Models of neuronal stimulus-response functions: elaboration, estimation and evaluation

Arne F. Meyer [1], **Ross S. Williamson** [2] **Jennifer F. Linden** [3,4], **and Maneesh Sahani** [1,*]

[1] *Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom*
[2] *Eaton-Peabody Laboratories, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA*
[3] *Ear Institute, University College London, London, United Kingdom*
[4] *Department of Neuroscience, Physiology & Pharmacology, University College London, London, United Kingdom*

Correspondence*:
Maneesh Sahani
Gatsby Computational Neuroscience Unit, University College London, 25 Howland Street,
London, W1T 4JG, United Kingdom, maneesh@gatsby.ucl.ac.uk

ABSTRACT

Rich, dynamic, and dense sensory stimuli are encoded within the nervous system by the time-varying activity of many individual neurons. A fundamental approach to understanding the nature of the encoded representation is to characterise the function that relates the moment-by-moment firing of a neuron to the recent history of a complex sensory input. This review provides a unifying and critical survey of the techniques that have been brought to bear on this effort thus far — ranging from the classical linear receptive field model to modern approaches incorporating normalisation and other nonlinearities. We address separately the structure of the models; the criteria and algorithms used to identify the model parameters; and the role of regularising terms or "priors". In each case we consider benefits or drawbacks of various proposals, providing examples for when these methods work and when they may fail. Emphasis is placed on key concepts rather than mathematical details, so as to make the discussion accessible to readers from outside the field. Finally, we review ways in which the agreement between an assumed model and the neuron's response may be quantified. Re-implemented and unified code for many of the methods, and example data sets, are made freely available.

Keywords: Receptive Field, Sensory System, Neural Coding

INTRODUCTION

Sensory perception involves not only extraction of information about the physical world from the responses of various sensory receptors (e.g., photoreceptors in the retina and mechanoreceptors in the cochlea), but also the transformation of this information into neural representations that are useful for cognition and behaviour. A fundamental goal of systems neuroscience is to understand the nature of stimulus-response transformations at various stages of sensory processing, and the ways in which the resulting neural representations shape perception.

In principle, the stimulus-response transformation for a neuron or set of neurons could be fully characterised if all possible stimulus input patterns could be presented and neural responses measured for each of these inputs. In practice, however, the space of possible inputs is simply too large to be experimentally accessible. Instead, a common approach is to present a rich and dynamic stimulus that spans a sizeable subset of the possible stimulus space, and then use mathematical tools to estimate a model relating the sensory stimulus to the neural response that it elicits.

Such functional models, describing the relationship between sensory stimulus and neural response, are the focus of this review. Unlike biophysical models that seek to describe the physical mechanisms of sensory processing such as synaptic transmission and channel dynamics, functional models typically do not incorporate details of how the response is generated biologically. Thus, in functional models, the model parameters do not reflect physical properties of the biological system, but are instead abstract descriptors of the stimulus-response transformation. An advantage of this abstraction is that functional models can be versatile and powerful tools for addressing many different questions about neural representation.

Another advantage of the abstract nature of functional models is that the power of recent statistical advances in machine learning can be leveraged to estimate model parameters. In this context it is important to clearly distinguish between models and methods. A model describes the functional form of the stimulus-response function (SRF), i.e. how the stimulus is encoded into a neural response. A method (or algorithm) is then used to find parameters that best describe the given model. Usually, there are a number of different methods that can be used to fit a specific model.

Different methods used for model fitting will involve different specific assumptions. For example, constraints may be placed on the statistical structure that the stimulus must take, or the exact shape of the SRF. Changes in the assumptions can produce different estimates of model parameters, even when the method for fitting remains the same. Therefore, it is crucial to employ techniques that can explicitly quantify how well a given model captures neural response properties. Such a quantification serves as a means of determining whether the fitted model is capable of providing an appropriate description of the underlying stimulus-response transformation.

A major goal of this review is to disentangle the existing arsenal of SRF models and estimation methods, and to provide examples that highlight when they work and when they fail. The first part of the review focuses on describing the different SRF models along with various methods that can be used to fit them. The second part of the review then describes techniques that can be used to evaluate the fitted models.

Statistical preliminaries

Although the subtleties of hypothesis testing (such as the statistics of multiple comparisons) are widely appreciated in the biological sciences, subtleties of model estimation are rarely discussed, even though the corresponding statistical theory is well-developed. Therefore, it will be useful to define some statistical concepts and terms at the outset of this review. Most models explicitly or implicitly define a probability

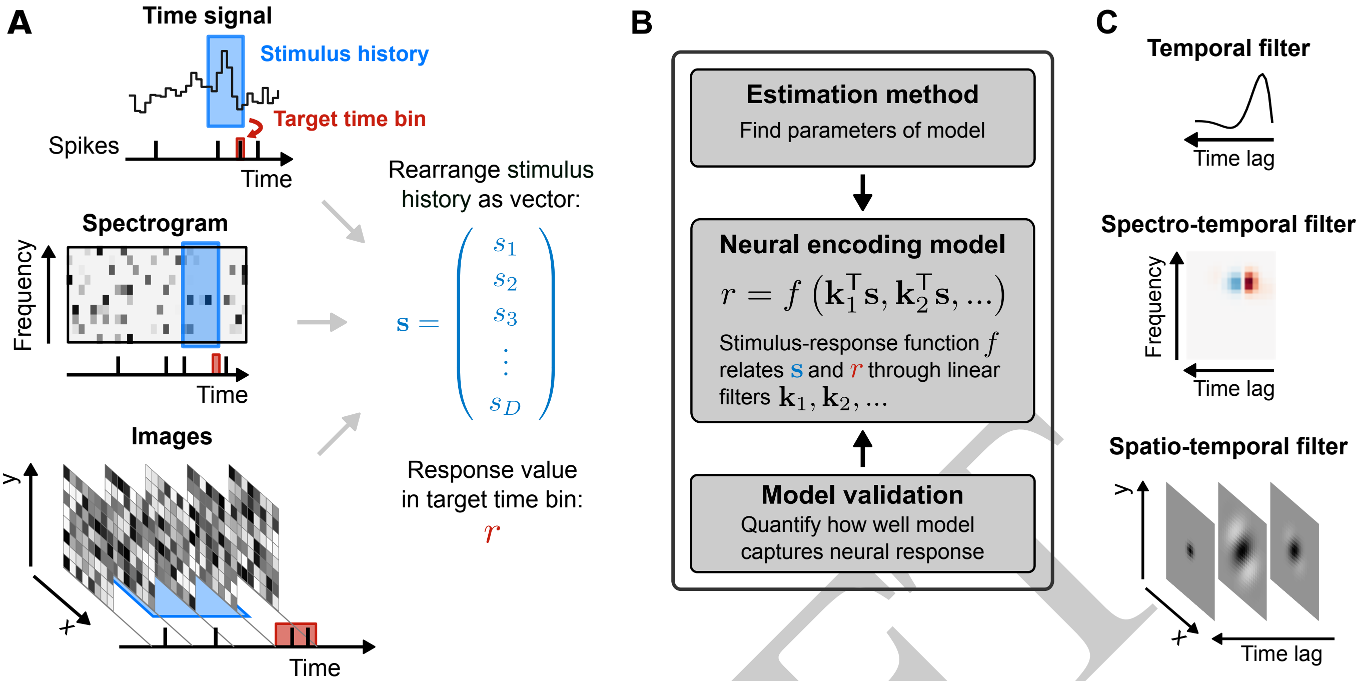
A
Time signal
Stimulus history
Target time bin
Spikes
Time
Spectrogram
Frequency
Time
Images
y
x
Time
Rearrange stimulus history as vector:
$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_D \end{pmatrix}$
Response value in target time bin:
r
B
Estimation method
Find parameters of model
Neural encoding model
$r = f\left(\mathbf{k}_1^\mathsf{T}\mathbf{s}, \mathbf{k}_2^\mathsf{T}\mathbf{s}, \ldots\right)$
Stimulus-response function f relates $\mathbf{s}$ and r through linear filters $\mathbf{k}_1, \mathbf{k}_2, \ldots$
Model validation
Quantify how well model captures neural response
C
Temporal filter
Time lag
Spectro-temporal filter
Frequency
Time lag
Spatio-temporal filter
y
x
Time lag

Figure 1. Sensory stimulus representation for stimulus–response functions. (**A**) Stimulus examples are sampled from the sensory stimulus representation, e.g., a time signal (top), a spectrogram (bottom), or a sequence of image patches (bottom), by rearranging the stimulus history (blue rectangle) as vector **s**. The spike response is usually binned at the temporal resolution of the stimulus, with the target spike bin indicated by the red rectangle. (**B**) The stimulus–response function describes the functional relationship between presented stimulus and measured response. In the models considered here, stimulus and response are related by a linear projection of the stimulus onto one or more linear filters $\mathbf{k}_1, \mathbf{k}_2, \ldots$. These filters represent the receptive field of the neuron. (**C**) Once the best parameters for the model have been identified, the representation of the linear filters in the original stimulus space can be interpreted as an an estimate of the stimulus sensitivities of the neuron. Examples of single filters are shown for each type of stimulus representation in **A**.

distribution of responses, given a stimulus and some parameters such as a tuning curve, or the weights of a receptive field. By evaluating the probability of the observed data under this distribution, for a known stimulus but varying parameters, we obtain the *likelihood* function over the model parameters. The parameter values which maximise this function, and thus the probability of the observed data, form the *maximum likelihood estimator* (MLE).

The MLE is not the only possible estimator, and we will sometimes discuss more than one way to estimate the parameters of the same model. An estimator is often evaluated in terms of its *bias* (the expected difference between an estimate based on a data set and the parameter value that actually generated those data), its *expected squared error* (bias squared plus variance), and its *consistency* (whether the bias and variance approach 0 when based on increasing amounts of data). However, it is important to realise that bias, variance and consistency are statistical confections. They only have meaning when data actually arise from a model of the form under consideration. Real neural data will *never* be completely and accurately described by abstract models of the type we discuss here; at best we expect the models to provide a decent approximation to the truth. Thus, while consistency and lack of bias are certainly characteristics of a good estimator, these favourable statistical features do not demonstrate “optimality” even within the assumed model form; the estimator may not select the parameters that provide the best model approximation to data generated by a different process.

Practical proof lies elsewhere, in predictive accuracy: how well can the parameters estimated predict a new response that was not used in the estimation process? This is often assessed by cross-validation. A data set is divided into segments; model parameters are estimated leaving out one of the segments; and the predictive quality of the model fit is evaluated on the segment left out. This procedure can be repeated leaving out each segment in turn and the prediction accuracy averaged to yield a more reliable number.

Ultimately, predictive measures such as these (sometimes in more elaborate guises discussed below) are needed to evaluate the quality of both model *and* estimator. Indeed, many pitfalls of interpretation can be avoided by remembering that all models are wrong, and so the only approachable question is: which one is most useful?

PART 1: ELABORATION AND ESTIMATION

Receptive-field-based stimulus–response function models

A stimulus–response function (SRF) model parametrises the response of a neural system to a rich input stimulus: usually a random, pseudo-random or natural sensory stimulus sequence presented under controlled conditions. Although many aspects of system response may be modelled — including behaviour, metabolic activity, and local field or surface potentials — we focus here on models that target the activity of individual neurons at the level of action potentials (“spikes”), membrane potential or cytoplasmic calcium concentration. Furthermore, we focus on SRF models that include one or more “spatiotemporal” linear filters. These filters encode the way in which the neural response integrates elementary inputs, say light at a point in the visual space or power at an acoustic frequency, over time and sensory space. In a sense, then, these filters represent estimates of the receptive field (RF) properties of a neuron, with each filter indicating a “dimension” or “feature” of the stimulus to which it is sensitive.

The choice of stimulus depends on the sensory modality being investigated and the specific question at hand. However, many stimuli can be represented in a common vector-based format, and then very similar, sometimes even identical, models and estimation methods can be applied across modalities to address a variety of questions. Stimulus sequences are usually represented in discretised time, at a rate dictated by the sampling frequency of the stimulus or else re-sampled to match the timescale on which the neural response varies. For simplicity, we assume that the response is measured with the same temporal precision as the stimulus.

The RF components of an SRF model are most often taken to have limited extent in time (technically, the impulse-response of the filters is finite). Thus, the input used by the model at time t, to describe the response $r(t)$, is limited to a “window” of stimulus values stretching from t to some maximal delay $\tau_{\max}$ time-steps in the past. The stimulus at each time in this window may be multidimensional, with one value for each pixel in an image, or each frequency band in a spectrogram. It is convenient to collect all such values falling with the window anchored at time t into a single (column) vector $\mathbf{s}(t) = (s_1(t), s_2(t), ..., s_D(t))^{\mathsf{T}}$, with dimension $D =$ (length of window) × (dimension of single stimulus frame). Thus, $s_1(t)$ might represent the power in the lowest audio frequency channel at time t, $s_{64}(t)$ the power in the highest frequency channel also at t, $s_{65}(t)$ low-frequency power at $t - 1$ and (say) $s_{640}(t)$ high-frequency power at $t - 9$. The process is illustrated in Figure 1 for different types of stimuli.

The discrete-time vector representation of the stimulus allows us to write the action of a single multi-channel linear filter as a inner or “dot” product between the stimulus vector and a vector of filter weights $\mathbf{k}$

arranged in the same way:

$$\mathbf{k}^{\mathsf{T}}\mathbf{s}(t) = \sum_{i=1}^{D} k_i s_i(t) = k_1 s_1(t) + k_2 s_2(t) + \dots + k_D s_D(t)\,, \tag{1}$$

thus providing a short-hand notation for integration over space (or channel) as well as over time. The filter **k** is often called a spatio-temporal or spectro-temporal receptive field (STRF) and the weights within it indicate the sensitivity of the neuron to inputs at different points of stimulus space and stimulus history.

Such discrete-time finite-window vector filtering lies at the heart of the majority of SRF models that have been explored in the literature, although these models may vary in the range of nonlinear transformations that they chain after or before the filtering process to form a "cascade". The cascades range from a simple point-by-point nonlinear transformation that acts on the output of a single linear filter — the linear-nonlinear or LN cascade often employed at earlier sensory stages — to more complicated series or parallel arrangements of filters with multiple intervening nonlinear functions. Some cascades are inspired by a feed-forward description of the sensory pathway, with architectures that recapitulate pathway anatomy. Nonetheless, the assumptions that integration within each stage is linear, often that the nonlinear functions fall within a constrained class, and particularly that responses do not depend on internal state or recurrence, mean that even anatomically-inspired SRF models should be regarded as abstract functional models of computation rather than as biologically plausible models of mechanism.

