r|\theta_0)$ describes the marginal distribution over r under the stimulus

distribution $p(\mathbf{s})$. The empirical or plug-in estimate of this information is equal to the LNX model log-likelihood plus the estimated marginal entropy:

$$\hat{I}_{lnx}(\theta) = \frac{1}{n}(\mathcal{L}_{lnx}(\theta; D) - \mathcal{L}_{lnx}(\theta_0; D)), \tag{47}$$

where n denotes the number of samples and θ_0 depends only on the marginal response distribution. The maximum likelihood estimate is therefore equally a maximal-information estimate.

Note that all of the dimensionality-reduction methods we have discussed treat neural responses as conditionally independent given the stimulus, meaning that they do not capture dependencies between spike counts in different time bins (e.g., due to refractoriness, bursting, adaptation, etc.). Spike-history dependencies can influence the single-bin spike count distribution—for example, a Bernoulli model is more accurate than a Poisson model when the bin size is smaller than or equal to the refractory period, since the Poisson model assigns positive probability to the event of having ≥ 2 two spikes in a single bin. The models we have considered can all be extended to capture spike history dependencies by augmenting the stimulus with a vector representation of spike history, as in both conditional renewal models and generalized linear models [10, 12, 27, 50–52].

Lastly, we have shown that viewing MID from a model-based perspective provides insight into how to overcome practical limitations on the number of filters that can be estimated. Standard implementations of MID employ histogram-based density estimators for $p(K^{\top}\mathbf{s})$ and $p(K^{\top}\mathbf{s}|spike)$. However, dimensionality and parameter count can be a crippling issue given limited data, and density estimation becomes intractable in dimensionalities > 3. Furthermore, the dependence of the information on the logarithm of the ratio of these densities amplifies sensitivity to errors in these estimates. The LNP-likelihood view suggests direct estimation of the nonlinearity f, rather than of the densities. Such estimates are naturally more robust, and are more sensibly regularized based on expectations about neuronal responses without reference to any regularities in the stimulus distribution. We have proposed a flexible yet tractable form for the nonlinearity in terms of linear combinations of basis functions cascaded with a second *output* nonlinearity. This approach yielded a flexible, computationally efficient, constrained version of MID that is able to estimate high-dimensional feature spaces. It is also general in the sense that it encompasses standard MID, generalized linear and quadratic models, and other constrained models that scale tractably to high-dimensional subspaces. Future work might seek to extend this flexible likelihood-based approach further, for example by including priors over the weights with which basis functions are combined to improve regularization, or perhaps by adjusting hyperparameters in a hierarchical model as has been successful with linear approaches [48, 49].

In recent years, the ability to successfully characterize low-dimensional neural feature spaces using MID has proved useful to address questions relating to multidimensional feature selectivity [53–56]. In all of these examples however, issues with dimensionality have prevented the estimation of feature spaces with more than two dimensions. The methods presented within this paper will help to overcome these issues, opening access to further important questions regarding the relationship between stimuli and their neural representation.

Methods

Bound on lost information under MID

Here we present a derivation of the lower bound on the fraction of total information carried by silences for a Bernoulli neuron, in the limit of rare spiking. For notational convenience, let $\rho = p(r = 1)$ denote the marginal probability of a spike, so the probability of silence is $p(r = 0) =$

$1 - \rho$. Let $Q_1 = p(\mathbf{s}|r = 1)$ and $Q_0 = p(\mathbf{s}|r = 0)$ denote the spike-triggered and silence-triggered stimulus distributions, respectively. Let $P_\mathbf{s} = p(\mathbf{s})$ denote the raw stimulus distribution. Note that we have the $P_\mathbf{s} = \rho Q_1 + (1 - \rho)Q_0$. The mutual information between the stimulus and one bin of the response (Equation 18) can then be written

$$I(\mathbf{s}, r) = \rho D_{KL}(Q_1 \,\|\, P_\mathbf{s}) + (1 - \rho)D_{KL}(Q_0 \,\|\, P_\mathbf{s}). \tag{48}$$

Note that this is a generalized form of the Jensen-Shannon (JS) divergence; the standard JS-divergence between Q_0 and Q_1 is obtained when $\rho = \frac{1}{2}$.

