Author/s:
Almasi, A;Sun, SH;Jung, YJ;Ibbotson, M;Meffin, H

Title:
Data-driven modelling of visual receptive fields: comparison between the generalized quadratic model and the nonlinear input model

Date:
2024

Citation:
Almasi, A., Sun, S. H., Jung, Y. J., Ibbotson, M. & Meffin, H. (2024). Data-driven modelling of visual receptive fields: comparison between the generalized quadratic model and the nonlinear input model. Journal of Neural Engineering, 21 (4), https://doi.org/10.1088/1741-2552/ad5d15.

Persistent Link:
https://hdl.handle.net/11343/353975

Journal of Neural Engineering

PAPER • **OPEN ACCESS**

Data-driven modelling of visual receptive fields: comparison between the generalized quadratic model and the nonlinear input model

To cite this article: Ali Almasi *et al* 2024 *J. Neural Eng.* **21** 046014

IOP Publishing J. Neural Eng. 21 (2024) 046014 https://doi.org/10.1088/1741-2552/ad5d15

Journal of Neural Engineering

OPEN ACCESS

RECEIVED
3 October 2023

REVISED
17 April 2024

ACCEPTED FOR PUBLICATION
28 June 2024

PUBLISHED
12 July 2024

PAPER

Data-driven modelling of visual receptive fields: comparison between the generalized quadratic model and the nonlinear input model

Ali Almasi[1,5], **Shi H Sun**[1,4,5], **Young Jun Jung**[1], **Michael Ibbotson**[1,2,5] **and Hamish Meffin**[1,3,5,*]

[1] National Vision Research Institute, Carlton, VIC 3053, Australia
[2] Department of Optometry and Vision Sciences, The University of Melbourne, Parkville, VIC 3010, Australia
[3] Department of Biomedical Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
[4] Current Affiliation: Department of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, United States of America
[5] These authors contributed equally to this work.
[*] Author to whom any correspondence should be addressed.

E-mail: hmeffin@unimelb.edu.au

Keywords: visual processing, receptive fields, model comparison

Abstract

Objective: Neurons in primary visual cortex (V1) display a range of sensitivity in their response to translations of their preferred visual features within their receptive field: from high specificity to a precise position through to complete invariance. This visual feature selectivity and invariance is frequently modeled by applying a selection of linear spatial filters to the input image, that define the feature selectivity, followed by a nonlinear function that combines the filter outputs, that defines the invariance, to predict the neural response. We compare two such classes of model, that are both popular and parsimonious, the generalized quadratic model (GQM) and the nonlinear input model (NIM). These two classes of model differ primarily in that the NIM can accommodate a greater diversity in the form of nonlinearity that is applied to the outputs of the filters. *Approach*: We compare the two model types by applying them to data from multielectrode recordings from cat primary visual cortex in response to spatially white Gaussian noise After fitting both classes of model to a database of 342 single units (SUs), we analyze the qualitative and quantitative differences in the visual feature processing performed by the two models and their ability to predict neural response. *Main results*: We find that the NIM predicts response rates on a held-out data at least as well as the GQM for 95% of SUs. Superior performance occurs predominantly for those units with above average spike rates and is largely due to the NIMs ability to capture aspects of the model's nonlinear function cannot be captured with the GQM rather than differences in the visual features being processed by the two different models. *Significance*: These results can help guide model choice for data-driven receptive field modelling.

1. Introduction

Visual perception depends on the processing of visual input through multiple stages. An important technique in understanding the visual processing at each of these stages is the characterization of neuronal receptive fields (RFs), which provide neural representations of the visual input using a set of image features that best describe the neural responses. The first attempts to characterize the processing that occurs in RFs of primary visual cortex (V1) was achieved using stimuli such as flashed bars or drifting gratings and showed that many of these cells are selective for the orientation features (e.g. Hubel and Wiesel 1962, Movshon *et al* 1978; see Carandini 2006 for a review). These studies identified RFs that operate using linear summation (simple cells), and others that operated using nonlinear summation mechanisms (complex cells). In contrast to simple cells, the nonlinear summation of complex cells allows them to respond with a degree of invariance to shifts in the position of an appropriately oriented bar/grating. These findings led to the formation of some phenomenological models for simple cells (DeAngelis *et al* 1993) and complex