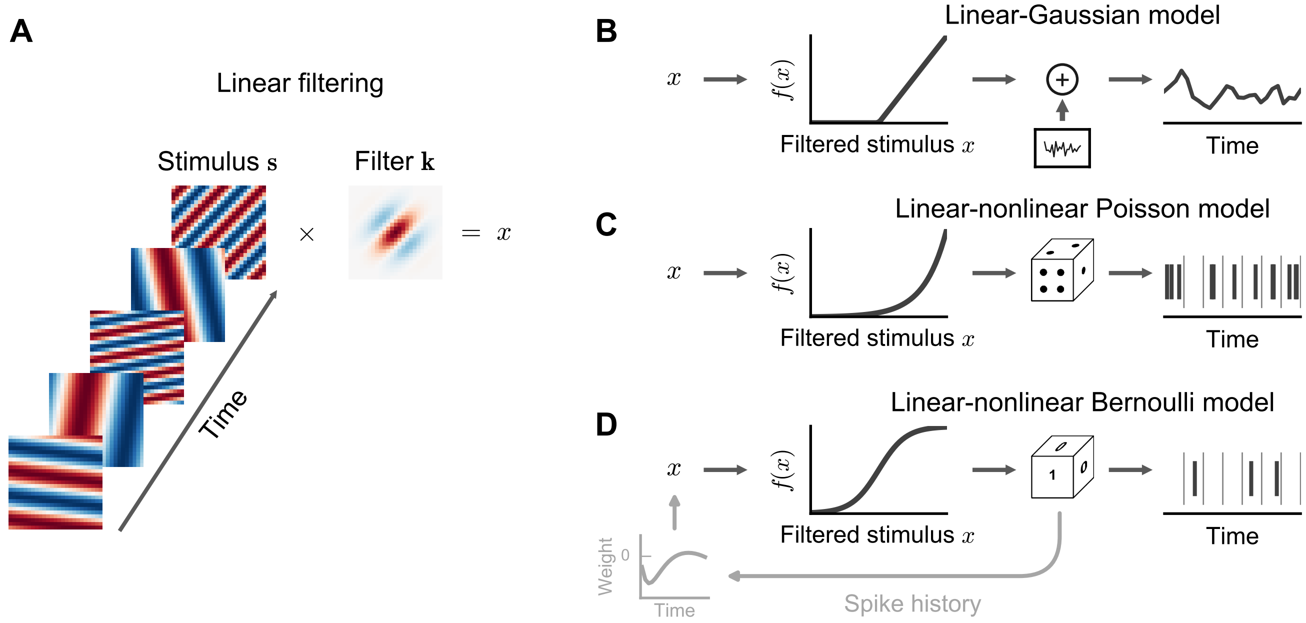

Figure 2. Common stimulus–response functions. A Filtering of stimulus examples through the linear filter k. **B** (Threshold-)Linear model with Gaussian noise. **C** Poisson model with exponential nonlinearity. **D** Bernoulli model. All models can be extended using a post-spike filter that indicates dependence of the model's output on the recent response history (light grey).

The linear-Gaussian model

In the simplest case the response is assumed to be modelled directly by the output of a single filter, possibly with a constant offset response:

$$r(t) \approx k^{(0)} + \mathbf{k}^{\mathsf{T}}\mathbf{s}(t)\,. \tag{2}$$

The constant offset $k^{(0)}$ can be conveniently absorbed into the RF vector $\mathbf{k}$ by setting an additional dimension in the stimulus vector $\mathbf{s}(t)$ to 1 at all times, so that the offset becomes the coefficient associated with this added dimension. Thus, we will typically omit explicit reference to (and notation of) the offset term.

In practice, most neurons do not respond the same way each time the same stimulus sequence is repeated, and so even if Eq. (2) were a correct model of the *mean* response, the actual response measured on one or a finite number of trials will almost surely be different. We reserve the notation $r(t)$ for the measured response and write $\hat{r}(t)$ for the SRF model prediction, so that for the linear model $\hat{r}(t) \equiv \mathbf{k}^{\mathsf{T}}\mathbf{s}(t)$.

Given a stimulus and a measured response, estimated filter weights $\widehat{\mathbf{k}}$ can be obtained by minimising the squared difference between the model output and the measured data:

$$\widehat{\mathbf{k}} = \operatorname*{argmin}_{\mathbf{k}} \sum_t \|r(t) - \mathbf{k}^{\mathsf{T}}\mathbf{s}(t)\|^2 = \left(\mathbf{S}^{\mathsf{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathsf{T}}\mathbf{r}\,, \tag{3}$$

where $\mathbf{S}$ is the stimulus design matrix formed by collecting the stimulus vectors as rows, $\mathbf{S} = (\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(T))^{\mathsf{T}}$, and $\mathbf{r}$ is a column vector of corresponding measured responses. The right-hand expression in Eq. (3) has a long history in neuroscience (Marmarelis and Marmarelis, 1978), and may be interpreted in many ways. It is the solution to a least-squares regression problem, solved by taking the Moore-Penrose pseudoinverse of $\mathbf{S}$; it is a discrete time version of the Wiener Filter; and, for spike responses, it may be seen as a scaled "correlation-corrected" spike-triggered average (deBoer and Kuyper, 1968; Chichilnisky, 2001). This latter interpretation follows as the matrix product $\mathbf{S}^{\mathsf{T}}\mathbf{r}$ gives the sum of all stimuli that evoked spikes (with stimuli evoking multiple spikes repeated for each spike in the bin); if divided by the total number of spikes this would be the spike-triggered average (STA) stimulus. The term $\mathbf{S}^{\mathsf{T}}\mathbf{S}$ is the stimulus auto-correlation matrix; pre-multiplying by its inverse removes any structure in the STA that might arise from correlations between different stimulus inputs, leaving an estimate of the SRF filter. In this way, the estimated model filter corresponds to a descriptive model of the receptive-field obtained by "reverse correlation" (deBoer and Kuyper, 1968) or "white noise analysis" (Marmarelis and Marmarelis, 1978).

Thus, the linear SRF model is attractive for its analytic tractability; its computational simplicity (although see the discussion of regularisation below); and its interpretability.

If the mean response of the neuron were indeed a linear function of the stimulus, then linear regression would provide an unbiased estimate of the true RF parameters, regardless of the statistical structure of the stimulus ensemble (Paninski, 2003a) and the nature of the neural response variability. More generally, Eq. (3) corresponds to the MLE (see Statistical preliminaries) for a model in which response variability is Gaussian-distributed with constant variance around the filter output:

$$r(t) = \mathbf{k}^{\mathsf{T}}\mathbf{s}(t) + \varepsilon(t), \quad \varepsilon(t) \sim \mathcal{N}(0, \sigma^2)\,. \tag{4}$$

By itself, this MLE property is of limited value in this case. The assumption of Gaussian response noise is inappropriate for single-trial spike counts, although it may be better motivated when the responses being modelled are trial-averaged mean rates (Theunissen et al., 2000; Linden et al., 2003), subthreshold membrane potentials (Machens et al., 2004), local field potentials (Mineault et al., 2013), or intracranial electrocorticographical recordings (Mesgarani and Chang, 2012); but even then the assumption of constant variance may be violated. Instead, the value of the probabilistic interpretation lies in access to a principled theory of stabilised (or "regularised") estimation, and to the potential generalisation to nonlinear and non-Gaussian modelling assumptions, both of which we discuss below.

The linear-nonlinear (LN) cascade model

Although valuable as a first description, a linear function rarely provides a quantitatively accurate model of neural responses (e.g., Sahani and Linden, 2003b; Machens et al., 2004). Particularly for spiking responses, an attractive extension is to assume that a linear process of stimulus integration within the RF is followed by a separate nonlinear process of response generation. This leads to the linear-nonlinear (or LN) cascade model:

$$\hat{r}(t) = f\left(\mathbf{k}^{\mathsf{T}}\mathbf{s}(t)\right), \tag{5}$$

where f is a static, memoryless nonlinear function. Unlike some more general nonlinear models described later, the input to the nonlinear stage of this LN cascade is of much lower dimension than the stimulus within the RF. Indeed, in Eq. (5) it is a single scalar product — although multi-filter versions are discussed below. This reduction in dimensionality allows both the parameters describing the RF filter $\mathbf{k}$ and any that describe the nonlinearity f to be estimated robustly from fewer data than would be required in the more general case.

Indeed, perhaps surprisingly, the linear estimator of Eq. (3) may sometimes also provide a useful estimate of the linear-stage RF within an LN model (Bussgang, 1952). To understand when and why, it is useful to visualise the analysis geometrically (Figure 3). Each stimulus vector is represented by a point in a D-dimensional space, centred such that origin lies at the mean of the stimulus distribution. Stimuli are coloured according to the response they evoke; for spike responses, this distinguishes stimuli associated with action potentials — the "spike-triggered" ensemble — from the "raw" distribution of all stimuli. An RF filter is also a D-dimensional vector, and so defines a direction within the space of stimuli. If the neural response can in fact be described by an LN process (with any variability only depending on the stimulus through the value of $\hat{r}(t)$), then by Eq. (5) the stimulus-evoked response will be fully determined by the orthogonal projection of the D-dimensional stimulus point onto this RF direction through the dot-product $\mathbf{k}^{\mathsf{T}}\mathbf{s}(t)$. Thus, averaging over response variability, the contours defining "iso-response" stimuli will be (hyper)planes perpendicular to the true RF direction.

Now, if the raw stimulus distribution is free of any intrinsic directional bias (that is, it is invariant to rotations about any axis in the D-dimensional space, or "spherically symmetric"), the distribution in any such iso-response plane will also be symmetric, so that its mean falls along the RF vector $\mathbf{k}$. It follows that the response-weighted mean of all stimuli lies along this same direction, and thus (as long as f is not a symmetric function) the empirical response-weighted average stimulus provides an unbiased estimate of the RF. For spike responses, this response-weighted stimulus mean is the STA (Figure 3**A**). The result can be generalised from spherically symmetric stimulus distributions (Chichilnisky, 2001) to those that can be linearly transformed to spherical symmetry (that is, are elliptically symmetric) (Paninski, 2003a), for which the "correlation-corrected" STA estimator of Eq. (3) is consistent.

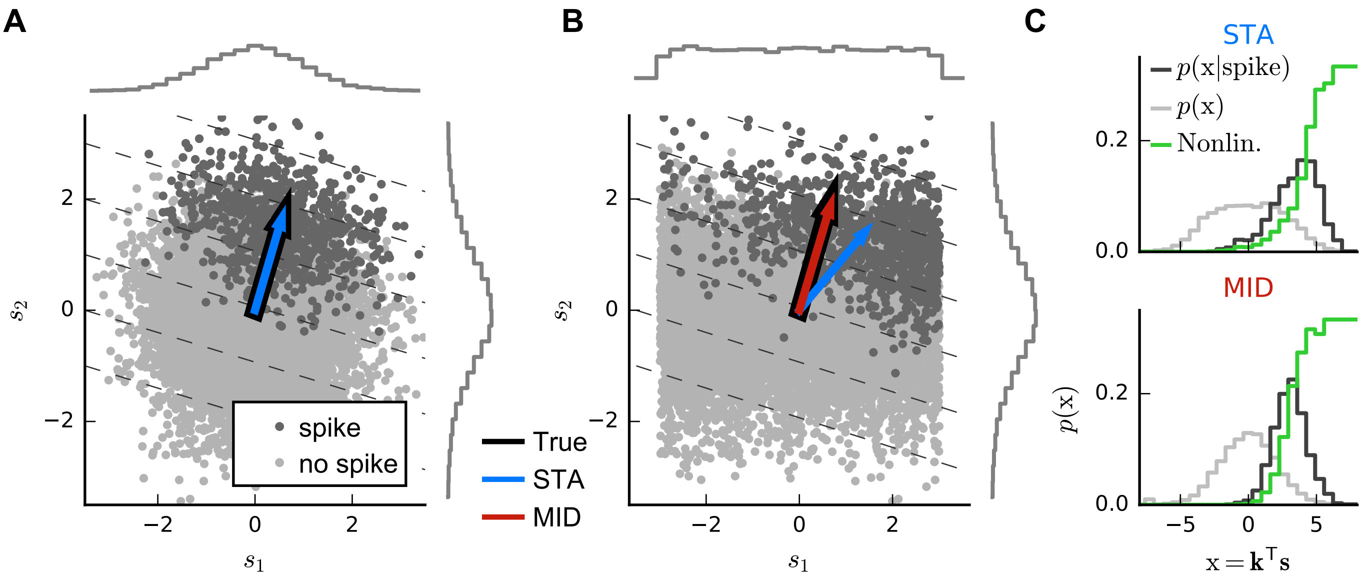

Figure 3. Geometric illustration of linear filter estimation in the LN model. (**A**) A two-dimensional stimulus sampled from a Gaussian distribution. Points indicate spike-eliciting (dark grey) and non-spike-eliciting (light grey) stimulus examples with true linear filter shown by the black arrow. For a Gaussian (or more generally, a spherically symmetric) stimulus, the spike-triggered average (STA; blue arrow), given by the mean of all spike-triggered stimuli, recovers the true linear filter. Histograms (insets) show the marginal distributions of stimulus values along each stimulus dimension. Dashed lines indicate "iso-response" hyperplanes (see main text). (**B**) The same as in **A** except that stimulus dimension s_1 follows a uniform distribution, resulting in a non-spherically symmetric stimulus distribution. The STA no longer points in the same direction as the true linear filter but the maximally informative dimensions (MID; red arrow) estimator is robust to the change in the stimulus distribution. (**C**) Spike-conditional distribution ($p(x|\text{spike})$, raw distribution ($p(x)$) of filtered stimuli, and histogram-based estimates of the spiking nonlinearity (solid green line) for the STA (top) and MID (bottom) for the example in **B**. MID seeks the filter that minimises the overlap between these distributions. The spiking nonlinearity has been rescaled for visualisation.

The symmetry conditions are important to these results. Even small asymmetries may bias estimates away from the true RF as the more heavily-sampled regions of the stimulus ensemble are over-weighted in the STA (Figure 3**B**). With more structured stimulus distributions, including "natural" movies or sounds, the effects of the bias in the STA-based estimators may be profound and misleading. For such stimuli, estimation of an LN model depends critically on assumptions about the functional form of the nonlinearity f and the nature of the variability in the response $r(t)$.

One intuitive approach is provided by information theory. Consider a candidate RF direction defined by vector $\tilde{\mathbf{k}}$, and let $\tilde{s} = \tilde{\mathbf{k}}^\mathsf{T}\mathbf{s}$ be the projection of a stimulus point $\mathbf{s}$ onto this direction. Again making the assumption that the true neural response (and its variability) depends only on the output of an LN process, the predictability of the neural response from $\tilde{s}$ will be maximal and equal to the predictability from the full stimulus vector $\mathbf{s}$ if and only if $\tilde{\mathbf{k}}$ is parallel to the true RF. This predictability can be captured by the mutual information between $\tilde{s}$ and the response, leading to the maximally informative dimensions (MID) estimation approach (Sharpee et al., 2004): identify the direction $\tilde{\mathbf{k}}$ for which the empirical estimate of the mutual information between $\tilde{s}(t)$ and the measured responses $r(t)$ is maximal.

While this basic statement is independent of assumptions about the nonlinearity or variability, the challenges of estimating mutual information from empirical distributions (Paninski, 2003b) mean that MID-based approaches invariably embody such assumptions in their practical implementations.

Parametric models for spike counts: linear-nonlinear-Poisson (LNP)

For spike-train responses, a natural first assumption is that spike times are influenced only by the stimulus, and are otherwise entirely independent of one another. This assumption requires that the distribution of spike times be governed by a Poisson (point) *process* conditioned on the stimulus, defined by an instantaneous rate function $\lambda(t)$. In turn, this means that the distributions of counts within response time bins of size Δ must follow a Poisson *distribution*:

$$P(r(t)|\mathbf{s}(t)) = \frac{1}{r(t)!} e^{-\lambda(t)\Delta} (\lambda(t)\Delta)^{r(t)}; \qquad \lambda(t) = f(\mathbf{k}^{\mathsf{T}}\mathbf{s}(t)) \,. \tag{6}$$

The most widely used definition of the MID is based on this assumption of spike-time independence. Again, letting $\tilde{\mathbf{k}}$ be a candidate RF direction, and $\tilde{s}$ the value of the projected stimulus, Sharpee et al. (2004) showed that the mutual information between the projected stimuli and independent (and so Poisson-distributed) spikes can be written as a Kullback-Leibler divergence D_{KL} between the spike-triggered distribution of projected stimuli, $p(\tilde{s}|\text{spike})$ and the raw distribution $p(\tilde{s})$:

$$I(\tilde{\mathbf{k}}) = D_{KL}\left[p(\tilde{s}|\text{spike})||p(\tilde{s})\right] = \int p(\tilde{s}|\text{spike}) \log \frac{p(\tilde{s}|\text{spike})}{p(\tilde{s})} \mathrm{d}\tilde{s} \,. \tag{7}$$

The spike-triggered and raw distributions must themselves be estimated to evaluate $I(\tilde{\mathbf{k}})$ and so to identify the MID. The common choice is to estimate each distribution by constructing a binned histogram; and so, in effect, the MID is defined to be the direction along which the histogram of the projected spike-triggered ensemble differs most from the raw stimulus histogram (Figure 3**BC**).