In the limit of small ρ (i.e., the Poisson limit), the mutual information is dominated by the first (Q_1) term. Here we wish to show a bound on the fraction of information carried by the Q_0 term. We can do this by computing a second-order Taylor expansion of $(1 - \rho)D_{KL}(Q_o|P_\mathbf{s})$ and $I(\mathbf{s},r)$ around $\rho = 0$, and show that their ratio is bounded below by $\rho/2$. Expanding in ρ, we have

$$(1 - \rho)D_{KL}(Q_o \,\|\, P_\mathbf{s}) = \tfrac{1}{2}\rho^2 V(Q_1, Q_0) + O(\rho^3), \text{ and} \tag{49}$$

$$I(\mathbf{s}, r) = \rho D_{KL}(Q_1 \,\|\, Q_0) - \tfrac{1}{2}\rho^2 V(Q_1, Q_0) + O(\rho^3), \tag{50}$$

where

$$V(Q_1, Q_0) = \int_\Omega Q_1\left(\frac{Q_1}{Q_0} - 1\right)d\mathbf{s}, \tag{51}$$

which is a an upper bound on the KL-divergence: $V(Q_1, Q_0) \geq D_{KL}(Q_1 \,\|\, Q_0)$, since $(z-1) \geq \log(z)$. We therefore have

$$\frac{(1 - \rho)D_{KL}(Q_o \,\|\, P_\mathbf{s})}{I(\mathbf{s}, r)} = \frac{\tfrac{1}{2}\rho^2 V(Q_1, Q_0) + O(\rho^3)}{\rho D_{KL}(Q_1 \,\|\, Q_0) - \tfrac{1}{2}\rho^2 V(Q_1, Q_0) + O(\rho^3)} \geq \frac{\rho V(Q_1, Q_0)}{2D_{KL}(Q_1 \,\|\, Q_0)} \geq \frac{\rho}{2} \tag{52}$$

in the limit $\rho \to 0$.

We conjecture that the bound holds for all values of ρ. For the case of $\rho = \frac{1}{2}$, this corresponds to an assertion about the relative contribution of each of the two terms in the JS divergence, that is:

$$\frac{D_{KL}\left(Q_1 \,\|\, \tfrac{1}{2}(Q_0 + Q_1)\right)}{D_{KL}\left(Q_1 \,\|\, \tfrac{1}{2}(Q_0 + Q_1)\right) + D_{KL}\left(Q_1 \,\|\, \tfrac{1}{2}(Q_0 + Q_1)\right)} \geq \frac{1}{4} \tag{53}$$

for any choice of distributions Q_0 and Q_1. We have been unable to find any counter-examples to this (or to the more general conjecture), but have so far been unable to find a general proof.

Single-spike information and Poisson log-likelihood

An important general corollary to the equivalence between MID and an LNP maximum likelihood estimate is that the standard single-spike information estimate $\hat{I}_{ss}$ based on a PSTH measured in response to repeated stimuli is also a Poisson log-likelihood per spike (plus a constant). Specifically, the empirical single-spike information is equal to the log-likelihood ratio between an inhomogeneous and homogeneous Poisson model of the repeat data (normalized by spike count):

$$\hat{I}_{ss} = \frac{1}{n_{sp}}\left(\mathcal{L}(\hat{\boldsymbol{\lambda}}_{ML}; \mathbf{r}) - \mathcal{L}(\bar{\boldsymbol{\lambda}}; \mathbf{r})\right), \tag{54}$$

where $\hat{\lambda}_{ML}$ denotes the maximum-likelihood or plug-in estimate of the time-varying spike rate (i.e., the PSTH itself), $\bar{\lambda}$ is the mean spike rate across time, and $\mathcal{L}(\lambda; \mathbf{r})$ denotes the log-likelihood of the repeat data $\mathbf{r}$ under a Poisson model with time-varying rate λ.

We can derive this equivalence as follows. Let $\{r_{jt}\}$ denote spike counts collected during a "frozen noise" experiment, with repeat index $j \in \{1, \ldots, n_{rpt}\}$ and index $t \in \{1, \ldots, n_t\}$ over time bins of width Δ. Then $T = n_t \Delta$ is the duration of the stimulus and $N = n_t n_{rpt}$ is the total number of time bins in the entire experiment. The single-spike information can be estimated with a discrete version of the formula for single-spike information provided in [26] (see eq. 2.5):

$$\hat{I}_{ss} = \frac{1}{n_t} \sum_{t=1}^{n_t} \frac{\hat{\lambda}(t)}{\bar{\lambda}} \log \frac{\hat{\lambda}(t)}{\bar{\lambda}}, \tag{55}$$

where $\hat{\lambda}(t) = \frac{1}{\Delta n_{rpt}} \sum_{j=1}^{n_{rpt}} r_{jt}$ is an estimate of the spike rate in the t'th time bin in response to the stimulus sequence s, and $\bar{\lambda} = (\sum_{t=1}^{n_t} \hat{\lambda}(t))/n_t$ is the mean spike rate across the experiment. Note that this formulation assumes (as in [26]) that T is long enough that an average over stimulus sequences is well approximated by the average across time.