cells. For simple cells a linear–nonlinear (LN) model captures linear spatial integration with a single spatial filter to model the RF, followed by a static nonlinearity to convert the filter output to a spike rate. A well-known phenomenological model of complex cells is the energy model (Adelson and Bergen 1985), which describes the responses of complex cells by using a pair of spatial filters with similar orientation and spatial frequency preferences but with spatial phases that are shifted by 90° relative to each other. The nonlinear summation occurs via squaring and summing the filter outputs, giving rise to spatial phase invariance. Other phenomenological models have achieved similar invariance properties via a nonlinear max-pooling operation over the outputs of a larger bank of oriented spatial filters with differing spatial phases (Einhäuser *et al* 2002, Serre and Riessenhuer 2004, Hansard and Horaud 2011; see Lindeberg 2020, for an account of generalizing invariance properties to rotations and scaling). Another class of models have shown how the requisite features and invariance can be learnt from natural images or movies through synaptic plasticity according to principles such as slow feature extraction or sparse coding (Einhaüser *et al* 2002, Körding *et al* 2004, Berkes and Wiskott 2005, Lian *et al* 2021). Others have shown that basic RFs can form even prior to visually structured input, consistent with developmental studies (Merolla and Boahn 2004).

In parallel to these phenomenological and learning-based approaches to modeling, another approach has been to estimate models of RFs directly from electrophysiological data of responses to large ensembles of more complex and statistically rich stimuli, including random combinations of bars, mixtures of gratings, white Gaussian noise (WGN) and natural scenes (e.g. Chichilnisky 2001, Touryan *et al* 2005). Such techniques provided a shift from quantitative RF mapping and qualitative models, to quantitative data-driven modeling that enables a more quantitative and objective estimate of RFs as filters and how those filter outputs are combined via nonlinear integration.

One of the earliest data-driven modeling studies of simple and complex cells was based on an adaptation of the Wiener kernel expansion (Emerson *et al* 1987). While this allowed estimation of multiple RF subunits, in practice only second order, i.e. quadratic, nonlinearities could be estimated given limited data. To accommodate more general non-linearities, a class of reverse correlation methods called spike-triggered analyses were developed, which analyses the average (STA) or covariance (STC) of stimuli that precede a neuronal response (Touryan *et al* 2002, 2005, Rust *et al* 2005, Schwartz *et al* 2006). These techniques are effective in characterizing the stimuli that drive a neuronal response, but suffer from some shortcomings. STA assumes linear spatial summation and, therefore, fails for cells that show nonlinear spatial summation properties. STC can overcome this nonlinear spatial summation property by finding multiple filters, but it struggles to accurately find the nonlinear component of processing as the number of filters increase, a problem that is commonly referred to as the 'curse of dimensionality' (Bellman 1957). Furthermore, STC is prone to artifacts when the stimulus is anything other than WGN (Schwartz *et al* 2006).

To combat the 'curse of dimensionality' and to successfully use stimuli other than Gaussian white noise, other techniques, such as maximum likelihood estimation, have employed probabilistic parametric models to guarantee consistent model inference. The most basic of these is the LN Poisson model, suitable for simple cells. As mentioned above, it uses a single spatial (or spatio-temporal) filter in combination with a static nonlinearity to predict spike rate, which, in an additional step, determines the spike probability via a Poisson process. For complex cells, this model is elaborated into a two-stage linear–nonlinear linear–nonlinear (LN–LN) cascade, in which output rates from multiple different LN models in a first LN stage, are combined via weighted summation and a further nonlinearity in the second LN stage. The final output rate is again fed into a Poisson process to determine spike probability. One popular LN–LN model that can combine multiple filters is the generalized quadratic model (GQM) (Fitzgerald *et al* 2011, Park *et al* 2013). Guided by the prior knowledge about V1 cells, the GQM assumes that the overall RF nonlinearity consists of linear and quadratic input functions to capture the response properties of simple and complex cells, respectively (based on LN models: Heeger 1992; and energy models: Adelson and Bergen 1985) (figure 1). Note that in the case of a single filter, the LN–LN cascade reduces to a single-stage LN model for simple cells.