Despite the information theoretic derivation, the Poisson-based information definition combined with histogram-based probability estimates makes the conventional MID approach mathematically identical to a likelihood-based method. Specifically, the histogram-based MID estimate equals the MLE of an LNP model in which the nonlinearity f is assumed to be piece-wise constant within intervals that correspond to the bins of the MID histograms (Figure 3C) (Williamson et al., 2015). A corollary is that if these assumptions do not hold, then this form of MID may also be biased. In practice, the approach is also complicated by the fragility of histogram-based estimates of information theoretic quantities, and by the fact that the objective function associated with such a flexible nonlinearity may have many non-global local maxima, making the true optimum difficult to discover.

Alternative approaches, based either on information theory or on likelihood, assume more restrictive forms of the nonlinearity.

For instance, assuming a Gaussian form for the distributions $p(\tilde{s}|\text{spike})$ and $p(\tilde{s})$ in Eq. (7), leads to an estimation procedure that combines both the STA and the spike-triggered-stimulus covariance (STC; see Multi-filter models) to identify the RF direction. This has been called "information-theoretic spike-triggered average and covariance" (iSTAC) analysis (Pillow and Simoncelli, 2006). Again, there is a link to a maximum likelihood estimate (this time assuming an exponentiated quadratic nonlinearity) although in this case equivalence only holds if the raw spike distribution is indeed Gaussian, and then too only in the limit as the number of stimuli grows to infinity.

If f is assumed to be monotonic and fixed (rather than being defined by parameters that must be fit along with the RF) then Eq. (6) describes an instance of a generalised linear model (GLM) (Nelder and Wedderburn, 1972), a widely-studied class of regression models. Many common choices of f result in a likelihood which is a convex function (Paninski, 2004), guaranteeing the existence of a single optimum that is easily found by standard convex optimisation techniques such as gradient ascent or Newton's method (see Parameter optimisation). The GLM formulation is also easy to extend to non-Poisson processes, by including probabilistic interactions between spikes in different bins that may be often reminiscent of cellular biophysical processes (see Interactions between bins).

Non-Poisson count models

The LNP model assumes that the exact times of individual spikes, whether in the same or different bins, are entirely statistically independent once their stimulus-dependence has been taken into account. While simple, this assumption is rarely biologically justified. Many biophysical and physiological processes lead to statistical dependence between spike times on both short and long timescales. These include membrane refractoriness, spike-rate adaptation, biophysical properties that promote bursting or oscillatory firing, and auto-correlated network input that fluctuates independently of the stimulus. Similar observations apply to other response measures — even to behavioural responses which exhibit clear decision-history dependence (Busse et al., 2011).

Bernoulli models

The refractoriness of spiking has a strong influence on counts within short time bins. Indeed, when the bin size corresponds to the absolute refractory period (around 1 ms), the observed spike-counts will all be either 0 or 1. If the spike probability is low, the difference between Poisson and binary predictions will be small, and so LNP estimators may still succeed. However, as the probability of spiking in individual bins grows large, an LNP-based estimator (such as MID or the Poisson GLM) may give biased results (Fig. 4).

For such short time bins, or for situations in which trial-to-trial variability in spike count is much lower than for a Poisson process (DeWeese et al., 2003), a more appropriate LN model will employ a Bernoulli distribution over the two possible responses $r(t) \in \{0, 1\}$:

$$\begin{aligned} \lambda(t) &= \frac{1}{\Delta} f(\mathbf{k}^{\mathsf{T}} \mathbf{s}(t)) \\ p(r(t)|\lambda(t)) &= (\lambda(t)\Delta)^{r(t)} \, (1 - \lambda(t)\Delta)^{1-r(t)} \ , \end{aligned} \tag{8}$$

where $\lambda(t)\Delta$ is now a probability between 0 and 1, and so the maximum possible rate is given by $1/\Delta$. As for the LNP model, the parameters of this linear-nonlinear-Bernoulli (LNB) model can be estimated using maximum-likelihood. The function f may be chosen to be piece-wise constant, giving an Bernoulli-based equivalent to the MID approach (Williamson et al., 2015). Alternatively, it may be a fixed, often sigmoid function with values between 0 and 1. In particular, if f is the logistic function, the LNB model corresponds to the GLM for logistic regression.

An alternative approach to estimation of the parameters of a binary encoding model is to reinterpret the problem as a classification task in which spike-eliciting and non-spike-eliciting stimuli are to be optimally discriminated (Meyer et al., 2014a). This approach is discriminative rather than probabilistic, and the model can be written as

$$r(t) = \mathrm{H}\left(\mathbf{k}^{\mathsf{T}} \mathbf{s}(t) - \eta + \varepsilon(t)\right) \tag{9}$$

where η is a spiking threshold and $\varepsilon(t)$ a random variable reflecting noise around the threshold. H is the step function which evaluates to 1 for positive arguments, and 0, otherwise. In this formulation, the RF vector $\mathbf{k}$ appears as the weight vector of a standard linear classifier. Optimal weights are determined by minimising

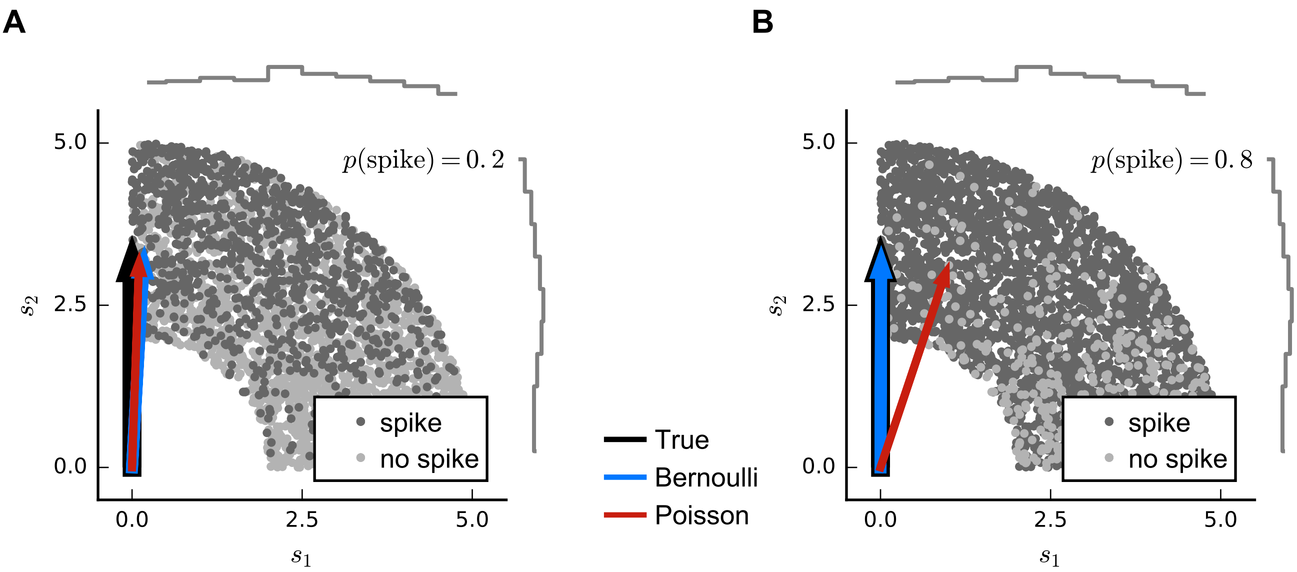

Figure 4. Simulated example illustrating failure of the Poisson model for Bernoulli distributed responses. (**A**) $N = 5000$ stimuli were drawn from a uniform distribution on a circular ring. A Bernoulli spike train with $p(\text{spike}) = 0.2$ was generated after filtering the 2D stimulus with a RF pointing along the y-axis and a subsequent sigmoid static nonlinear function. Both Poisson GLM (red arrow) and Bernoulli GLM (blue arrow) reliably recover the true filter (black arrow). (**B**) Same as in **A** but for $p(\text{spike}) = 0.8$. The Poisson GLM estimator fails to recover the true linear filter because its neglects information from silences which are more informative when $p(\text{no spike}) = 1 - p(\text{spike})$ is low (see text). The Bernoulli GLM accounts for silences and thus reliably reconstructs the true linear filter.

a cost function that depends on the locations of spike-labelled stimuli relative to the resulting classification boundary. Robust classifiers often favour a large *margin*; that is, they set the classification boundary so that stimuli that fall nearby are classified as spike-eliciting or not with as little ambiguity as possible. This large-margin approach can be seen as a form of regularisation (see the section on Regularisation below). Meyer et al. (2014a) report that a large-margin classifier returns robust RF estimates for simulated data generated using a wide range of different neural nonlinearities, while a point-process GLM is more sensitive to mismatch between the nonlinearity assumed by the model and that of the data — particularly when working with natural stimuli. On the other hand, the loss function associated with logistic regression (a binary-ouput GLM) also favours large margins (Rosset et al., 2003) and results for the simulations shown here for the Bernoulli model were virtually identical to those obtained using the classification-based approach described by Meyer et al. (2014a) (data not shown).

Over-dispersed and general count models

Longer bins, for example those chosen to match the refresh rate of a stimulus, may contain more than one spike; but even so the expected distribution of binned counts in response to repeated presentations of the same stimulus will not usually be Poisson.

One form of non-Poisson effect may result from the influence of variability in the internal network state (for instance the “synchronised” and “desynchronised” states of cortical activity; Harris and Thiele 2011), which may appear to multiplicatively scale the mean of an otherwise Poisson-like response. This additional variance leads to *over-dispersion* relative to the Poisson; that is the Fano factor (variance divided by the mean) exceeds 1. Such over-dispersion within individual bins may be modelled using a “negative binomial” or Polya distribution (Scott and Pillow, 2012). However, the influence of such network effects often extends over many bins or many cells (if recorded together), in which case it may be better modelled as an explicit unobserved variable contributing correlated influence.

More generally, for moderate-length bins where the maximal possible spike count is bounded by refractoriness, the neural response may be described by an arbitrary distribution over the possible count values $j \in \{0, ..., r_{\text{max}}\}$. A linear-nonlinear-count (LNC) model can then be defined as:

$$\lambda^{(j)}(t) = f^{(j)}\big(\mathbf{k}^{\mathsf{T}}\mathbf{s}(t)\big)$$
$$p\big(\, r(t){=}j \mid \lambda^{(j)}(t)\,\big) = \lambda^{(j)}(t) \tag{10}$$

with the added constraint on the functions $f^{(j)}$ that $\sum_{j=0}^{r_{\text{max}}} f^{(j)}(x) = 1$ for all x, to ensure that the probabilities over the different counts sum to 1 for each stimulus. This model includes the LNB model as a special case and, as before, the model parameters can be estimated using maximum-likelihood. Furthermore, if the functions f are assumed to be piece-wise constant, the LNC model estimate of $\mathbf{k}$ corresponds to a non-Poisson information maximum analogous to the MID. Thus, there is a general and exact equivalence between likelihood-based and information-based estimators for each LN structure (Williamson et al., 2015).

Interactions between bins

If responses are measured in short time-bins then longer-term firing interactions such as adaptation, bursting or intrinsic membrane oscillations will induce dependence between counts in different bins. In general, any stimulus-dependent point process can be expressed in a form where the instantaneous probability of spiking depends jointly on the stimulus history and the history of previous spikes, although the spike-history dependence might not always be straightforward. However, a useful approach is to assume a particular parametric form of dependence on past spikes, essentially incorporating these as additional inputs during estimation.

This formulation is perhaps most straightforward within the GLM framework(Chornoboy et al., 1988; Truccolo et al., 2005). For a fixed nonlinearity $f()$ we have

$$\lambda(t) = f(\mathbf{k}^{\mathsf{T}}\mathbf{s}(t) + \mathbf{g}^{\mathsf{T}}\mathbf{h}(t))$$

where $\mathbf{g}$ is a vector of weights and $\mathbf{h}(t)$ is a vector representing the history of spiking at time prior to time t; this may be a time-window of response bins stretching some fixed time into the past (as for the stimulus) or may be the outputs of a fixed bank of filters which integrate spike history on progressively longer timescales.

In effect, the combination of $\mathbf{g}$ and any filters that define $\mathbf{h}$ serves to implement a “post-spike” filtered input to the intensity function. It is tempting to interpret such filters biophysically as action-potential related influences on the membrane potential of the cell; indeed this model may be seen as a probabilistic version of the spike-response model of Gerstner and Kistler (2002). Suitable forms of post-spike filters may implement phenomena such as refractoriness, bursting or adaptation.

Multi-filter models

Many LN models can be generalised to incorporate multiple filters acting within the same RF, replacing the single filter $\mathbf{k}$ by a the matrix $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, ...]$ where each column represents a different filter (Figure 5. Conceptually, each of these filters may be understood to describe a specific feature to which the neuron is sensitive, although in many cases it is only the subspace of stimuli spanned by the matrix $\mathbf{K}$ which can be determined by the data, rather than the specific filter shapes themselves. In general, the assumptions embodied in the model or estimators, e.g., regarding the statistical structure of the stimulus, are similar to those made for the single-filter estimation. In particular, the directions in stimulus space (in the sense of Figure 3) along which the spike-triggered covariance (STC) of the stimulus vectors differs from the overall covariance of all stimuli used in the experiment provides one estimate of the columns of $\mathbf{K}$ in an LNP model (Brenner et al., 2000). This approach to estimation is often called STC analysis. The STC estimate is unbiased provided the overall stimulus distribution is spherically or elliptically symmetric (as was the case for the STA estimator of a single-filter model) *and* the stimulus dimensions are independent or can be linearly transformed to be independent of each other (Paninski, 2003a; Schwartz et al., 2006). These conditions are met only by a Gaussian stimulus distribution, and in other cases the bias can be very significant (Paninski, 2003a; Fitzgerald et al., 2011a)

The MID approach can also be extended to the multi-filter LNP case, defining a subspace projection for a candidate matrix $\tilde{\mathbf{K}}$ to be $\tilde{\mathbf{s}}(t) = \tilde{\mathbf{K}}^{\mathsf{T}}\mathbf{s}(t)$ and adjusting $\tilde{\mathbf{K}}$ to maximise the Kullback-Leibler divergence between the distributions $p(\tilde{\mathbf{s}}|\text{spike})$ and $p(\tilde{\mathbf{s}})$. Unfortunately, estimation difficulties make it challenging to use MID to robustly estimate the numbers of filters that might be needed to capture realistic responses (Rust et al., 2005). The problem is not the number of filter parameters *per se* (these scale linearly with stimulus dimensionality), but rather the number of parameters that are necessary to specify the densities $p(\tilde{\mathbf{s}})$ and $p(\tilde{\mathbf{s}}|\text{spike})$. For common histogram-based density estimators, the number of parameters grows exponentially with dimension (m bins for p filters requires m^p parameters), e.g., a model with four filters and 25 histogram bins would require fitting 390625 parameters, a clear instance of the “curse of dimensionality”.

In this context, the likelihood-based LN approaches may provide more robust estimates. Rather than depending on estimates of the separate densities, the LN model framework directly estimates a single nonlinear function $f(\tilde{\mathbf{s}})$. This immediately halves the number of parameters needed to characterise the relationship between $\tilde{\mathbf{s}}$ and the response. Furthermore, for larger numbers of filters, f may be parametrised using sets of basis functions whose numbers grow less rapidly than the number of histogram bins, and which can be tailored to a given data set. This allows estimates of multi-filter LNP models for non-Gaussian stimulus distributions to be extended to a greater number of filters than would be possible with histogram-based MID (Williamson et al., 2015).