The plug-in (ML) estimator for spike rate can be read off from the peri-stimulus time histogram (PSTH). It results from averaging the response across repeats for each time bin:

$$\hat{\lambda}(t) = \frac{1}{n_{rpt}\Delta} \sum_{j=1}^{n_{rpt}} r_{jt}. \tag{56}$$

Clearly, $\bar{\lambda} = \frac{n_{sp}}{N\Delta}$, where $n_{sp} = \Sigma_{j,t} \, r_{jt}$ is the total spike count. This allows us to rewrite single-spike information (Equation 55) as:

$$\hat{I}_{ss} = \frac{n_{rpt}\Delta}{n_{sp}} \sum_{t=1}^{n_t} \hat{\lambda}(t) \log \hat{\lambda}(t) - \log \frac{n_{sp}}{N\Delta}. \tag{57}$$

Now, consider the Poisson log-likelihood $\mathcal{L}$ evaluated at the ML estimate $\hat{\lambda} = (\hat{\lambda}(1), \ldots, \hat{\lambda}(n_t))$, i.e., the conditional probability of the response data $\mathbf{r} = \{r_{jt}\}$ given rate vector $\hat{\lambda}$. This is given by:

$$\begin{aligned}
\mathcal{L}(\hat{\lambda}; \mathbf{r}) &= \sum_{t=1}^{n_t} \sum_{j=1}^{n_{rpt}} \left(r_{jt} \log \left(\hat{\lambda}(t)\Delta \right) - \hat{\lambda}(t)\Delta - \log r_{jt}! \right) \\
&= \sum_{t=1}^{n_t} \left(\sum_{j=1}^{n_{rpt}} r_{jt} \right) \log \hat{\lambda}(t) - n_{sp} + n_{sp} \log \Delta - \sum_{t,j} \log r_{jt}! \\
&= n_{rpt}\Delta \sum_{t=1}^{n_t} \hat{\lambda}(t) \log \hat{\lambda}(t) - n_{sp} + n_{sp} \log \Delta - \sum_{t,j} \log r_{jt}! \\
&= n_{sp}\hat{I}_{ss} + n_{sp} \log \frac{n_{sp}}{N} - n_{sp} - \sum_{t,j} \log r_{jt}! \\
&= n_{sp}\hat{I}_{ss} + \mathcal{L}(\bar{\lambda}; \mathbf{r}), \tag{58}
\end{aligned}$$

which is identical to relationship between single-spike information and Poisson log-likelihood expressed in Equation 13. Thus, even when estimated from raster data, I_{ss} is equal to the difference between Poisson log-likelihoods under an inhomogeneous (rate-varying) and a homogeneous (constant rate) Poisson model, divided by spike count (see also [57]). These normalized log-likelihoods can be conceived as entropy estimates, with $-\frac{1}{n_{sp}}\mathcal{L}(\bar{\lambda}; \mathbf{r})$ providing an estimate

for prior entropy, measuring the prior uncertainty about spike times given the mean rate, and $-\frac{1}{n_{sp}}\mathcal{L}(\hat{\lambda};\mathbf{r})$ corresponding to posterior entropy, measuring the posterior uncertainty once we know the time-varying spike rate.

A similar quantity has been used to report the cross-validation performance of conditionally Poisson models, including the GLM [13, 58]. To penalize over-fitting, the empirical single-spike information is evaluated using the rate estimate $\hat{\lambda}$ obtained with parameters fit to training data and responses $\mathbf{r}$ from unseen test data. This results in the "cross-validated" single-spike information:

$$\hat{I}_{ss}^{[xv]} = \frac{1}{n_{sp}^{[test]}}\left(\mathcal{L}(\hat{\lambda}^{[train]};\mathbf{r}^{[test]}) - \mathcal{L}(\bar{\lambda}^{[test]};\mathbf{r}^{[test]})\right). \tag{59}$$

This can be interpreted as the predictive information (in bits-per-spike) that the model captures about test data, above and beyond that captured by a homogeneous Poisson model with correct mean rate. Note that this quantity can be negative in cases of extremely poor model fit, that is, when the model prediction on test data is worse than of the best constant-spike-rate Poisson model. Cross-validated single-spike information provides a useful measure for comparing models with different numbers of parameters (e.g., a 1-filter vs. 2-filter LNP model), since units of "bits" are more interpretable than raw log-likelihood of test data. Generally, $\hat{I}_{ss}^{[xv]}$ can be considered to a lower bound on the model's true predictive power, due to stochasticity in both training and test data. By contrast, the empirical I_{ss} evaluated on training data tends to over-estimate information due to over-fitting.