One critique of the GQM is that the linear or quadratic input functions used in the model, although simplifying, are not the only functions observed biologically. In principle, many other function types are at least as biologically plausible as the linear or quadratic kind. Hence, this simplifying assumption about input functions is a limitation of the GQM. The nonlinear input model (NIM) rectifies this issue by making the input functions completely general (i.e. non-parametric), rather than fixed (McFarland *et al* 2013) (figure 1). The modeling framework assumes that the input functions can be estimated given the neural data, thus providing a more general and physiologically plausible modeling framework. While there are other.

In this study, we will characterize the RFs of primary visual cortical neurons in anaesthetized cats using the NIM and the GQM: two general and robust RF characterization techniques. While the NIM provides the more general model framework, being able to capture a greater diversity of response

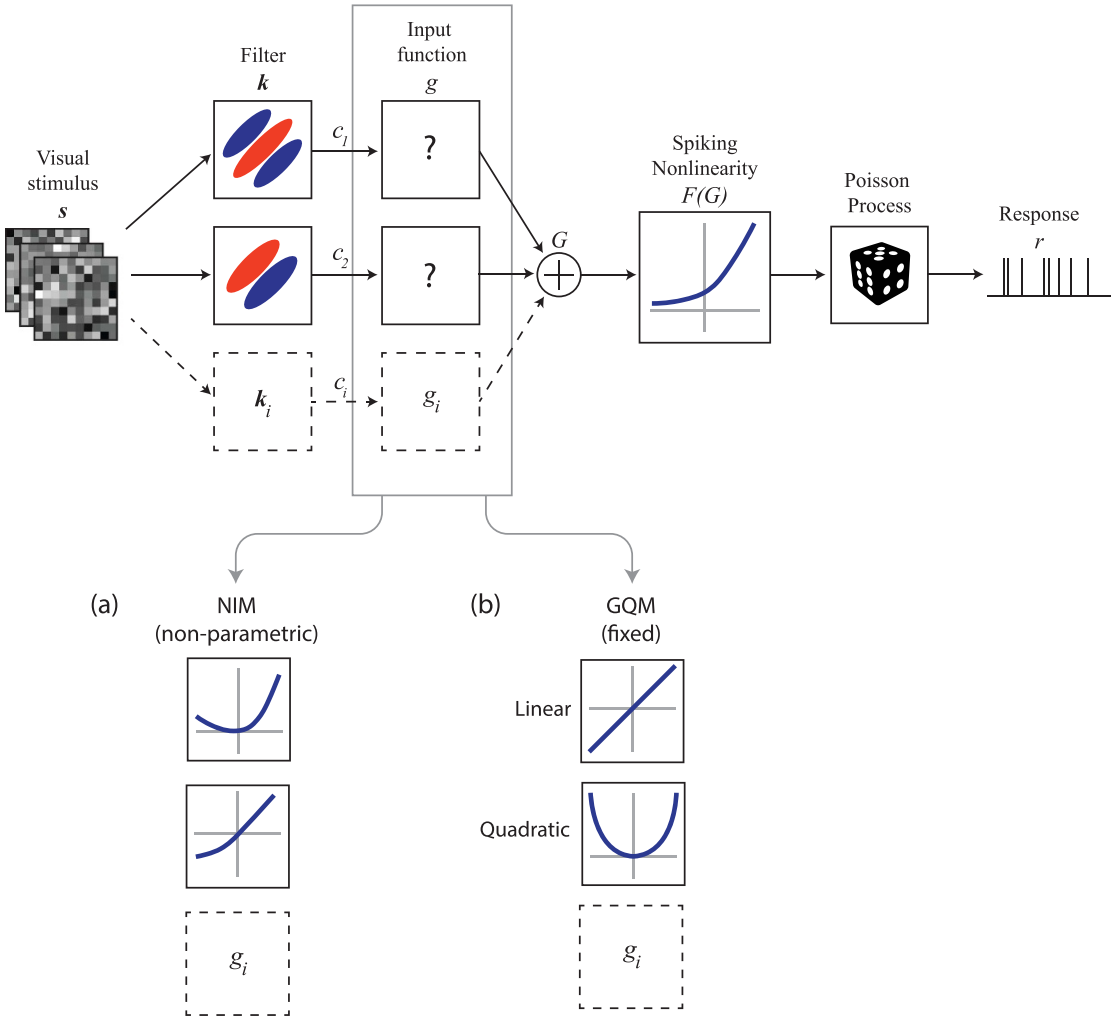

Figure 1. The RF models of V1 cells. The input functions applied on outputs of the spatial RF filters differentiate between the generalized quadratic model and the nonlinear input model.