In general, multi-filter LN models in which the form of the nonlinearity f is fixed have been considered much less widely than in the single filter case. In part this is because such fixed-f models are not GLMs (except in the trivial case where the multiple filter outputs are first summed and then transformed, which is no different to a model with a single filter $\mathbf{k} = \sum_n \mathbf{k}_n$). Thus, likelihood-based estimation does not benefit from the structural guarantees conferred by the GLM framework. However, there are a few specific forms of nonlinearity which have been considered. One appears in certain models of stimulus-strength gain control, which are considered next. Furthermore, some Input nonlinearity models, discussed later, combine multiple filters in more complicated arrangements. Finally, low-rank versions of quadratic, generalised-quadratic and higher-order models (see Quadratic and higher-order models) can also be seen as forms of multi-filter LNP model with fixed nonlinearity.

Gain control models

Neurons throughout the nervous system exhibit nonlinear behaviours that are not captured by the cascaded models with linear filtering stage or have a more specialised structure than the general multi-filter models described above. For example, the magnitude of the linear filter in a LN model may change with the amplitude (or contrast) of the stimulus (Rabinowitz et al., 2011), or the response may be modulated

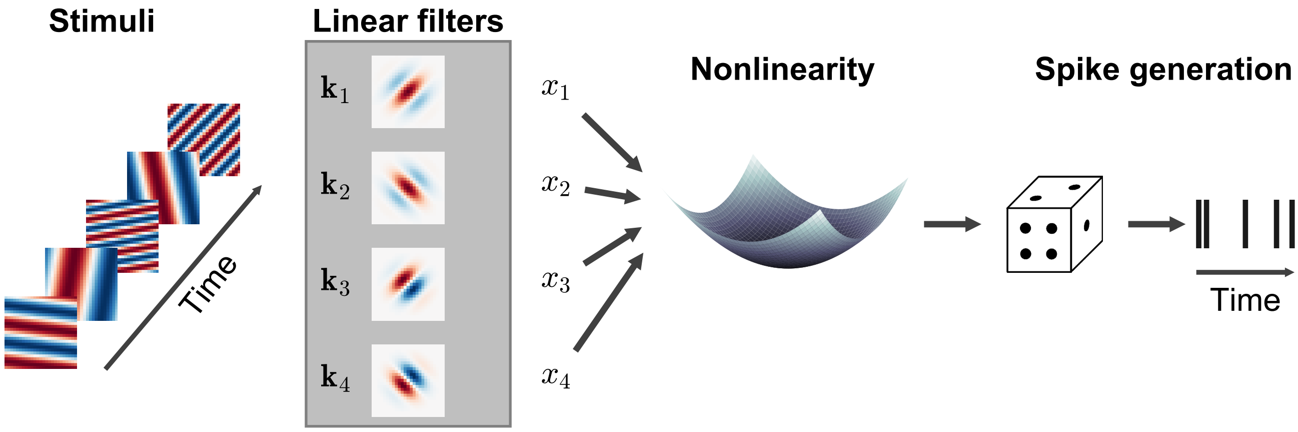

Figure 5. Illustration of a multi-filter linear-nonlinear Poisson encoding model. Each input stimulus (here represented by sinusoidal gratings) is filtered by a number of linear filters $\mathbf{k}_1, \mathbf{k}_2$, ... representing the receptive fields of the neuron. The output of the filters, $x_1, x_2, \ldots$ is transformed by a nonlinearity into an instantaneous spike rate that drives an inhomogeneous Poisson process.

by stimulus components outside the neuron's excitatory RF (.e.g., Chen et al. (2005)). These nonlinear behaviours can be attributed to a mechanism known as gain control, in which the neural response is (usually suppressively) modulated by the magnitude of a feature of the stimulus overall. Gain control is a specific form of normalisation, a generic principle that is assumed to underlie many computations in the sensory system (for a review see Carandini and Heeger 2012).

While there are a number of models specific to particular sensory areas and modalities, most gain control models assume the basic form

$$\hat{r}(t) = f\left(\frac{\mathbf{k}_0^{\mathsf{T}}\mathbf{s}(t) - u\big(\mathbf{s}(t)\big)}{v\big(\mathbf{s}(t)\big)}\right) \tag{11}$$

where $\mathbf{k}_0$ is the excitatory RF of the neuron, and $u(\mathbf{s})$ and $v(\mathbf{s})$ shift and scale the filter output, respectively, depending on the stimulus $\mathbf{s}$. As for the LN model, the adjusted filtered stimulus can be related to the response through a static nonlinear function $f(\cdot)$.

Schwartz et al. (2002) estimated an excitatory RF filter by the STA $\mathbf{k}_0$ and a set of suppressive filters $\{\mathbf{k}_n\}$ by looking for directions in which the STC (built from stimuli orthogonalised with respect to $\mathbf{k}_0$) was smaller than the overall stimulus covariance. They then fit a nonlinearity of the form

$$r(t) = \frac{\left[\mathbf{k}_0^{\mathsf{T}}\mathbf{s}(t)\right]_+^p}{\left(\sum_n w_n|\mathbf{k}_n^{\mathsf{T}}\mathbf{s}(t)|^2\right)^{p/2} + \sigma^2} \tag{12}$$

finding MLEs for the exponent p, which determines the shape of the contrast-response function, the constant σ, and the weights w_n are the coefficients with which each of the suppressive filters $\mathbf{k}_n$ affect the gain.

While in the above example the excitatory and the suppressive filters acted simultaneously on the stimulus, the gain can also depend on the recent stimulation history. Recent studies demonstrated that a gain control model as in Eq. (11) can also account for a rescaling of response gain of auditory cortical neurons depending on the recent stimulus contrast (Rabinowitz et al., 2011, 2012). Specifically, contrast-dependent changes in neural gain could be described by the model

$$r(t) = r_0 + \frac{c}{1 + \exp\left(-\left(\frac{\mathbf{k}_0^{\mathsf{T}}\mathbf{s}(t) - u(\mathbf{s}(t))}{v(\mathbf{s}(t))}\right)\right)} \tag{13}$$

where r_0 is the spontaneous rate, c a constant, $\mathbf{k}_0$ is the STRF, and u and v are linear functions of a single "contrast kernel" that characterises sensitivity to the recent stimulus contrast. In this specific case, the nonlinear function f is taken to be the logistic function.

Input nonlinearity models

LN models assume that any nonlinearity in the neural response can be captured after the output of an initial linear filtering stage. In fact, nonlinear processes are found throughout the sensory pathway, from logarithmic signal compression at the point of sensory transduction, through spiking and circuit-level nonlinearities at intermediate stages, to synaptic and dendritic nonlinearities at the immediate inputs to the cells being studied. *Input* nonlinearities such as these are not captured by a LN model and even the incorporation of a simple static nonlinearity prior to integration (an NL cascade model) can increase the performance of a linear or LN model considerably (Gill et al., 2006; Ahrens et al., 2008a; Willmore et al., 2016).

In the simplest case, the same nonlinear function $g()$ may be assumed to apply pointwise to each dimension of $\mathbf{s}$. For an input nonlinearity model with a single integration filter, we write: $\hat{r}(t) = \mathbf{k}^{\mathsf{T}} g\big(\mathbf{s}(t)\big)$. For $g()$ to be estimated, rather than assumed, it must be parametrised — but many parametric choices lead to difficult nonlinear optimisations. Ahrens et al. (2008b) suggest a tractable form, by parametrising $g()$ as a linear combination of B fixed basis functions g_i, so that $g() = \sum_{i=1}^{B} b_i g_i()$. This choice leads to the *multilinear* model

$$\hat{r}(t) = \mathbf{k}^{\mathsf{T}} \sum_{i=1}^{B} b_i g_i\big(\mathbf{s}(t)\big)\ , \tag{14}$$

which is linear in each of the parameter vectors $\mathbf{k}$ and $\mathbf{b} = [b_1, b_2, \ldots, b_B]$ separately. Least-squares estimates of the parameters can be obtained by alternation: $\mathbf{b}$ is fixed at an arbitrary initial choice, and a corresponding value for $\mathbf{k}$ found by ordinary least squares; $\mathbf{k}$ is then fixed at this value and $\mathbf{b}$ updated to the corresponding least-squares value; and these alternating updates are continued to convergence. The resulting least-squares estimates at convergence correspond to the MLE for a model assuming constant variance Gaussian noise; however a similar alternating strategy can also be used to find the MLE for a generalised multilinear model with a fixed nonlinearity and Poisson or other point-process stochasticity (Ahrens et al., 2008b). Bayesian regularisation (see Regularisation) can be incorporated into the estimation process by an approximate method knows as variational Bayes (Sahani et al., 2013).

The multilinear or generalised multilinear formulation may be extended to a broader range of input nonlinearity models. Ahrens et al. (2008a) discuss variants in which different nonlinearities apply at different time-lags or to different input frequency bands in an auditory setting. Indeed, in principle a different combination of basis functions could apply to each dimension of the input (Figure 6), although the number of parameters required in such a model makes it practical only for relatively small stimulus dimensionalities.

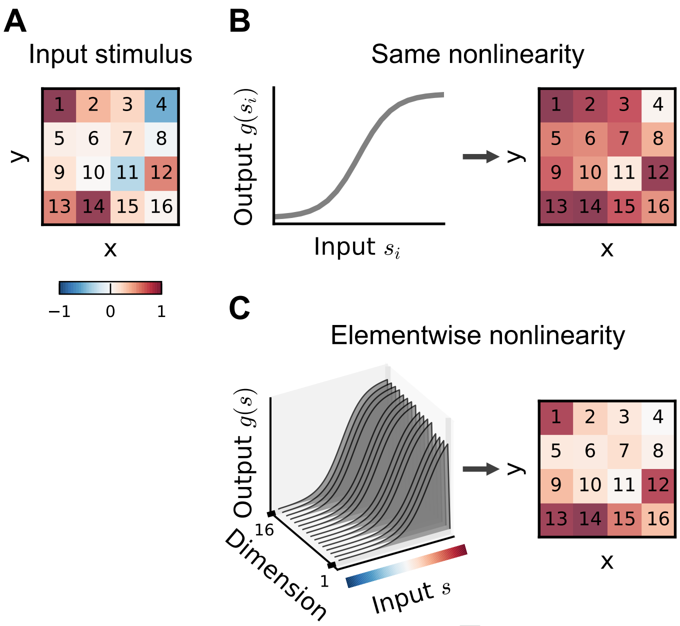

Figure 6. Illustration of input nonlinearity models. **A** Example image patch stimulus. Numbers indicate dimension indices. **B** Input nonlinearity model in which the same nonlinearity (left) acts on all stimulus dimensions, resulting in a transformed stimulus (right). **C** Example where the nonlinearity depends on the y dimensions of the stimulus. Colourbar indicates stimulus values in **A**.

Ahrens et al. (2008a) and Williamson et al. (2016) also introduce multilinear models to capture input nonlinearities in which the sensitivity to each input within the RF is modulated by the local context, for example through multiplicative suppression of repeated inputs (Brosch and Schreiner, 1997; Sutter et al., 1999). The general form of these models is

$$\hat{r}(t) = \sum_i k_i g(s_i(t)) \cdot \text{Context}_i(t) \tag{15}$$

where the term $\text{Context}_i(t)$ itself depends on a second local integration field surrounding the i^{th} stimulus element (called the contextual gain field or CGF by Williamson et al. 2016). The model as described by Williamson et al. is illustrated in Figure 7 for an acoustic stimulus. A local window around each input element of the stimulus is weighted by the CGF and integrated to yield a potentially different value of $\text{Context}_i(t)$ at each element. This value multiplicatively modulates the gain of the response to the element, and the gain-modulated input values are then integrated using weights given by the principal receptive field or PRF. As long as the parameters within $\text{Context}_i(t)$ appear linearly, the overall model remains multilinear, and can also be estimated by alternating least squares.

Nonlinearities prior to RF integration could also result from more elaborate physiological mechanisms. A simple case might be where an early stage of processing is well described by an LN cascade, and the output from this stage is then integrated at the later stage begin modelled. A natural model might then be an LNLN cascade:

$$\hat{r}(t) = f\left(\sum_{n=1}^{N} w_n g_n\left(\mathbf{k}_n^{\mathsf{T}} \mathbf{s}(t)\right)\right) \tag{16}$$

where $\mathbf{k}_n$ describes the linear filter and g_n the output nonlinearity of one of the N input neurons, and their outputs are combined using weights w_n before a final nonlinear transformation f. Such a model has

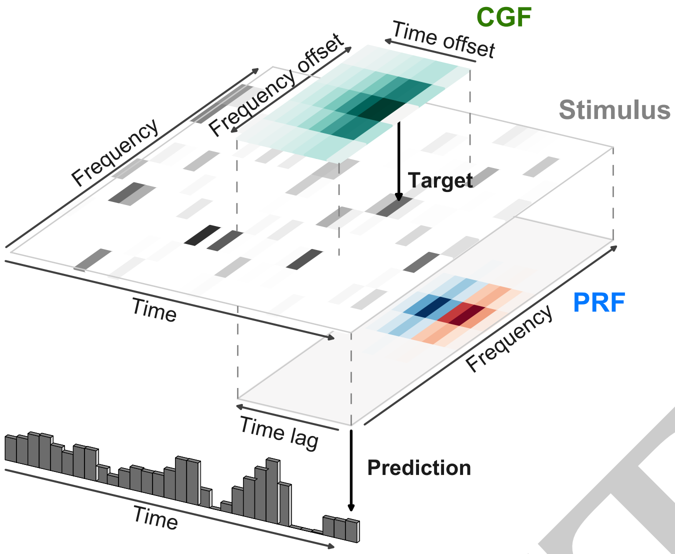

Figure 7. Modelling of local contextual modulation of the stimulus. Each value of the input stimulus (here: target tone of an acoustic stimulus) is modulated according to its context using a contextual gain field (CGF). The modulated stimulus is then transformed into a neural response using a principal receptive field (PRF). While each of these stages is linear, the resulting model is nonlinear in the stimulus.

also been called a generalised nonlinear model (GNM) (Butts et al., 2007, 2011; Schinkel-Bielefeld et al., 2012), or nonlinear input model (NIM) (McFarland et al., 2013) and model parameters may be estimated by maximising the spike-train likelihood of an inhomogeneous Poisson model with rate given by Eq. (16) — often using a process of alternation similar to that described above.

Quadratic and higher-order models

The cascade nonlinearity models described to this point have been designed to balance biological fidelity and computational tractability in different ways. In principle, it is also possible to characterise nonlinear neural response functions using generic function expansions that are not tailored to any particular expected nonlinear structure.

One approach is to use a polynomial extension of the basic linear model:

$$\hat{r}(t) = k^{(0)} + \sum_{i=1}^{D} k_i^{(1)} s_i(t) + \sum_{i,j=1}^{D} k_{ij}^{(2)} s_i(t) s_j(t) + \sum_{i,j,l=1}^{D} k_{ijl}^{(3)} s_i(t) s_j(t) s_l(t) + \dots, \tag{17}$$

where we have re-introduced the explicit constant offset term. Recall that the stimulus vector $\mathbf{s}(t)$ typically includes values drawn from a window in time preceding t. This means that the sums range in part over a time index, and so implement (possibly multidimensional) convolutions. Such a convolutional series expansion of a mapping from one time series (the stimulus) to another (the response) is known as a Volterra expansion (Marmarelis and Marmarelis, 1978) and the parameters $k^{(n)}$ as the Volterra kernels.