Computation of model-based information quantities

To gain intuition for the different information measures we have considered (Poisson, Bernoulli, and categorical or "count"), it is useful to consider how they differ for a simple idealized example. Consider a world with two stimuli, 'A' and 'B', and two possible discrete stimulus sequences, $s_1 = AB$ and $s_2 = BA$, each of which occurs with equal probability, so $p(s_1) = p(s_2) = 0.5$. Assume each sequence lasts $T = 2$s, so the natural time bin size for considering the spike response is $\Delta = 1$s. Suppose that stimulus A always elicits 3 spikes, while B always elicits 1 spike. Thus, when sequence s_1 is presented, we observe 3 spikes in the first time interval and 1 spike in the second interval; when s_2 is presented, we observe 1 spike in the first time interval and 3 spikes in the second.

Single-spike information can be computed exactly from $\lambda_1(t)$ and $\lambda_2(t)$, the spike rate in response to stimulus sequence s_1 and s_2, respectively. For this example, $\lambda_1(t)$, takes the value 3 during (0,1] and 1 during (1,2], while $\lambda_2(t)$ takes values 1 and 3 during the corresponding intervals. The mean spike rate for both stimuli is $\bar{\lambda} = 2$ sp/s. Plugging these into Equation 54 gives single-spike information of $I_{ss} = 0.19$ bits/spike. This result is slightly easier to grasp using an equivalent definition of single-spike information as the mutual information between the stimulus s and a single spike time τ (see [26]). If one were told that a spike, sampled at random from the four spikes present during every trial, occurred during [0, 1], then the posterior $p(s|\tau = 1)$ attaches 3/4 probability to $s = s_1$ and 1/4 to $s = s_2$. The posterior entropy is therefore $-0.25 \log 0.25 - 0.75 \log 0.75 = 0.81$ bits. We obtain the same entropy if the spike occurs in the second interval, so $H(s|\tau) = 0.81$. The prior entropy is $H(s) = 1$ bit, so once again we have $I_{ss} = 1 - 0.81 = 0.19$ bits/spike.

The Bernoulli information, by contrast, is undefined, since r takes values outside the set {0,1}, and therefore cannot have a Bernoulli distribution. To make Bernoulli information well

defined, we would need to either truncate spike counts above 1 (*e.g.*, [59]), or else use smaller bin size so that no bin contains more than one spike. In the latter case, we would need to provide more information about the distribution of spike times within these finer bins. If, for example, the three spikes elicited by *A* are evenly spaced within the interval and we use bins equal to 1/3s, then the Bernoulli information will clearly exceed single-spike information, since the time of a no-spike response ($r = 0$, a term neglected by single-spike information) provides perfect information about the stimulus, since it occurs only in response to *B*.

Lastly, the count information is easy to compute from the fact that count *r* carries perfect information about the stimulus, so the mutual information between stimulus (*A* or *B*) and *r* is 1 bit. We defined I_{count} to be the mutual information normalized by the mean spike count per bin (Equation 35). Thus, $I_{count} = 0.5$ bits/spike, which is more than double the single-spike information.

Gradient and Hessian of LNP log-likelihood

Here we provide formulas useful for fitting the the many-filter LNP model with cylindrical basis function (CBF) nonlinearity. We performed joint optimization of filter parameters *K* and basis function weights $\{\alpha_i\}$ using MATLAB's `fminunc` function. We found this approach to converge much more rapidly than alternating coordinate ascent. We used analytically computed gradient and Hessian of the joint-likelihood to speed up performance, which we provide here.

Given a dataset $\{(\mathbf{s}_t, r_t)\}_{t=1}^{n_t}$, define $\mathbf{r} = (r_1, \ldots, r_{n_t})^\top$ and $\lambda = (f(K^\top \mathbf{s}_1), \ldots, f(K^\top \mathbf{s}_{n_t}))^\top$, where nonlinearity $f = g(\sum_i \alpha_i \varphi_i)$ depends on basis function $\Phi = \{\varphi_i\}$ and weights $\boldsymbol{\alpha} = \{\alpha_i\}$ (Equation 39). We can write the log-likelihood for the many-filter LNP model (from Equations 38–40) as:

$$\mathcal{L}(\theta) = \mathbf{r}^\top \log \lambda - (\Delta)\mathbf{1}^\top \lambda \tag{60}$$

where $\theta = \{K, \boldsymbol{\alpha}\}$ are the model parameters, Δ is the time bin size, and $\mathbf{1}$ denotes a vector of ones. The first and second derivatives of the log-likelihood are given by

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \left(\frac{\partial \lambda}{\partial \theta_i}\right)^\top \left(\frac{\mathbf{r}}{\lambda} - \Delta 1\right) \tag{61}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} = \left(\frac{\partial^2 \lambda}{\partial \theta_i \partial \theta_j}\right)^\top \left(\frac{\mathbf{r}}{\lambda} - \Delta 1\right) + \left(\frac{\partial \lambda}{\partial \theta_i}\frac{\partial \lambda}{\partial \theta_j}\right)^\top \left(\frac{\mathbf{r}}{\lambda^2}\right), \tag{62}$$

where multiplication, division, and exponentiation operations on vector quantities indicate component-wise operations.