nonlinearities, it does so at the expense of using a larger number of parameter than the GQM, which may require more data or result in overfitting. Here will compare the two techniques' ability to characterize RFs using the same neural data obtained using stimulation with WGN. We aim to address the following two questions: *What are the qualitative differences in characterization between the GQM and the NIM*, and *Which is quantitatively better at predicting responses*? Our results show that for 67% of V1 cells the two models are comparable in performance. However, for most remaining V1 cells the NIM tends to outperform the GQM. We will also discuss how the availability of more neural data for the model fitting better favors the NIM compared to the GQM.

2. Methods

All of our experimental procedures and analysis are explained in detail in our previous studies (Almasi *et al* 2020, 2022, Sun *et al* 2021). Here we will describe briefly.

2.1. Experimental design

2.1.1. Animal preparation

Recordings were made from ten adult cats (*Felis catus*, M/F, 2–6 kg), maintained in the facility at the National Vision Research Institute approved by the Animal Care Ethics Committee at The University of Melbourne (ethics ID 1814462.1). Animals were anaesthetized with ketamine (10 mg kg^{-1}; i.m), medetomidine (15 μg kg^{-1}; i.m) and methadone (0.4 mg kg^{-1}; i.m). Following intravenous cannulation, anesthesia was maintained using gaseous isoflurane (1.0%–1.5% in O_2). A craniotomy and durotomy (10 mm anterior posterior and 8 mm wide) were then performed. During electrophysiological recordings, anesthesia was moved to halothane (0.5%–1.0% in a 2:1 mixture of O_2 and N_2O) and the animal paralyzed by a continuous infusion of vecuronium bromide (0.1 mg kg^{-1} h^{-1}) to prevent eye movement. Daily injections of atropine (0.05 mg/kg/; s.c), dexamethasone phosphate (1.5 mg kg/; i.m) and Clavulox (0.2 ml kg^{-1}; i.m) were administered daily to reduce salivation, cerebral edema and infection, respectively. Contact lenses were inserted to prevent the eyes from

dehydrating and eye drops (1% atropine sulfate; 10% phenylephrine hydrochloride) were given daily to dilate the pupils and retract the nictitating membranes. Refractive errors were assessed by reverse ophthalmoscopy and corrected using spherical lenses placed in front of the eyes. At the conclusion of the experiment, animals were euthanized with an intravenous injection of an overdose of barbiturate (sodium pentobarbital; 150 mg kg^{-1}).

2.1.2. Visual stimuli

Visual stimuli were presented on a gamma corrected LED monitor (ASUS VG248QE, 1920 × 1080 pixels, 60 Hz, 1 ms response time) with a ViSaGe visual stimulus generator (Cambridge Research Systems, Cambridge, UK) at a viewing distance of 57 cm. WGN stimuli were generated through MATLAB© with a mean value equal to the mean luminance of the display monitor. Dimensions of WGN images ranged from 45 × 45 to 90 × 90 pixels (with 15 pixel increments) and were kept the same for the duration of recording. We often used 90 × 90 pixels images as they provided higher spatial resolution stimuli. However, occasionally for units with bigger RFs lower resolution images were employed. WGN stimuli were presented over 30° of the visual field at 30 Hz. Each noise image was followed by a 1/30 s blank screen to increase the responsiveness of the units.