While the mapping is clearly nonlinear in the stimulus, Eq. (17) is nonetheless linear in the kernel parameters $k^{(n)}$. Thus, in principle, the MLE of the Volterra expansion truncated at a fixed order p could be found by Eq. (3), with the parameters concatenated into a single vector: $\check{\mathbf{k}} = [k^{(0)}, k_1^{(1)}, k_2^{(1)}, \dots, k_D^{(1)}, k_{11}^{(2)}, k_{12}^{(2)}, \dots, k_{DD}^{(2)}, \dots, k_{DD\dots D}^{(p)}]$; and the stimulus vector augmented to incorporate higher-order combinations: $\check{\mathbf{s}} = [1, s_1, s_2, \ldots, s_D, s_1^2, s_1 s_2, \ldots, s_D^2, \ldots, s_D^p]$. In practice, this approach raises a number of challenges.

Even if the stimuli used in the experiment are distributed spherically or independently, the ensemble of augmented stimulus vectors $\check{\mathbf{s}}(t)$ will have substantial and structured correlation as the higher order elements depend on the low-order ones. One consequence of this correlation is that the optimal value of any given Volterra kernel depends on the order at which the expansion is truncated; for instance, the linear kernel within the best second-order model will generally differ from the optimal linear fit. If the stimulus distribution is known, then it may be possible to redefine the stimulus terms in Eq. (17) (and the entries of $\check{\mathbf{s}}$) so that each successive order of stimulus entries is made orthogonal to all lower-order values. This re-written expansion is known as a Wiener series, and the corresponding coefficients are the Wiener kernels. The Wiener expansion is best known in the case of Gaussian-distributed stimuli (Rieke et al., 1997), but can also be defined for alternative stimulus classes (Pienkowski and Eggermont, 2010). The orthogonalised kernels can then be estimated in sequence: first the linear, then the quadratic and so on, with the process terminated at the desired maximal order.

However, even if orthogonalised with respect to lower-order stimulus representations, the individual elements of the augmented stimulus at any non-linear order will still be correlated amongst themselves, and so STA (or STC) based analyses will be biased. Thus, estimation depends on explicit least-squares or other maximum-likelihood approaches. This raises a further difficulty, in that computation of the inverse auto-correlation $\left(\mathbf{S}^\mathsf{T}\mathbf{S}\right)^{-1}$ in Eq. (3) may be computationally burdensome and numerically unstable. Park et al. (2013) suggest replacing this term, which depends on the particular stimuli used in the experiment, by its expectation under the distribution used to generate stimuli; which, for some common distributions, may be found analytically. They call this approach maximum expected likeihood (MEL). In a sense, MEL provides an extension of the expected orthogonalisation of the Wiener series to structure within a single order of expansion.

The underlying parametric linearity of the Volterra expansion also makes it easy to "generalise" by introducing a fixed, cascaded, output nonlinearity. Although theoretically redundant with the fully general nonlinear expansion already embodied in the Volterra series, this approach provides a simple way to introduce more general nonlinearities when truncating the Volterra expansion at low order. In particular, collecting the second-order Volterra kernel in a matrix $\mathbf{K}^{(2)} = [k_{ij}^{(2)}]$ we can write a generalised quadratic model (GQM):

$$\hat{r}(t) = f\left({\mathbf{k}^{(1)}}^\mathsf{T}\mathbf{s}(t) + \mathbf{s}(t)^\mathsf{T}\mathbf{K}^{(2)}\mathbf{s}(t)\right). \tag{18}$$

Again, as the parameters appear linearly in the exponent, this is a GLM in the (second-order) augmented stimulus $\check{\mathbf{s}}$, guaranteeing convexity for appropriate choices of $f()$ and noise distribution, and rendering the MLE relatively straightforward—although concerns regarding numerical stability remain. Park et al. (2013) show that MEL can be extended to the GQM for particular combinations of stimulus distribution and nonlinear function f. The GQM, with logistic nonlinearity and Bernoulli noise, is also equivalent to an information-theoretic approach that seeks to maximise the "noise entropy" of a second-order model of binary spiking (Fitzgerald et al., 2011b).

An obvious further challenge to estimation of truncated Volterra models is the volume of data needed to estimate a number of parameters that grows exponentially in the order p. Indeed, this has limited most practical exploration of such expansions to second (i.e., quadratic) order, and often required treatment of stimuli of restricted dimensions (e.g., spectral or temporal, rather than spectro-temporal acoustic patterns, Yu and Young 2000). One strategy to alleviate this challenge is to redefine the optimisation in terms of polynomial "kernel" inner products (a different use of "kernel" from the Volterra parameters) evaluated with respect to each input data point (Sahani, 2000; Franz and Schölkopf, 2006). This approach, often called the "kernel trick", makes it possible to estimate that part of the higher-order expansion which is determined by the data (a result called the "representer theorem"), and gives access to a powerful theory of optimisation and regularisation. A second strategy is to parametrise the higher-order kernels so that they depend on a smaller number of parameters. Many such parametrisations lead to versions of cascade model. Indeed the context-modulated input gain model of Williamson et al. (2016) can be seen as a specific parametrisation of the second-order kernel $\mathbf{K}^{(2)}$. Alternatively, "low-rank" parametrisations of kernels as sums of outer- or tensor-products of vectors lead to versions of LN cascade with polynomial or generalised polynomial nonlinearities. Park et al. (2013) suggest that low-rank quadratic models may be estimated by first estimating the full matrix $\mathbf{K}^{(2)}$ using MEL, and then selecting the eigenvectors of this matrix corresponding to the largest magnitude eigenvalues. Although consistent, in the sense that the procedure will converge to the generating parameters in artificial data drawn from a low-rank quadratic model, these significant eigenvectors do not generally give to the optimal low-rank approximation to real data generated according to some other unknown response function. Instead estimates must be found by direct numerical optimisation of the likelihood or expected likelihood. For models of even rank, this optimisation may exploit an alternating process similar to that used for multilinear NL formulations (See Williamson et al., 2016, ,Supplementary methods).

Time-varying models

The models described so far seek to characterise neural mechanisms through a combination of linear and nonlinear transformations. These stimulus-response relationships are assumed to be an invariant (stationary) property of the neuron, i.e., the linear filters and nonlinearities do not change with time. While this assumption might be reasonable for early sensory areas, neurons at higher stages of sensory processing may have more labile, adaptive and plastic response properties, which fluctuate with changes in stimulus statistics, attentional state, and task demands (e.g., Atiani et al. (2009); Rabinowitz et al. (2011); Fritz et al. (2003); David et al. (2012)

The simplest approach to investigating changes in SRF parameters over time is to split the data into different parts, either sequentially using a moving window (Sharpee et al., 2006) or by experimental condition (Fritz et al., 2003). A separate SRF is then estimated for each part, under the assumption that the underlying system is stationary in each part. This "naive" approach can only reveal changes in response properties on rather long time scales, because some minimum amount of data (typically on the order of a few minutes) is required in order to reliably fit model parameters. Temporal resolution can be increased substantially (to about 5 – 20 seconds) by exploiting the fact that fluctuations in SRF parameters are typically rather small, and adopting the strategy of characterising deviations from the long-term SRF estimate (estimated using all data) instead of a "naive" estimate for each part of the data (Meyer et al., 2014b). Previous work using such an approach has shown that accounting for temporal fluctuations in auditory cortical responses can provide a considerably better description of measured responses (Meyer et al., 2014b).

Tracking the evolution of SRFs on a millisecond timescale requires a model that is capable of describing these changes on a moment-by-moment basis. A number of approaches developed to address this issue, ranging from recursive least-squares filtering (Stanley, 2002) to adaptive point-process estimation techniques (Brown et al., 2001; Eden et al., 2004), can be described in the state-space model framework.

State-space methods are a very broad class of methods for analysing neural responses in a dynamic fashion in time and space (or frequency) . Basically, it is assumed that the temporal variation in model parameters is generated through Markovian dynamics; at every time-step, the parameters of the model are determined only by their previous values (and a transition probability). Famous examples are Hidden Markov Models and the Kalman filter with discrete and continuous states, respectively. While the details are beyond the scope of this overview, it should be noted that many models described here (specifically probabilistic models like GLMs) can be formulated as state-space models, and that principled ways of estimating model parameters exist for such models (Paninski et al., 2010).

Ongoing fluctuations in network dynamics, for instance related to synchronisation and desynchronisation in sensory cortex, may also contribute to SRF variability; response properties are seen to change qualitatively with cortical state (Pachitariu et al., 2015). Recent work has demonstrated that these fluctuations are reflected in the local field potential (Saleem et al., 2010), and that including local-field-potential phase information in a simple GLM with a fixed filter provides a better description of neural responses in the anaesthetised auditory cortex (Kayser et al., 2015). While the parameters of the model are time-invariant, the output of the model depends on network dynamics. Such an approach makes it possible to disentangle intrinsic properties of the neuron from (potentially global) network effects.

Regularisation

Even a linear RF filter is often high-dimensional, possibly containing hundreds or even thousands of elements — particularly when it extends in time as well as over sensory space. To accurately estimate so many parameters requires a large amount of data. In a space of stimuli such as that drawn in Figure 3, the number of dimensions corresponds to the number of RF parameters. To properly estimate the RF direction in this space, whether by STA, MID or MLE, it is necessary that all directions in this very high-dimensional space be explored with a reasonable number of samples, so that the effects of variability in response on the estimate of the component of the RF along that direction are averaged away. However, the difficulty of maintaining stable neural recordings over long times, or other constraints of experimental design, often limit the data available in real experiments. With limited data in very many dimensions, it becomes very likely that random variability along some dimensions will happen to fall in a way that appears to be dependent on stimulus value. Simple STA, MID or MLE estimates cannot distinguish between such random alignment and genuine stimulus-dependence, and so *overfit* to the noise, leading to poor estimates of RF parameters. By construction the overfit model appears to fit data in the training sample as well as possible, but its predictions of responses will fail to generalise to new out-of-sample measurements. The noisy RF estimates might also be biologically implausible, with a “speckled” structure of apparently random sensitivities in time and space (Figure 8).

Regularised estimators incorporate strategies to combat overfitting. Two approaches to regularisation have seen widespread use in SRF estimation: early stopping and the incorporation of penalty terms in cost functions. In early stopping, the parameters are found by an iterative process, most often gradient ascent in an objective function such as the likelihood or single-spike information. Following each iteration, the predictions of the current parameters are tested on a separate held-out data set (see section for more on the effect of noise in training and testing data on model fitting). Once these validation predictions no longer improve the iterations are stopped and the current parameters are taken to be the regularised estimate (Sharpee et al., 2004; David et al., 2007; Fitzgerald et al., 2011b).

The second approach to regularisation augments the objective function with additional terms or *regularisers* that penalise implausible values of the parameters. In the context of estimation theory, the addition of a regulariser introduces bias into estimates but reduces variance, and so frequently reduces the expected squared error of the estimate. Furthermore, if the magnitude of the regulariser is independent of the number of data, while the scale of the original objective function (such as log-likelihood) grows with the data volume, the regulariser has little impact on the optimum for large data sets, and estimators remain consistent. In practical settings, where the responses do not in fact arise from an instance of the model, regularised estimates are almost always found to generalise more accurately to novel data than unregularised ones.

In the likelihood setting, the regulariser may be interpreted as a *prior* belief about the plausibility of parameter values. Then, by Bayes' rule, the regularised objective corresponds (up to a constant) to the *posterior* belief about the parameters given data, and the maximum of this objective is called the *maximum a posteriori* or MAP estimate:

$$\begin{aligned}\hat{\mathbf{k}}_{\mathrm{MAP}} &= \operatorname{argmax} \left[p(\mathbf{r}|\mathbf{S},\mathbf{k})p(\mathbf{k}|\Theta)\right] &(19)\\ &= \operatorname{argmax} \left[\log p(\mathbf{r}|\mathbf{S},\mathbf{k}) + \log p(\mathbf{k}|\Theta)\right] &(20)\end{aligned}$$

where $p(\mathbf{r}|\mathbf{S},\mathbf{k})$ is the probability of the observed response given the stimulus $\mathbf{S}$ under the model parameters $\mathbf{k}$ (and so the likelihood function for $\mathbf{k}$), and $p(\mathbf{k}|\Theta)$ is regularising prior which may depend on *hyperparameters* Θ (note that taking the logarithm in the second line does not change the location of the maximum). The hyperparameters may be adjusted to refine the penalty term based on the data themselves: either by selecting the values that lead to estimators that generalise best when measured by cross-validation on the training data (Theunissen et al., 2000; Machens et al., 2004; Meyer et al., 2014a); or by a process known variously as *evidence optimisation*, *maximum marginal likelihood*, or sometimes *empirical Bayes* (Sahani and Linden, 2003a). For more complex models such as the multilinear approaches used for NL cascades, the corresponding approach relies on an approximation known as *variational Bayes* (Sahani et al., 2013).

For RF models, appropriate choices of regulariser may favour fields which vary smoothly in time and sensory space, which have relatively few non-zero elements and in which these non-zero elements are concentrated within a single local region. Terms incorporating these constraints can be refined by evidence optimisation leading to schemes known respectively as *automatic smoothness determination* (ASD), *automatic relevance determination* (ARD) (both discussed by Sahani and Linden, 2003a), and *automatic locality determination* (ALD) (Park and Pillow, 2011); each of which leads to improved model estimates, particularly for small sample sizes. Results are illustrated in Figure 8 for a linear-Gaussian model with a zero-mean Gaussian prior (known as "ridge regression"), ASD, and ALD.

Parameter optimisation

Many of the parameter estimators discussed in this review are defined by the optima of likelihood or other objective functions. How easy are these optima to find?

For linear models estimated by least-squares, corresponding to the MLE under the assumption of fixed-variance Gaussian noise, the optimum is available in closed form by Eq. (3). This analytic result can be extended to the MAP estimate under the assumption of a fixed zero-mean Gaussian prior with inverse covariance matrix $\mathbf{A}$ on the RF weights, for which we obtain:

$$\hat{\mathbf{k}}_{\mathrm{MAP}} = (\mathbf{S}^{\mathsf{T}}\mathbf{S} + \lambda\mathbf{A})^{-1}\mathbf{S}^{\mathsf{T}}\mathbf{r}\,. \tag{21}$$

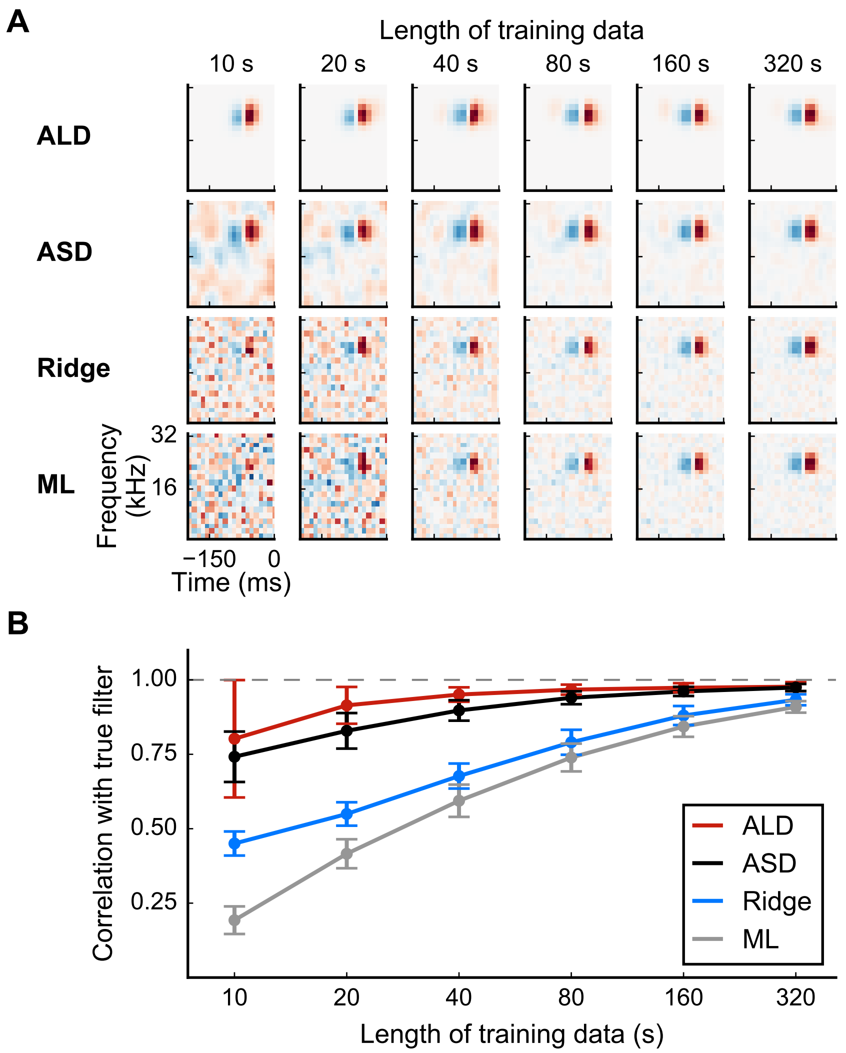

Figure 8. Simulated example illustrating the effect of priors on linear filter estimation. Responses were simulated using a linear-Gaussian model with different spectro-temporal receptive fields (STRFs) and a noise-like input stimulus. (**A**) STRF estimates for different data sizes obtained using a linear model with different priors. Maximum-likelihood (ML) estimation and Ridge regression appear noisy for small sample sizes. Estimators using more structured priors like automatic smoothness determination (ASD) and automatic locality determination (ALD) yield better estimates in the small sample regime. (**B**) Mean correlation of estimated STRFs with the true STRF across for different neuron. Error bars indicate one standard deviation.