Let $\mathbf{k}_1, \ldots, \mathbf{k}_m$ denote the linear filters, i.e., the *m* columns of *K*. Then the required gradients of λ with respect to the model parameters can be written:

$$\frac{\partial \lambda}{\partial \mathbf{k}_i} = S^\top (\lambda' \circ \Phi^{(i)}\alpha) \tag{63}$$

$$\frac{\partial \lambda}{\partial \alpha} = \Phi^\top \lambda' \tag{64}$$

where *S* denotes the ($n_t \times D$) stimulus design matrix, Φ denotes the ($n_t \times n_\varphi$) matrix whose (t, j)'th entry is $\varphi_j(K^\top \mathbf{s}_t)$, and $\Phi^{(i)}$ denotes a matrix of the same size, formed by the point-wise derivative of Φ with respect to its *i*'th input component, evaluated at each projected stimulus $K^\top \mathbf{s}_t$. Finally, $\lambda' = g'(\Phi\boldsymbol{\alpha})$ is a ($n_t \times 1$) vector composed of the point-wise derivatives of the

inverse-link function g at its input, and '$\circ$' denotes Hadamard or component-wise vector product.

Lastly, second derivative blocks, which can be plugged into Equation 62 to form the Hessian, are given by

$$\frac{\partial^2 \boldsymbol{\lambda}}{\partial \mathbf{k}_i \partial \mathbf{k}_j} = S^\top \operatorname{diag}\left(\left[\boldsymbol{\lambda}'' \circ (\Phi^{(i)}\boldsymbol{\alpha}) \circ (\Phi^{(j)}\boldsymbol{\alpha})\right] + \left[\boldsymbol{\lambda}' \circ \Phi^{(i,j)}\boldsymbol{\alpha}\right]\right)S \tag{65}$$

$$\frac{\partial^2 \boldsymbol{\lambda}}{\partial \boldsymbol{\alpha}^2} = \Phi^\top \operatorname{diag}(\boldsymbol{\lambda}'')\Phi \tag{66}$$

$$\frac{\partial^2 \boldsymbol{\lambda}}{\partial \mathbf{k}_i \partial \boldsymbol{\alpha}} = S^\top \left(\operatorname{diag}(\boldsymbol{\lambda}'' \circ (\Phi^{(i)}\boldsymbol{\alpha}))\Phi + \operatorname{diag}(\boldsymbol{\lambda}')\,\Phi^{(i)}\right), \tag{67}$$

where $\boldsymbol{\lambda}'' = g''(\Phi\boldsymbol{\alpha})$ and $\Phi^{(i,j)}$ is a matrix of point-wise second-derivatives of Φ with respect to i'th and j'th inputs, evaluated for each projected stimulus $K^\top \mathbf{s}_t$.

V1 data analysis

To examine performance in recovering high-dimensional subspaces, we analyzed data from macaque V1 cells, driven by 1D binary white noise "flickering bars" stimulus, presented at a frame rate of 100 Hz (data published in [29]). The spatiotemporal stimulus had between 8 and 32 spatial bars and we considered 10 time bins for the temporal integration window. This made for a stimulus space with dimensionality ranging from 80 to 320.

The cbf-LNP model was implemented with a cylindrical basis function (CBF) nonlinearity using three first-order CBFs per filter. For a k-filter model, this resulted in $3k$ parameters for the nonlinearity, and $(240+3)k$ parameters in total for a stimulus with 24 bars.

The traditional MID estimator (rbf-LNP) was implemented using radial basis functions (RBFs) to represent the nonlinearity. Unlike the histogram-based parametrization discussed in the manuscript (which produces a piece-wise constant nonlinearity), this results in a smooth nonlinearity and, more importantly, a smooth log-likelihood with tractable analytic gradients. We defined a grid of RBFs with three grid points per dimension, so that CBF and RBF models were identical for a 1-filter model. For a k-filter model, this resulted in 3^k parameters for the nonlinearity, and $240k+3^k$ parameters in total, for a stimulus with 24 bars.

For both models, the basis function responses were combined linearly and transformed by a "soft-rectification" function: $g(\cdot) = \log(1+\exp(\cdot))$, to ensure positive spike rates. We also evaluated the performance of an exponential function, $g(\cdot) = \exp(\cdot)$, which yielded slightly worse performance (reducing single-spike information by ~ 0.02 bits/spike).