2.1.3. Data acquisition

Two types of 32-channel multi-electrode arrays (MEAs) from NeuroNexus were used: single linear probe with 32 sites (1 × 32,) with 100 μm spacing; or four linear probes with eight sites on each (4 × 8, one penetration) spaced at 100 μm. MEAs were inserted into the cortex by a piezoelectric drive (Burleigh inchworm and 6000 controller, Burleigh Instruments, Rochester, NY) and signals were acquired by a CerePlex acquisition system and Central software (Blackrock Microsystems, Salt Lake City, Utah) sampled at 30 kHz.

2.1.4. Post-processing and spike sorting

Spike sorting of recordings was performed using KiloSort (Pachitariu *et al* 2016) and the graphical user interface *phy* (Rossant *et al* 2016). Single units (SUs) were identified as previously described in Almasi *et al* (2020) and Sun *et al* (2021).

2.2. RF analysis

Spikes were pooled into spike counts at the optimum response latency (defined at the dip in the post-stimulus time histogram from stimulus onset to 1/15 s in 1 ms bins) within a window of 1/30 s. The optimum response latency used for spike pooling in RF estimation of SUs was calculated by observing the time when neurons start to spike following a stimulus presentation.

Before model estimation, all SUs were assigned a spatial region of interest within the input image that covered the RF by finding the linear RF using the spike-triggered average (STA) and manually drawing a window around the spatial RF.

2.2.1. Nonlinear input model (NIM)

RFs were then estimated using a modified version of the NIM (McFarland *et al* 2013) as described in our previous study (Almasi *et al* 2020, 2022). The NIM has five main processing stages as shown in figure 1: (1) the stimulus is processed by the spatial filters; (2) passed through the corresponding input functions; (3) combined into a generator signal via summation; (4) a spiking nonlinearity is applied to produce a spike rate; and (5) the firing rate is used to randomly generate spikes according to a Poisson point process. This model is described by the response r given the visual stimulus $\boldsymbol{s}$ as:

$$\begin{aligned} p(r|\mathbf{s}) &\sim \text{Poisson}(\eta), \\ \eta &= F\left(\sum_{i=1}^{m} g_i(\mathbf{k}_i \cdot \mathbf{s})\right), \\ p(r|\mathbf{s};\eta) &= \frac{\eta^r \exp(-\eta)}{r!} \end{aligned} \tag{1}$$

where η represents the firing rate in response to $\boldsymbol{s}$ and m is the total number of filters. $\mathbf{k}_i$ represents the ith spatial filter that processes the stimulus to produce feature-contrast (step 1), defined as their inner product $c_i = \mathbf{k}_i \cdot \mathbf{s}$. The $g_i(\cdot)$ is the input function that is applied on the feature-contrast of ith filter that assumes an essentially arbitrary form (step 2). The input signals are then combined to form a generator signal ($G = \sum_{i=1}^{m} g_i(\mathbf{k}_i \cdot \mathbf{s})$) (step 3), $F(\cdot)$ is the spiking nonlinearity that converts the generator signal to the spike rate η (step 4), and then the *Poisson* process converts spike rate into spike times (step 5; last line in equation (1)).

The estimation of the parameters for the NIM (filters $\mathbf{k}_i$, input functions $g_i(\cdot)$, and spiking nonlinearity $F(\cdot)$) was done by maximizing the log-likelihood (LL) of the model given the neural data and stimuli, which is:

$$LL(\mathbf{k}_i, g_i, F|r, s) = \sum_{t=1}^{N} [r_t \log(\eta) - \eta] + \text{constant} \tag{2}$$

where N are mutually independent instances of the stimuli whose responses are $r_1, \ldots, r_N$, i.e. spike counts. Note that the constant term $\log(r!)$ can be neglected in equation (2) since it does not depend on the model parameters. Maximization was performed via gradient ascent. All details regarding fitting the NIM are identical to those describe previously (Almasi *et al* 2020) and readers are referred to this paper for comprehensive details, including sections in that paper on Model Definition,

Mathematical Representation of the filters, input functions and spiking nonlinearity, Estimating Best Maxima, including Identifying the Subspace of RF Filters and the Brute Force Search to Find the Best Local Maximum through a simultaneous estimation procedure, and finally Avoiding Model Degeneracy. In brief, we note that the filters were represented using a spatial Fourier basis, the input functions were represented with a tent basis and the spiking nonlinearity was fit to a log-exponential function $F(v) = \alpha \log\left[1+\exp\left(\frac{v-\gamma}{\alpha}\right)\right]+\delta$. An additional summary of how the optimal number of filters was obtained is provided below for convenience that follows Almasi *et al* (2020).