The regularisation parameter λ is equal to the assumed variance of the Gaussian output noise. In ridge regression, $\mathbf{A}$ is taken to be the identity matrix, and λ either set arbitrarily or chosen by cross-validation. For adaptive regulariser approaches, including ASD, ARD and ALD, λ and the matrix $\mathbf{A}$ must first be found by maximising the model "evidence" by iterative numerical methods.

With the exception of the STA- and STC-based approaches, estimators for non-linear models require iterative optimisation. For a Poisson GLM, possibly with spike-history terms, the log-likelihood function is concave provided that the static nonlinearity assumed is convex and log-concave (Paninski, 2004). This concavity property extends naturally to the log-posterior under a log-concave prior. Such functions have a single unconstrained maximum, which is easily found by gradient-based methods (e.g. Boyd and Vandenberghe, 2004). In particular, a standard algorithm from the GLM literature known as *iteratively* *reweighted least squares* (IRLS; Green 1984) exploits information about the expected local curvature of the likelihood to converge rapidly on the optimum. For specific static nonlinearities known as “canonical” (these include the exponential function for Poisson models, and logistic function for Bernoulli models), IRLS corresponds exactly the Newton method of optimisation. In these cases, and if stimuli are drawn randomly from a known and simple distribution, estimation can be further accelerated by maximising the “expected likelihood” with only a small cost in accuracy (Ramirez and Paninski, 2014). Alternatively, stochastic gradient techniques estimate gradients using random subsets of the data, converging stably for convex optimisation problems. These techniques are simple and scalable, making them particularly well-suited to large data sets, and they also facilitate online monitoring of SRF parameters during experiments through their batch-based structure (Meyer et al., 2015).

For estimators based on non-convex objective functions, such as MID, general LN likelihood models, or multilinear NL models, as well as the evidence-optimisation stage of some adaptive regularisers, the results of iterative optimisation may depend on the parameter value from which the iterations begin. Thus, additional steps are needed to ensure that the local optimum found is likely to represent a good parameter or hyperparameter choice. One approach is to repeat the iterative optimisation starting from a number of different initial parameter values, accepting the results of the run that leads to the best value of the objective function (or, as a form of regularisation, the best validation performance; compare the discussion of early stopping above). Alternatively, stochastic gradient methods, particularly incorporating momentum terms, may escape poor local extrema and approach the true optimal parameter values. A similar idea, albeit without explicit use of gradient information, underlies stochastic search methods such as simulated annealing. In the general case, however, no approach beyond exhaustive search can guarantee that the value obtained will be the true global optimum of the objective function.

PART 2: EVALUATION

Once we have a found the parameters of a model for a set of neural data, there remains the important task of validating the quality of the model fit. In this section, we discuss different methods for quantifying how well a fitted model captures the neural response.

There are different settings in which model performance needs to be evaluated. The relatively straightforward scenario is when we wish to compare the performance of two or more estimators for a specific model, e.g., different regression-based estimators of the linear-Gaussian model (see Eq. 2). In this case, the log-likelihood provides a convenient measure for comparing the *relative* performance of the estimators on the same set of validation data. However, often we are interested in finding which model amongst a number of different models provides the best description of the neural response. Again, this is a relative comparison, but in this case of the models rather than the estimators; therefore a *model-independent* measure is required, such as the single-spike information (Brenner et al., 2000; Sharpee et al., 2004).

Ultimately, however, the goal is not only to identify the best model for a recorded set of data but also to quantify the fraction of the response captured by the model. This scenario — evaluation of *absolute* model performance — is more complicated, because response prediction errors arise not only from inaccurate model assumptions but also from variability in neural responses. While these variations might represent an important aspect of the neural response, from a modelling perspective they are usually treated as “noise” (unless the variations are under control of the experimenter or are related to observable variables), and the impact of this “noise” has to be taken into account when evaluating absolute model performance.

In the following, we will provide an overview of common measures used to evaluate performance of the different stimulus-response function models reviewed above. We will also provide an intuitive outline of a methodology that allows the separation of response prediction errors arising from inaccurate model assumptions from errors arising from noise inherent in neuronal spike trains (Sahani and Linden, 2003b).

Rate-based approaches

Mean squared error (MSE)

For continuous responses such as spike rates or local field potentials, a natural measure for the quality of an estimated model is the mean squared error (MSE; σ_e^2) between the estimated response $\hat{r}$ and the measured response r,

$$\sigma_e^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{r}(t) - r(t)\right)^2 \tag{22}$$

$$= \left\langle \left(\hat{r}(t) - r(t)\right)^2 \right\rangle, \tag{23}$$

with $\langle\cdot\rangle$ used to denote average over time. The MSE is a common measure of error used in many estimation problems, and is also closely related to the negative log-likelihood of the linear-Gaussian model.

The MSE measures the mean error per sample (Figure 9 **A**) but it is not bounded above; higher variability in the recordings will produce higher MSE estimates for equivalent data sizes. This limitation makes it difficult to compare MSE values across different brain areas or even across different recordings from the same area. The coefficient of determination, or R^2 statistic, normalises the MSE by the variance in the neural response,

$$R^2 = \frac{\sigma_r^2 - \sigma_e^2}{\sigma_r^2} \tag{24}$$

where $\sigma_r^2 = \left\langle (r(t) - \langle r\rangle)^2 \right\rangle$ is the variance in the neural response with mean $\langle r\rangle = 1/T\sum_{t=0}^{T} r(t)$ and σ_e^2 is the MSE. Unfortunately, R^2 cannot be used directly to quantify how well a model reproduces the recorded response as it does not distinguish stimulus-dependent variance in the response from stimulus-independent variability ("noise"). A modification of Eq. (24) described below (see Quantifying stimulus-dependent coding in the presence of noise) makes it possible to measure the fraction of *explainable* variance in the data captured by a specific model in the presence of such stimulus-independent variability.

Correlation and coherence

Correlation measures the degree of linear dependence between two variables. For a predicted and observed time-varying firing rates, the sample correlation coefficient, also known as the Pearson correlation, is calculated as

$$\rho_{r,\hat{r}} = \frac{\mathrm{cov}(r,\hat{r})}{\sigma_r \sigma_{\hat{r}}} \tag{25}$$

where the cross-covariance and the standard deviations are replaced by their sample estimates: $\mathrm{cov}(r,\hat{r}) = 1/N\sum_{t=1}^{T}(r(t) - \langle r\rangle)(\hat{r}(t) - \langle\hat{r}\rangle)$ and $\sigma_r = \sqrt{1/N\sum_{t=1}^{T}(r(t) - \langle r\rangle)^2}$, and similarly for $\sigma_{\hat{r}}$. The correlation coefficient is bounded between -1 and 1, and with a correlation of 1 indicating a perfect linear relation between predicted and actual response, and values close to zero indicating that the responses are linearly unrelated. An example for a cortical neuron is shown in Figure 9 **B**.

The correlation coefficient is centred and normalised and therefore does not depend on mean or scaling of the signals. In settings where the focus is on capturing the temporal modulation of the firing rate rather than its overall magnitude, this may provide an advantage over the MSE. Introducing a time lag between the two signals, and computing a correlation at each lag yields a function known as the crosscorrelogram. This may reveal temporal relationships between the two prediction and measurement, e.g., temporal offsets of correlation lengths, that are not evident from the correlation at zero time lag.

An alternative formulation of the linear dependency between two signals is the magnitude-squared coherence,

$$\gamma^2(\omega) = \frac{|S_{r\hat{r}}(\omega)|^2}{S_r(\omega)S_{\hat{r}}(\omega)}, \tag{26}$$

where $S_{r\hat{r}}(\omega)$ is the cross-spectrum of r and $\hat{r}$, and $S_r(\omega)$ and $S_{\hat{r}}(\omega)$ are the autospectra of r and $\hat{r}$, respectively (Gardner, 1992). The coherence measures of the strength of linear relationship between two processes as a function of frequency. While it can be more expensive to compute, it has several important advantages over the time-domain correlation. First, for spike data, the correlation coefficient and correlogram require binned spike counts and their values depend on the bin size. As Fourier transforms of spike-train signals can be found without explicit discretisation or smoothing, computation of the coherence does not require binning; and is less sensitive to the bin size if the data have been pre-binned. The temporal scale of the correlation is instead implicit in the frequency range over which the coherence is considered. Thus, the coherence may be diagnostically valuable: revealing, for instance, that a model accurately predicts slow fluctuations in response while missing many short time-scale events. (Figure 9 **C**). For nonstationary signals, such as stimulus-driven firing rates, the coherence has to be estimated from continuous time-varying quantities. Common approaches for obtaining continuous firing rate estimates include moving-window averaging, wavelet-based filtering, and multitaper techniques (Brown et al., 2004).

Spike-based approaches

Single-spike information

The single-spike information (Eq. 7) maximised by the MID estimator of the LNP model provides a measure of the mutual information between stimulus and response, regardless of the shape of the neural nonlinearity. Furthermore, it does not depend on the scaling of the linear filter(s) which might be inherently different for different estimators. Therefore, it is a useful measure to use for comparing different LNP models.

However, empirical estimation of information-theoretic quantities from finite data is non-trivial. Histogram-based estimation of single-spike information values can result in substantial upward bias in information estimates (Brenner et al., 2000; Paninski, 2003b). While it is possible to correct for this bias to some degree (Paninski, 2003b), the optimal number of histogram bins also depends on the amount of data (Paninski, 2003b; Williamson et al., 2015). Thus, the parametrisation of the histogram-based estimator must be chosen carefully, or investigated as a variable.

Once an appropriate parametrisation has been identified, the single-spike information can be normalised by the total information in the response (Brenner et al., 2000). The total information can be estimated from a large number (e.g. 50–150) of repetitions of a short stimulus segment (Sharpee et al., 2008), using

$$I_{\text{resp}} = \frac{1}{T}\int \mathrm{d}t \frac{r(t)}{\langle r\rangle} \log_2 \frac{r(t)}{\langle r\rangle}, \tag{27}$$

where $r(t)$ is the time-varying firing rate for the stimulus segment averaged over all stimulus repetitions, and $\langle r \rangle$ is the overall mean firing rate across time and repetitions. Finite data effects both in the single-spike information and the total information in the response can be reduced by (linear) extrapolation to infinite data (Sharpee et al., 2008).

Receiver-operating characteristic analysis

The problem of correctly predicting a spike can also be phrased in terms of a detection task with the goal of successfully detecting events (spikes) against a background (no spikes). In signal detection theory, the successful detection of a spike can be described by the receiver operating characteristic (ROC) curve, which is generated by plotting the fraction of correctly detected spike examples ("true positive rate") versus the fraction of falsely detected non-spike examples ("false positive rate") for different spiking thresholds (Green and Swets, 1966; Meyer et al., 2014a). Because the output of most binary SRF models depends only on the filtered stimulus, this is equivalent to "shifting" the threshold along the axis defined by the filter and estimating the rates from the conditional distributions.

This is illustrated in Figure 9 for an example auditory cortical neuron. The overlap between the distributions can be quantified by integrating over all thresholds (e.g., using the trapezoid rule) yielding the area under the ROC curve (AUC). A value close to 1 corresponds to a small overlap, whereas a value close to 0.5 indicates highly overlapping distributions indicating that the fitted model cannot distinguish between spikes and silence.

From a model perspective, the overlap determines the amount of noise in the system. The discriminative approach described above (see Bernoulli models) seeks to find the filter in stimulus space that minimises the overlap between the distributions, which is equivalent to finding the model with minimum coding noise (Meyer et al., 2014a). Note that the single-spike information seeks to minimise the overlap between similar conditional distributions in terms of the Kullback-Leibler divergence (see Figure 3 **D**). In case the number of spikes is small relative to the number of bins, which is typically the case for small enough bin sizes, AUC and single-spike information are highly correlated, with AUC exhibiting considerably smaller bias and variance for small sample sizes (Meyer et al., 2013).

Quantifying stimulus-dependent coding in the presence of noise

The response of a neuron to repeated presentations of the same physical stimulus can vary considerably, even in anaesthetised preparations (Tolhurst et al., 1983; Goris et al., 2014). This variability makes it difficult both to estimate the parameters of the model in the first place, and then to quantify the extent to which a given model or class of models has captured the true response of the neuron. Here, we describe a three-step procedure for finding the fraction of the explainable component in the response that can be captured by a model, for a population of similar neurons (e.g., from a specific brain area). We also illustrate the principles on simulated data.

Estimation of signal and noise power

Suppose that we have available the responses of a population of neurons to N repetitions of the same stimulus. (It is not essential that all neurons were recorded at once as the analysis is performed treating each neuron as a separate sample.) Our objective is to measure the performance of a predictive model in terms of the fraction of the neuron's response that it successfully predicts. Following Sahani and Linden (2003b) we focus on the *response power* or response variance σ_r^2 (see Eq. 24).