The cbf- and rbf-LNP models were both fit by maximizing the likelihood for the model parameters $\theta = \{K, \boldsymbol{\alpha}\}$. Both models were fit incrementally, with the $N+1$ dimensional model being initialized with the parameters of the N dimensional model, plus one additional filter (initialized with the iSTAC filter that provided the greatest increase in log-likelihood). The joint likelihood in K and $\boldsymbol{\alpha}$ was ascended using MATLAB's `fminunc` optimization function, which exploits analytic gradients and Hessians. The models were fit to 80% of the data, with the remaining 20% used for validation.

In order to calculate information contributed by excitatory filters under the cbf-LNP model (Fig. 8F), we removed each filter from the model and refit the nonlinearity (using the training data) using just the other filters. We quantified the information contributed by each filter as the difference between log-likelihood of the full model and log-likelihood of the reduced model

(on test data). We sorted the filters by informativeness and computed the cumulative sum of information loss to obtain the trace shown in (Fig. 8F).

Measurements of computation time (Fig. 8D) were averaged over 100 repetitions using different random seeds. For each cell, four segments of activity were chosen randomly with fixed lengths of 5, 10, 20 and 30 minutes, which contained between about 22000 and 173000 spikes. Even with 30 minutes of data, 8 filters could be identified within about 4 hours on a desktop computer, making the approach tractable even for large numbers of filters.

Code will be provided at http://pillowlab.princeton.edu/code.html.

Acknowledgments

Author Contributions

Conceived and designed the experiments: RSW MS JWP. Performed the experiments: RSW MS JWP. Analyzed the data: RSW JWP. Contributed reagents/materials/analysis tools: RSW JWP. Wrote the paper: RSW MS JWP.

References

1. de Ruyter van Steveninck RR, Bialek W (1988) Real-time performance of a movement-senstivive neuron in the blowfly visual system: coding and information transmission in short spike sequences. Proceedings of the Royal Society B 234: 379–414. doi: 10.1098/rspb.1988.0055

2. Aguera y Arcas B, Fairhall AL (2003) What causes a neuron to spike? Neural Computation 15: 1789–1807. doi: 10.1162/08997660360675044 PMID: 14511513

3. Aguera y Arcas B, Fairhall AL, Bialek W (2003) Computation in a single neuron: Hodgkin and Huxley revisited. Neural Computation 15: 1715–1749. doi: 10.1162/08997660360675017 PMID: 14511510

4. Simoncelli EP, Pillow JW, Paninski L, Schwartz O (2004) Characterization of neural responses with stochastic stimuli. In: Gazzaniga M, editor, The Cognitive Neurosciences, III, Cambridge, MA: MIT Press, chapter 23 pp. 327–338.

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

6. Chichilnisky EJ (2001) A simple white noise analysis of neuronal light responses. Network: Computation in Neural Systems 12: 199–213. doi: 10.1080/net.12.2.199.213

7. Schwartz O, Pillow JW, Rust NC, Simoncelli EP (2006) Spike-triggered neural characterization. Journal of Vision 6: 484–507. doi: 10.1167/6.4.13 PMID: 16889482

8. Saleem AB, Krapp HG, Schultz SR (2008) Receptive field characterization by spike-triggered independent component analysis. Journal of Vision 8. doi: 10.1167/8.13.2 PMID: 19146332

9. Brillinger DR (1988) Maximum likelihood analysis of spike trains of interacting nerve cells. Biological Cybernetics 59: 189–200. doi: 10.1007/BF00318010 PMID: 3179344

10. Kass RE, Ventura V (2001) A spike-train probability model. Neural Computation 13: 1713–1720. doi: 10.1162/08997660152469314 PMID: 11506667

11. Paninski L (2004) Maximum likelihood estimation of cascade point-process neural encoding models. Network: Computation in Neural Systems 15: 243–262. doi: 10.1088/0954-898X/15/4/002

12. Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN (2005) A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. Journal of Neurophysiology 93: 1074–1089. doi: 10.1152/jn.00697.2004 PMID: 15356183

13. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, et al. (2008) Spatio-temporal correlations and visual signaling in a complete neuronal population. Nature 454: 995–999. doi: 10.1038/nature07140 PMID: 18650810

14. Park IM, Pillow JW (2011) Bayesian spike-triggered covariance analysis. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K, editors, Advances in Neural Information Processing Systems 24. Curran Associates, Inc. pp. 1692–1700.

15. McFarland JM, Cui Y, Butts DA (2013) Inferring nonlinear neuronal computation based on physiologically plausible inputs. PLoS Computational Biology 9: e1003143. doi: 10.1371/journal.pcbi.1003143 PMID: 23874185

16. Cui Y, Liu LD, Khawaja FA, Pack CC, Butts DA (2013) Diverse suppressive influences in area MT and selectivity to complex motion features. Journal of Neuroscience 33: 16715–16728. doi: 10.1523/JNEUROSCI.0203-13.2013 PMID: 24133273

17. Paninski L (2003) Convergence properties of some spike-triggered analysis techniques. Network: Computation in Neural Systems 14: 437–464. doi: 10.1088/0954-898X/14/3/304

18. Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Computation 16: 223–250. doi: 10.1162/089976604322742010 PMID: 15006095

19. Pillow JW, Simoncelli EP (2006) Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis. Journal of Vision 6: 414–428. doi: 10.1167/6.4.9 PMID: 16889478

20. Kouh M, Sharpee TO (2009) Estimating linear-nonlinear models using renyi divergences. Network: Computation in Neural Systems 20: 49–68. doi: 10.1080/09548980902950891 PMID: 19568981

21. Fitzgerald JD, Rowekamp RJ, Sincich LC, Sharpee TO (2011) Second order dimensionality reduction using minimum and maximum mutual information models. PLoS Computational Biology 7: e1002249. doi: 10.1371/journal.pcbi.1002249 PMID: 22046122

22. Rajan K, Bialek W (2013) Maximally informative stimulus energies in the analysis of neural responses to natural signals. PLoS ONE 8: e71959. doi: 10.1371/journal.pone.0071959 PMID: 24250780

23. Park IM, Archer EW, Priebe N, Pillow JW (2013) Spectral methods for neural characterization using generalized quadratic models. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K, editors, Advances in Neural Information Processing Systems 26, Curran Associates, Inc. pp. 2454–2462.

24. Rajan K, Marre O, Tkačik G (2013) Learning quadratic receptive fields from neural responses to natural stimuli. Neural Computation 25: 1661–1692. doi: 10.1162/NECO_a_00463 PMID: 23607557

25. Kinney J, Tkačik G, Callan C (2007) Precise physical models of protein–DNA interaction from high-throughput data. Proceedings of the National Academy of Sciences 104: 501–506. doi: 10.1073/pnas.0609908104

26. Brenner N, Strong SP, Koberle R, Bialek W, de Ruyter van Steveninck RR (2000) Synergy in a neural code. Neural Computation 12: 1531–1552. doi: 10.1162/089976600300015259 PMID: 10935917

27. Maimon G, Assad JA (2009) Beyond Poisson: increased spike-time regularity across primate parietal cortex. Neuron 62: 426–440. doi: 10.1016/j.neuron.2009.03.021 PMID: 19447097

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

29. Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque V1 receptive fields. Neuron 46: 945–956. doi: 10.1016/j.neuron.2005.05.021 PMID: 15953422

30. Rust NC, Mante V, Simoncelli EP, Movshon JA (2006) How MT cells analyze the motion of visual patterns. Nature Neuroscience 9: 1421–1431. doi: 10.1038/nn1786 PMID: 17041595

31. Sugiyama M, Kawanabe M, Chui PL (2010) Dimensionality reduction for density ratio estimation in high-dimensional spaces. Neural Networks 23: 44–59. doi: 10.1016/j.neunet.2009.07.007 PMID: 19631506

32. Sugiyama M, Suzuki T, Kanamori T (2012) Density ratio estimation in machine learning. Cambridge University Press.

33. Suzuki T, Sugiyama M (2013) Sufficient dimension reduction via squared-loss mutual information estimation. Neural Computation 25: 725–758. doi: 10.1162/NECO_a_00407 PMID: 23272920

34. Vintch B, Zaharia A, Movshon JA, Simoncelli EP (2012) Efficient and direct estimation of a neural subunit model for sensory coding. In: Bartlett P, Pereira F, Burges C, Bottou L, Weinberger K, editors, Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp. 3104–3112.

35. Theis L, Chagas AM, Arnstein D, Schwarz C, Bethge M (2013) Beyond GLMs: A generative mixture modeling approach to neural system identification. PLoS Computational Biology 9. doi: 10.1371/journal.pcbi.1003356 PMID: 24278006

36. Rad KR, Paninski L (2010) Efficient, adaptive estimation of two-dimensional firing rate surfaces via Gaussian process methods. Network: Computation in Neural Systems 21: 142–168.

37. Park M, Horwitz G, Pillow JW (2011) Active learning of neural response functions with Gaussian processes. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K, editors, Advances in Neural Information Processing Systems 24, Curran Associates, Inc., pp. 2043–2051.