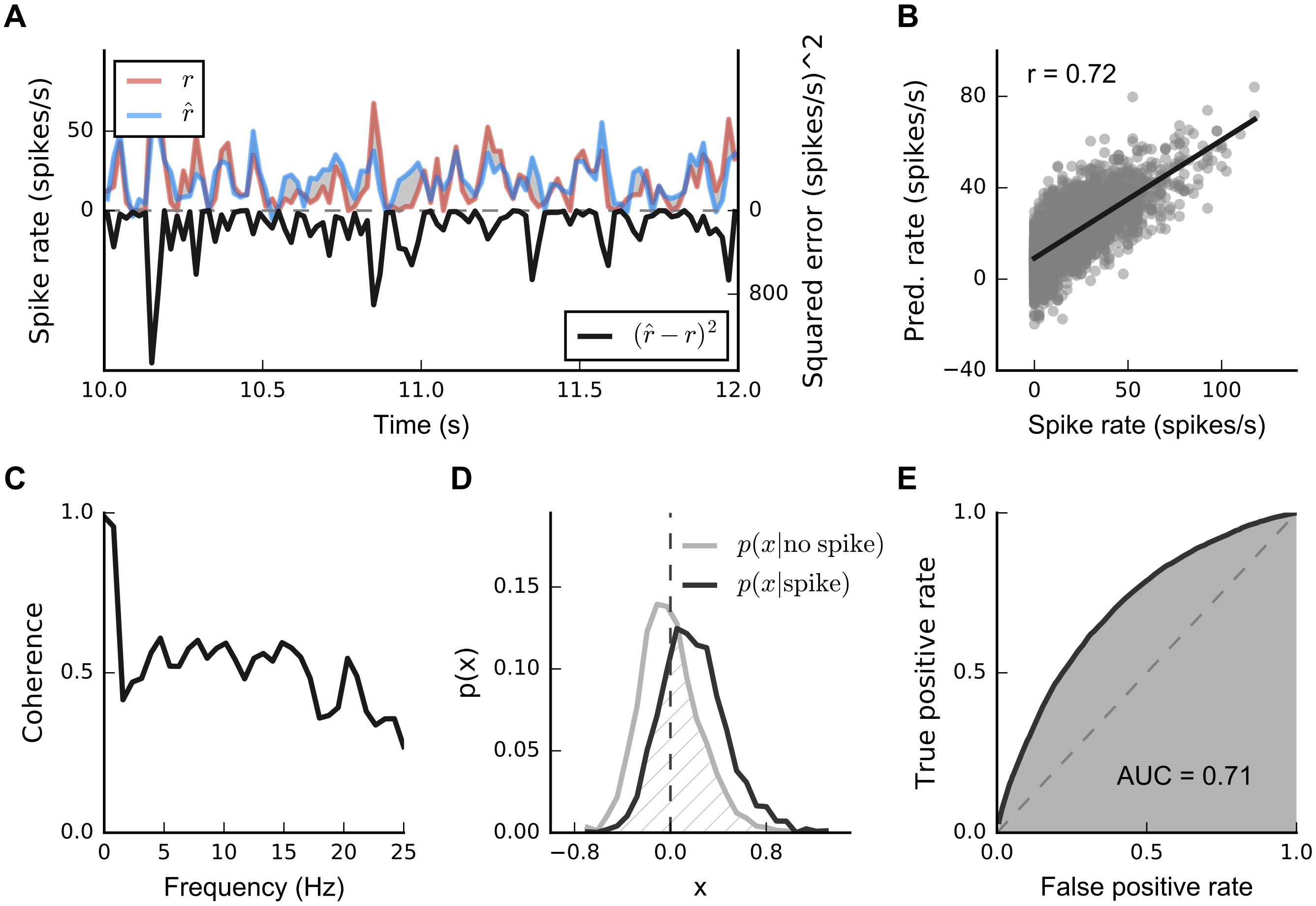

Figure 9. Common techniques for evaluating SRFs. (**A**) The mean squared error (MSE) measures the squared error (black line) per time step between the measured rate (red line) and predicted rate (blue line). The error between the rates (grey shaded area) depends on mean and scaling of the rates. (**B**) The correlation coefficient reflects the linearity between measured and predicted response, indicated by the least squares line fit. The correlation is invariant to linear transformations, i.e., its value does not depend on the mean and the scaling of the responses. (**C**) The coherence assesses the linear relation between two variables in frequency space; i.e., it is a frequency-dependent correlation measure. (**D**) Conditional distribution of filtered stimuli that elicited a spike (black line) or no spike (grey line). The hatched area indicates the overlap between the two distributions, which is related to prediction performance in a binary coding model (see text for details). (**E**) A receiver-operating characteristic curve (ROC) can be constructed from the distributions in **D** by computing false positive and true positive rates for all possible thresholds along the x-axis. The area under the ROC curve (AUC; shaded grey area) provides a scalar measure of the prediction performance of the fitted model

From a modelling perspective, the response to the nth stimulus repetition, $r^{(n)}(t)$ can be divided into a reliable (stimulus-driven *signal*) part $\mu(t)$ and a variable (*noise*) component $\eta^{(n)}(t)$,

$$r^{(n)}(t) = \mu(t) + \eta^{(n)}(t)\,. \tag{28}$$

We define $\mu(t)$ to be the expected response to the stimulus — the average we would obtain from an infinite collection of responses to the same stimulus — and so $\eta^{(n)}(t)$ has an expected value of zero for all t and n. The signal $\mu(t)$ reflects the time-locked, stimulus-driven part of the response of the neuron under consideration, and it is thus the component of the response that is (in theory) predictable by a model of the cell's SRF. However, the average of a finite number of trial responses collected within experimental constraints will retain a contribution from the noise, and thus the true signal response cannot be determined. Nevertheless, it is possible to form an unbiased estimator of the *power* or variance in that response $\sigma_\mu^2 = \left\langle (\mu(t) - \langle \mu \rangle)^2 \right\rangle$ as follows.

First, the simple property of additivity of variances implies that on each trial $\sigma_r^2 \overset{\mathcal{E}}{=} \sigma_\mu^2 + \sigma_\eta^2$ where σ_η^2 is the average squared deviation from $\mu(t)$. Note that σ_r^2 and σ_η^2 depend on the particular response on a single trial, while σ_μ^2 is a property of the idealised response. Thus, $\overset{\mathcal{E}}{=}$ means "equal in expectation"; i.e., the equality may not hold on any trial, but the expected values of the left- and right-hand sides are equal. This relationship depends only on the noise component having been defined to have zero expectation, and holds even if the variance or other property of the noise depends on the signal strength as would be expected for a Poisson noise process (see the simulated example in Fig. 10 **A–C**). We now construct two trial-averaged quantities, similar to the sum-of-squares terms used in the analysis of variance (ANOVA) (e.g. Lindgren, 1993): the power of the average response $\sigma_{\bar{r}}^2$, and the average power per response $\overline{\sigma_r^2}$, with $\bar{\cdot}$ indicating trial averages:

$$\sigma_{\bar{r}}^2 \overset{\mathcal{E}}{=} \sigma_\mu^2 + \sigma_{\bar{\eta}}^2 \qquad \text{and} \qquad \overline{\sigma_r^2} \overset{\mathcal{E}}{=} \sigma_\mu^2 + \overline{\sigma_\eta^2}.$$

Assuming that the noise in each trial is independent, although the noise in different time bins within a trial need not be, we have: $\sigma_{\bar{\eta}}^2 \overset{\mathcal{E}}{=} \overline{\sigma_\eta^2}/N$. Then solving these two equations for σ_μ^2 suggests the following estimator for the signal power:

$$\widehat{\sigma_\mu^2} = \frac{1}{N-1}\left(N\sigma_{\bar{r}}^2 - \overline{\sigma_r^2}\right). \tag{29}$$

A similar estimator for the *noise power* is obtained by subtracting this expression from $\overline{\sigma_r^2}$. Thus, the resulting estimator of the fraction of explainable response power captured by a model, the *predictive power*, is given by

$$\beta = \frac{\sigma_r^2 - \sigma_e^2}{\widehat{\sigma_\mu^2}}. \tag{30}$$

This corresponds to the R^2 estimator (Eq. 24) except that the explained variance is measured against an estimate of the stimulus-driven power (or variance) instead of the total response variance, which overestimates the signal power by the noise power (Fig. 10 **C**).

Hsu et al. (2004) applied a similar idea to the coherence measure (see Eq. 26) to obtain an estimate of the coherence between model prediction and signal-driven response. However, it is important to note that whereas the estimator for the signal power itself (Eq. 29) depends depends linearly on the measured power in single responses and their trial average and so is unbiased, estimators for the predictive power (Eq. 30) and coherence (**?**) which involve nonlinear transformation are at best consistent. However simulations (Fig. 10 **D** and Hsu et al. 2004) suggest that any finite-data biases might be small for typical data volumes.

David and Gallant (2005) study the bias in the correlation coefficient between (unregularised) prediction and validation measurements, using an analysis similar to the predictive power. They focus separately on the prediction errors introduced directly by noise in the measured validation data and by mis-estimation of model parameters from noisy training data, and propose two different schemes for extrapolation in number of trials or training time (though not in the population noise level as described below). While they arrive at the correct estimate of the correlation coefficient of the ideal model this approach makes assumptions that might not hold for many experimental data sets. First, the unregularised model is assumed to be predictive which is often not the case for realistic data sizes (see Regularisation). Second, the (linear) model fit is assumed to be the same in the noise-free training and validation sets. This is approximately true for large training and validation data sets but unlikely for rather limited amounts of data as stimuli in the two sets might differ substantially and neural models are stimulus dependent (Christianson et al., 2008).

Upper and lower estimates of model predictive power

Model parameters (such as the weights or coefficients of the SRF) are commonly estimated by minimising the mean squared error of the model prediction on the training data. By definition, these least-mean-squares (LMS) parameters produce model predictions for the training data that have minimum possible error, and therefore maximal predictive power. Of course, the resulting maximal value, the training predictive power, will inevitably include an element of overfitting to the training data, and so will overestimate the true predictive power of the model with ideal parameters (i.e., the model that would perform best on average for all possible stimulus-response combinations, not just the training data). More precisely, the expected value of the training predictive power of the LMS parameters is an upper bound on the predictive power of the model with ideal parameters. Thus, the measured training predictive power can be considered an upper estimate of the true predictive power of the model class (light grey dots in Figs. 10 **D** and **E**).

We can also obtain a lower estimate, defined similarly, by empirically measuring the generalisation performance of the model by cross-validation. Cross-validation provides an unbiased estimate of the average generalisation performance of the fitted models (as obtained from the training fraction of the available data). Since these models are inevitably overfit to their training data, not the test data, the expected value of this cross-validation predictive power bounds the predictive power of the model with ideal parameters from below, and thereby provides the desired lower estimate of the true predictive power of the model class (dark grey dots in Figs. 10 **D** and **E**).

Population extrapolation to zero-noise limit

For any one recording of finite length, the true predictive power of the model class (i.e., the predictive power of the version of the model with ideal parameters) can only be bracketed between the upper and lower estimates defined above. The looseness of these estimates will depend on the variability or noise in the recording. For a recording with high trial-to-trial variability, the model parameters will be more strongly overfit to the noise in the training data. Thus we expect the training predictive power on such a recording to appear high relative to the signal power, and the cross-validation predictive power to appear low. Indeed, in very high-noise conditions, the model may primarily describe the stimulus-independent noisy part of the training data, and so the training predictive power might exceed the estimated signal power (P_{signal}), while the cross-validation predictive power may fall below zero (that is, the predictions made by the model may be worse than a simple unchanging mean rate prediction). Thus, the estimates may not usefully constrain the predictive power measure for a particular recording.

However, if we assume that the true fractional predictive power of the model is similar for the entire population of neurons recorded, then it is possible to tighten the estimates of model predictive power for the population as a whole, by extrapolating across the population to a hypothetical recording with no noise. In other words, assuming the population of recorded neurons is relatively homogeneous, we can plot upper and lower estimates of model predictive power as a function of noise level for all the neurons recorded, and then extrapolate to the point of zero noise level to obtain a relatively tight estimated range within which the optimal population mean predictive performance of the model must lie. This is illustrated for two simulated populations in Figures 10 **D** and **E**.

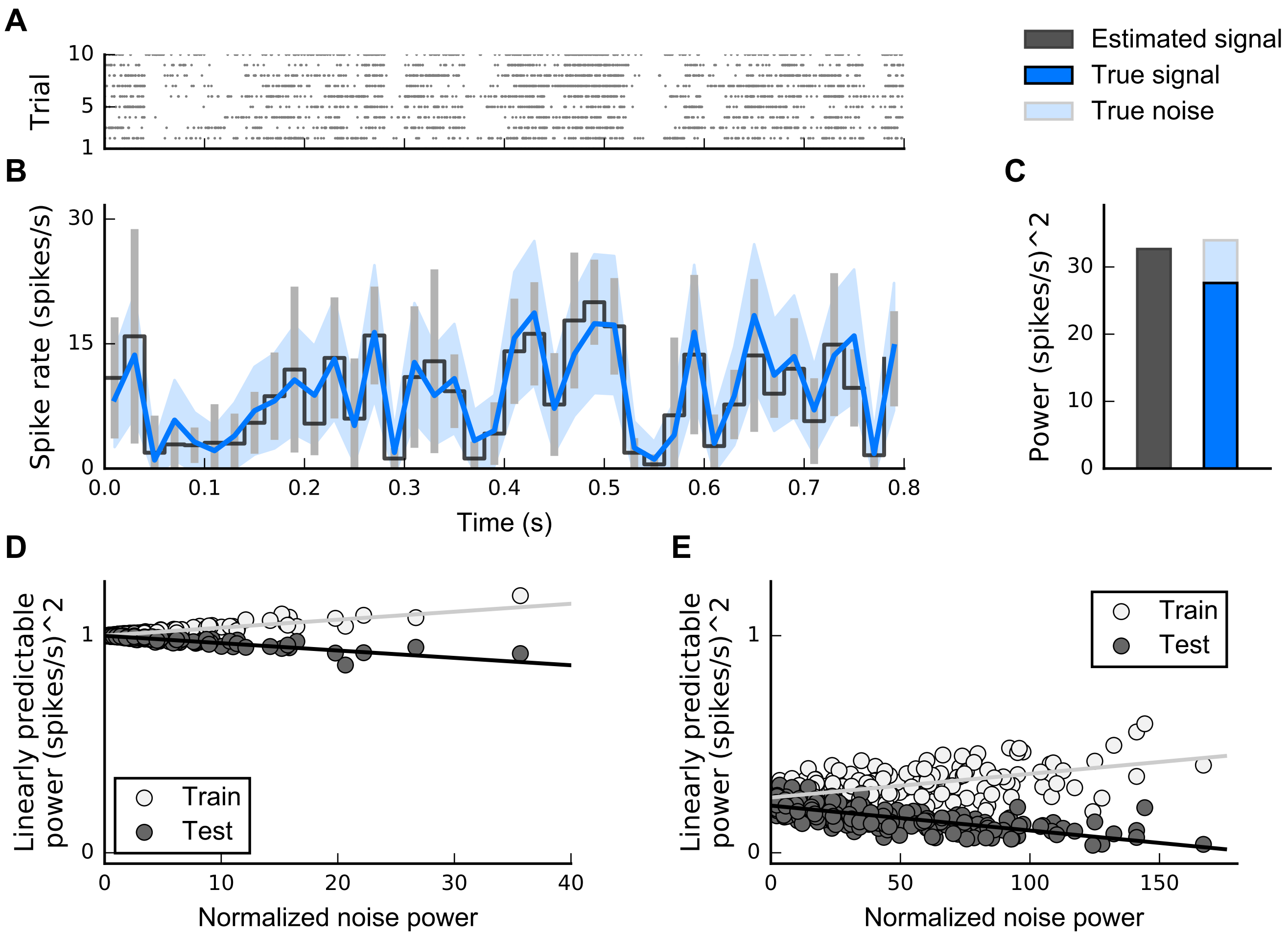

Figure 10. Signal power, noise power, and population extrapolation. Simulated data illustrate principle of quantification of predictable signal power. (**A**) Raster plot showing Poisson spike trains for 10 presentations of the same stimulus. (**B**) True signal (solid blue line) that was used to generate the spike trains, together with true noise (blue shaded area). The spike rate (black line) and the standard deviation across trials (grey bars) were estimated by counting spikes in discrete bins. (**C**) Power of estimated response, true signal, and true noise. The estimated response power overestimates the true signal power by the noise power (additivity of variances, see text). The estimated signal power is found by subtracting the noise power from the estimated response power. (**D**) Normalised predictive power for a population of 200 simulated linear-Gaussian cells. Predictions on training data (light grey circles) and testing data dark grey circle) were done using a linear estimator. As expected, extrapolation to zero noise power reveals that the model accounts for the maximum linearly predictable power. (**E**) The same as in **D** but responses were simulated using 200 linear-nonlinear Bernoulli cells and a non-Gaussian stimulus (similar to the stimulus in Figure 3**B**). Extrapolation to the zero noise condition indicates that imperfect model performance is due to an incorrect model assumption (linear-Gaussian model) rather than to noise.

The upper and lower estimates of model predictive power in this zero-noise limit provide the desired noise-independent measure of model predictive performance. Crucially, low predictive power values in the zero-noise limit indicate that the assumed model does not provide a good description of the underlying responses, as shown in Figure 10 **E**. Moreover, the assumption that the neuronal population is homogeneous (with regard to predictive power of the ideal version of model for each recording) can be directly assessed by examination of the population spread in the extrapolated values. If the assumption is valid, then the population spread should be unimodal and relatively small.

Consequences of model mismatch

DISCUSSION

Abstract stimulus–response function models can be versatile and powerful tools for addressing many different questions about sensory processing and neural representation. The great advantage of these models is that their parameters can be estimated from experimentally feasible amounts of data, but nevertheless can describe neuronal responses across a large subset of a high-dimensional stimulus space. The disadvantage is the flipside of this advantage; the same abstract formulation that permits robust and efficient parameter estimation from limited data also requires assumptions that can produce potentially misleading results arising from mismatch with biological reality.

Unlike biophysical models that describe actual low-level mechanisms of sensory processing such as synaptic transmission and channel dynamics, functional models are abstract descriptors of the stimulus–response function transformation. In general, then, the estimated parameters of functional models should not be interpreted as estimates of specific physical properties of the biological system. The true test of a stimulus–response function model is not whether the fitted parameters can be mapped onto low-level biological mechanisms, but whether the model can successfully predict neuronal responses to novel instances of the sensory input. This review has included a summary of means by which the quality of model predictions can be rigorously and systematically quantified, in a manner robust to the level of stimulus-independent "noise" in the neuronal responses. Such methods for evaluating model predictive power — combined with a healthy appreciation for the potential issues arising from model mismatch — help to make abstract stimulus–response function models an essential tool in the arsenal of methods for analysis of neural systems.

Data Sharing

Python code implementing many of the estimators described above and some example data sets will be made available on-line at http://www.gatsby.ucl.ac.uk/resources/srf/.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

AM and MS implemented the estimation methods. AM implemented and conducted all simulations and analyses and created all figures in this manuscript. All authors wrote the manuscript.

FUNDING

AM and MS were supported by Gatsby Charitable Foundation and Simons Foundation (SCGB323228). JL was supported by Action on Hearing Loss .

REFERENCES

Ahrens, M. B., Linden, J. F., and Sahani, M. (2008a). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *J Neurosci* 28, 1929–1942. doi:10.1523/JNEUROSCI.3377-07.2008

Ahrens, M. B., Paninski, L., and Sahani, M. (2008b). Inferring input nonlinearities in neural encoding models. *Network* 19, 35–67. doi:http://dx.doi.org/10.1080/09548980701813936

Atiani, S., Elhilali, M., David, S. V., Fritz, J. B., and Shamma, S. A. (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron* 61, 467–480. doi:10.1016/j.neuron.2008.12.027

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization* (Cambridge University Press)

Brenner, N., Strong, S. P., Koberle, R., Bialek, W., and de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Comput* 12, 1531–1552

Brosch, M. and Schreiner, C. E. (1997). Time course of forward masking tuning curves in cat primary auditory cortex. *Journal of neurophysiology* 77, 923–943

Brown, E. N., Kass, R. E., and Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nat Neurosci* 7, 456–461. doi:10.1038/nn1228

Brown, E. N., Nguyen, D. P., Frank, L. M., Wilson, M. A., and Solo, V. (2001). An analysis of neural receptive field plasticity by point process adaptive filtering. *Proc Natl Acad Sci U S A* 98, 12261–12266. doi:10.1073/pnas.201409398

Busse, L., Ayaz, A., Dhruv, N. T., Katzner, S., Saleem, A. B., Schölvinck, M. L., et al. (2011). The detection of visual contrast in the behaving mouse. *J Neurosci* 31, 11351–11361. doi:10.1523/JNEUROSCI.6689-10.2011

Bussgang, J. J. (1952). *Crosscorrelation functions of amplitude-distorted Gaussian signals*. Tech. rep., Res. Lab. Elec., Mas. Inst. Technol., Cambridge MA

Butts, D. A., Weng, C., Jin, J., Alonso, J.-M., and Paninski, L. (2011). Temporal precision in the visual pathway through the interplay of excitation and stimulus-driven suppression. *J Neurosci* 31, 11313–11327. doi:10.1523/JNEUROSCI.0434-11.2011

Butts, D. A., Weng, C., Jin, J., Yeh, C.-I., Lesica, N. A., Alonso, J.-M., et al. (2007). Temporal precision in the neural code and the timescales of natural vision. *Nature* 449, 92–95. doi:10.1038/nature06105

Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat Rev Neurosci* 13, 51–62. doi:10.1038/nrn3136

Chen, G., Dan, Y., and Li, C.-Y. (2005). Stimulation of non-classical receptive field enhances orientation selectivity in the cat. *J Physiol* 564, 233–243. doi:10.1113/jphysiol.2004.080051

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

Chornoboy, E. S., Schramm, L. P., , and Karr, A. F. (1988). Maximum likelihood identification of neural point process systems 59, 265–275

Christianson, G. B., Sahani, M., and Linden, J. F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *J Neurosci* 28, 446–455. doi:10.1523/JNEUROSCI.1775-07.2007

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc Natl Acad Sci U S A* 109, 2144–2149. doi:10.1073/pnas.1117717109

David, S. V. and Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network* 16, 239–260

David, S. V., Mesgarani, N., and Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network* 18, 191–212. doi:10.1080/09548980701609235 deBoer, E. and Kuyper, P. (1968). Triggered correlation. *IEEE Transactions on Biomedical Engineering* BM15, 169–179

DeWeese, M. R., Wehr, M., and Zador, A. M. (2003). Binary spiking in auditory cortex. *J Neurosci* 23, 7940–7949

Eden, U. T., Frank, L. M., Barbieri, R., Solo, V., and Brown, E. N. (2004). Dynamic analysis of neural encoding by point process adaptive filtering. *Neural Comput* 16, 971–998. doi:10.1162/089976604773135069

Fitzgerald, J. D., Rowekamp, R. J., Sincich, L. C., and Sharpee, T. O. (2011a). Second order dimensionality reduction using minimum and maximum mutual information models. *PLoS Comput Biol* 7, e1002249. doi:10.1371/journal.pcbi.1002249

Fitzgerald, J. D., Sincich, L. C., and Sharpee, T. O. (2011b). Minimal models of multidimensional computations. *PLoS Comput Biol* 7, e1001111. doi:10.1371/journal.pcbi.1001111

Franz, M. O. and Schölkopf, B. (2006). A unifying view of wiener and volterra theory and polynomial kernel regression. *Neural Comput.* 18, 3097–3118. doi:http://dx.doi.org/10.1162/neco.2006.18.12.3097

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci* 6, 1216–1223. doi:10.1038/nn1141

Gardner, W. A. (1992). A unifying view of coherence in signal processing. *Signal Processing* 29, 113–140

Gerstner, W. and Kistler, W. (2002). *Spiking Neuron Models: An Introduction* (New York, NY, USA: Cambridge University Press)

Gill, P., Zhang, J., Woolley, S. M. N., Fremouw, T., and Theunissen, F. E. (2006). Sound representation methods for spectro-temporal receptive field estimation. *J Comput Neurosci* 21, 5–20. doi:10.1007/s10827-006-7059-4

Goris, R. L. T., Movshon, J. A., and Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nat Neurosci* 17, 858–865. doi:10.1038/nn.3711

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Wiley)

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B* 46, 149–192

Harris, K. D. and Thiele, A. (2011). Cortical state and attention. *Nat Rev Neurosci* 12, 509–523. doi:10.1038/nrn3084

Hsu, A., Borst, A., and Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network* 15, 91–109

Kayser, C., Wilson, C., Safaai, H., Sakata, S., and Panzeri, S. (2015). Rhythmic auditory cortex activity at multiple timescales shapes stimulus-response gain and background firing. *J Neurosci* 35, 7750–7762. doi:10.1523/JNEUROSCI.0268-15.2015

Linden, J. F., Liu, R. C., Sahani, M., Schreiner, C. E., and Merzenich, M. M. (2003). Spectrotemporal structure of receptive fields in areas ai and aaf of mouse auditory cortex. *J Neurophysiol* 90, 2660–2675. doi:10.1152/jn.00751.2002

Lindgren, B. W. (1993). *Statistical Theory* (Boca Raton, FL: Chapman & Hall/CRC), 4th edn.

Machens, C. K., Wehr, M. S., and Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *J Neurosci* 24, 1089–1100. doi:10.1523/JNEUROSCI.4445-03.2004

Marmarelis, P. Z. and Marmarelis, V. Z. (1978). *Analysis of Physiological Systems* (New York: Plenum Press)

McFarland, J. M., Cui, Y., and Butts, D. A. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput Biol* 9, e1003143. doi:10.1371/journal.pcbi.1003143

Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* doi:10.1038/nature11020

Meyer, A. F., Diepenbrock, J.-P., Happel, M. F., Ohl, F. W., and Anemüller, J. (2014a). Discriminative learning of receptive fields from responses to non-gaussian stimulus ensembles. *PLOS ONE* 9, e93062. doi:10.1371/journal.pone.0093062

Meyer, A. F., Diepenbrock, J.-P., Ohl, F. W., and Anemüller, J. (2013). Quantifying neural coding noise in linear threshold models. In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*. 1127–1130. doi:10.1109/NER.2013.6696136

Meyer, A. F., Diepenbrock, J.-P., Ohl, F. W., and Anemüller, J. (2014b). Temporal variability of spectro-temporal receptive fields in the anesthetized auditory cortex. *Frontiers in Computational Neuroscience* 8. doi:10.3389/fncom.2014.00165

Meyer, A. F., Diepenbrock, J.-P., Ohl, F. W., and Anemüller, J. (2015). Fast and robust estimation of spectro-temporal receptive fields using stochastic approximations. *J Neurosci Methods* doi:10.1016/j.jneumeth.2015.02.009

Mineault, P. J., Zanos, T. P., and Pack, C. C. (2013). Local field potentials reflect multiple spatial scales in v4. *Front Comput Neurosci* 7, 21. doi:10.3389/fncom.2013.00021

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General* 135, 370–384

Pachitariu, M., Lyamzin, D. R., Sahani, M., and Lesica, N. A. (2015). State-dependent population coding in primary auditory cortex. *J Neurosci* 35, 2058–2073. doi:10.1523/JNEUROSCI.3318-14.2015

Paninski, L. (2003a). Convergence properties of three spike-triggered analysis techniques. *Network* 14, 437–464

Paninski, L. (2003b). Estimation of entropy and mutual information. *Neural Comput.* 15, 1191–1253. doi:10.1162/089976603321780272

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

Paninski, L., Ahmadian, Y., Ferreira, D. G., Koyama, S., Rad, K. R., Vidne, M., et al. (2010). A new look at state-space models for neural data. *Journal of Computational Neuroscience* 29, 107–126

Park, I. M., Archer, E. W., Priebe, N., and Pillow, J. W. (2013). Spectral methods for neural characterization using generalized quadratic models. In *Advances in Neural Information Processing Systems 26*, eds. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc.). 2454–2462

Park, M. and Pillow, J. W. (2011). Receptive field inference with localized priors. *PLoS Comput Biol* 7, e1002219. doi:10.1371/journal.pcbi.1002219

Pienkowski, M. and Eggermont, J. J. (2010). Nonlinear cross-frequency interactions in primary auditory cortex spectrotemporal receptive fields: a Wiener-Volterra analysis. 28, 285–303. doi:10.1007/s10827-009-0209-8

Pillow, J. W. and Simoncelli, E. P. (2006). Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis. *J Vis* 6, 414–428. doi:10.1167/6.4.9

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron* 70, 1178–1191. doi:10.1016/j.neuron.2011.04.030

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2012). Spectrotemporal contrast kernels for neurons in primary auditory cortex. *J Neurosci* 32, 11271–11284. doi:10.1523/JNEUROSCI.1715-12.2012

Ramirez, A. D. and Paninski, L. (2014). Fast inference in generalized linear models via expected log-likelihoods. *J Comput Neurosci* 36, 215–234. doi:10.1007/s10827-013-0466-4

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

Rosset, S., Zhu, J., and Hastie, T. (2003). Margin maximizing loss functions. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*. 1237–1244

Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron* 46, 945–956. doi:10.1016/j.neuron.2005.05.021

Sahani, M. (2000). Kernel regression for neural systems identification. In *Workshop on Information and Statistical Structure in Spike Trains, NIPS*

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

Sahani, M. and Linden, J. F. (2003b). How linear are auditory cortical responses? (Cambridge, Massachusetts: MIT Press), vol. 15, 109–116

Sahani, M., Williamson, R. S., Ahrens, M. B., , and Linden, J. F. (2013). Probabilistic methods for linear and multilinear models. In *Handbook of Modern Techniques in Auditory Cortex*, eds. D. Depirieux and M. Elhilahi (Hauppage, NY: Nova)

Saleem, A. B., Chadderton, P., Apergis-Schoute, J., Harris, K. D., and Schultz, S. R. (2010). Methods for predicting cortical up and down states from the phase of deep layer local field potentials. *J Comput Neurosci* 29, 49–62. doi:10.1007/s10827-010-0228-5

Schinkel-Bielefeld, N., David, S. V., Shamma, S. A., and Butts, D. A. (2012). Inferring the role of inhibition in auditory processing of complex natural stimuli. *J Neurophysiol* 107, 3296–3307. doi:10.1152/jn.01173.2011

Schwartz, O., Chichilnisky, E. J., and Simoncelli, E. P. (2002). Characterizing neural gain control using spike-triggered covariance. In *Advances in Neural Information Processing Systems 14*, eds. T. Dietterich, S. Becker, and Z. Ghahramani (MIT Press). 269–276

Schwartz, O., Pillow, J. W., Rust, N. C., and Simoncelli, E. P. (2006). Spike-triggered neural characterization. *J Vis* 6, 484–507. doi:10.1167/6.4.13

Scott, J. and Pillow, J. W. (2012). Fully bayesian inference for neural models with negative-binomial spiking. In *Advances in Neural Information Processing Systems 25*, eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.). 1898–1906

Sharpee, T., Rust, N. C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput* 16, 223–250. doi:10.1162/089976604322742010

Sharpee, T. O., Miller, K. D., and Stryker, M. P. (2008). On the importance of static nonlinearity in estimating spatiotemporal neural filters with natural stimuli. *J Neurophysiol* 99, 2496–2509. doi:10.1152/jn.01397.2007

Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., and Miller, K. D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature* 439, 936–942. doi:10.1038/nature04519

Stanley, G. B. (2002). Adaptive spatiotemporal receptive field estimation in the visual pathway. *Neural Comput* 14, 2925–2946. doi:10.1162/089976602760805340

Sutter, M. L., Schreiner, C. E., McLean, M., O'connor, K. N., and Loftus, W. C. (1999). Organization of inhibitory frequency receptive fields in cat primary auditory cortex. *J Neurophysiol* 82, 2358–2371

Theunissen, F. E., Sen, K., and Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J Neurosci* 20, 2315–2331

Tolhurst, D. J., Movshon, J. A., and Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res* 23, 775–785

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J Neurophysiol* 93, 1074–1089. doi:10.1152/jn.00697.2004

Williamson, R. S., Ahrens, M. B., Linden, J. F., and Sahani, M. (2016). Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. *Neuron* doi:10.1016/j.neuron.2016.05.041

Williamson, R. S., Sahani, M., and Pillow, J. W. (2015). The equivalence of information-theoretic and likelihood-based methods for neural dimensionality reduction. *PLoS Comput Biol* 11, e1004141. doi:10.1371/journal.pcbi.1004141

Willmore, B. D. B., Schoppe, O., King, A. J., Schnupp, J. W. H., and Harper, N. S. (2016). Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. *J Neurosci* 36, 280–289. doi:10.1523/JNEUROSCI.2441-15.2016

Yu, J. J. and Young, E. D. (2000). Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus 97, 11780–11786. doi:10.1073/pnas.97.22.11780

FIGURES