38. Rice J (1964) The approximation of functions: linear theory, volume 1. Addison-Wesley.

39. Park J, Sandberg I (1991) Universal approximation using radial-basis-function networks. Neural Computation 3: 246–257. doi: 10.1162/neco.1991.3.2.246

40. Korenberg M, Bruder S, Mcllroy P (1988) Exact orthogonal kernel estimation from finite data records: Extending Weiner's identification of nonlinear systems. Annals of Biomedical Engineering 16: 201–214. doi: 10.1007/BF02364581 PMID: 3382067

41. Victor J (1991) Asymptotic approach of generalized orthogonal functional expansions to Wiener kernels. Annals of Biomedical Engineering 19: 383–399. doi: 10.1007/BF02584317 PMID: 1741523

42. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Computation 14: 1771–1800. doi: 10.1162/089976602760128018 PMID: 12180402

43. Rowekamp R, Sharpee T (2011) Analyzing multicomponent receptive fields from neural responses to natural stimuli. Network: Computation in Neural Systems 7: 1–29.

44. Park IM, Pillow JW (2011) Bayesian Spike-Triggered Covariance Analysis. In: Shawe-Taylor J, Zemel RS, Bartlett P, Pereira F, Weinberger K, editors, Advances in Neural Information Processing Systems 24, Curran Associates, Inc. pp. 1692–1700.

45. Fitzgerald JD, Sincich LC, Sharpee TO (2011) Minimal models of multidimensional computations. PLoS Computational Biology 7: e1001111. doi: 10.1371/journal.pcbi.1001111 PMID: 21455284

46. Fitzgerald JD, Rowekamp RJ, Sincich LC, Sharpee TO (2011) Second order dimensionality reduction using minimum and maximum mutual information models. PLoS Computational Biology 7: e1002249. doi: 10.1371/journal.pcbi.1002249 PMID: 22046122

47. Christianson GB, Sahani M, Linden JF (2008) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. Journal of Neuroscience 28: 446–455. doi: 10.1523/JNEUROSCI.1775-07.2007 PMID: 18184787

48. Sahani M, Linden J (2003) Evidence optimization techniques for estimating stimulus-response functions. In: Becker S, Thrun S, Obermayer K, editors, Advances in Neural Information Processing Systems 15 MIT Press. pp. 317–324.

49. Park M, Pillow JW (2011) Receptive field inference with localized priors. PLoS Computational Biology 7: e1002219. doi: 10.1371/journal.pcbi.1002219 PMID: 22046110

50. Reich DS, Mechler F, Purpura KP, Victor JD (2000) Interspike intervals, receptive fields, and information encoding in primary visual cortex. Journal of Neuroscience 20: 1964–1974. PMID: 10684897

51. Barbieri R, Quirk MC, Frank LM, Wilson MA, Brown EN (2001) Construction and analysis of non-poisson stimulus-response models of neural spiking activity. Journal of Neuroscience Methods 105: 25–37. doi: 10.1016/S0165-0270(00)00344-7 PMID: 11166363

52. Pillow JW (2009) Time-rescaling methods for the estimation and assessment of non-poisson neural encoding models. In: Bengio Y, Schuurmans D, La erty J, Williams CKI, Culotta A, editors, Advances in Neural Information Processing Systems 22. Curran Associates, Inc., pp. 1473–1481.

53. Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. Neuron 58: 956–66. doi: 10.1016/j.neuron.2008.04.026 PMID: 18579084

54. Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. Proceedings of the National Academy of Sciences 106: 21894–21899. doi: 10.1073/pnas.0908383106

55. Sharpee TO, Atencio CA, Schreiner CE (2011) Hierarchical representations in the auditory cortex. Current Opinion in Neurobiology 21: 761–767. doi: 10.1016/j.conb.2011.05.027 PMID: 21704508

56. Atencio CA, Sharpee TO, Schreiner CE (2012) Receptive field dimensionality increases from the auditory midbrain to cortex. Journal of Neurophysiology 107: 2594–2603. doi: 10.1152/jn.01025.2011 PMID: 22323634

57. Fernandes HL, Stevenson IH, Phillips AN, Segraves MA, Kording KP (2013) Saliency and saccade encoding in the frontal eye field during natural scene search. Cerebral Cortex 24: 3232–3245.

58. Paninski L, Fellows M, Shoham S, Hatsopoulos N, Donoghue J (2004) Superlinear population encoding of dynamic hand trajectory in primary motor cortex. Journal of Neuroscience 24: 8551–8561. doi: 10.1523/JNEUROSCI.0919-04.2004 PMID: 15456829

59. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440: 1007–1012. doi: 10.1038/nature04701 PMID: 16625187