2.2.2. Generalized quadratic model (GQM)

This model is in fact a subclass of the NIM in which input functions are held fixed to a combination of linear and quadratic functions, as:

$$\eta = F\left(w_0\left(\mathbf{k}_0\cdot\mathbf{s}\right)+\sum_{i=1}^{n}\left(w_i\left(\mathbf{k}_i\cdot\mathbf{s}\right)\right)^2\right), \qquad (3)$$

where w_i denotes the corresponding weight of filter $\mathbf{k}_i$ and n is the total number of quadratic filters (equal to the total number of filters minus one, $m-1$). In above, the filters are normalized, i.e. they have unit norm $\|\mathbf{k}_i\|=1$. The positive/negative sign of the weights of quadratic filters indicate their excitatory/inhibitory contribution to the overall response.

The model was fit with the same routine as used for NIM, namely maximizing log-likelihood via gradient ascent. Other relevant details were also identical for GQM and NIM, and readers are referred to Almasi *et al* (2020) for a comprehensive description. In summary the following were identical: model representation of filters, $\mathbf{k}_i$, and the spiking nonlinearity, F; procedures for identifying the optimal number of filters and the corresponding optimal filter subspace; methods to search for the best maxima of the log-likelihood through multiple, systematic initializations of the filters across the optimal filter subspace prior to gradient ascent; and finally avoiding degeneracy in the model by choosing the filters to have unit norm $\|\mathbf{k}_i\|=1$. As the input nonlinearities, g_i, are fixed to be either linear or quadratic in the GQM, procedures used in the NIM to optimize the parameters for the g_i or avoid degeneracies associated with the g_i were unnecessary for the GQM.

Following optimization of the GQM to obtain the filters, $\mathbf{k}_i$, the weights, w_i, and the spiking nonlinearity, F, it is useful to transform the model to a mathematically equivalent form with a unique set of transformed filters, $\mathbf{h}_i$, that are guaranteed to be orthonormal, as follows. This version of the GQM model uses the transformed filters, $\mathbf{h}_i$, as a basis to represent the same subspace spanned by the filters, $\mathbf{k}_i$. The transformation to an orthonormal representation allows both easier visualization and comparison to the NIM, such as shown in figure 7. The quadratic filters, $\mathbf{k}_i$, span a space to which we refer as the quadratic feature space. Mathematically, the quadratic feature space is the column space of the matrix collecting all the spatial RF filters as its columns $K=[\mathbf{k}_1\cdots\mathbf{k}_n]$, which contains all possible linear combinations of the filters. The model response may also be written as

$$\eta = F\left(w_0\left(\mathbf{k}_0\cdot\mathbf{s}\right)+\ \mathbf{s}^T Q\mathbf{s}\right), \qquad (4)$$

where $Q=KWK^{\mathrm{T}}$ and $W=\mathrm{diag}\left([w_1\ldots w_n]\right)$ is the matrix that accommodates all the weights of quadratic filters on its main diagonal. The quadratic filters in the GQM are not necessarily orthogonal to each other since there is no orthogonality constraint between the spatial filters in the NIM. Although the quadratic feature space is uniquely represented by the set of quadratic filters, the filters themselves are not unique as any set of filters that span the same quadratic space can provide a representation for this space. To provide a unique and orthonormal representation of the quadratic spatial filters, only for visualization purposes, the eigenvalue decomposition of the matrix Q can be sought as a below

$$Q = H\Lambda H^{\mathrm{T}} \qquad (5)$$

where $\Lambda=\mathrm{diag}([\lambda_1,\ldots,\lambda_n])$ and $H=[\mathbf{h}_1,\ldots,\mathbf{h}_n]$, in which λ_i and $\mathbf{h}_i$ indicate the ith largest eigenvalue and its corresponding eigenvector of Q, respectively. In fact, $\{\mathbf{h}_1,\ldots,\mathbf{h}_n\}$ gives the set of unique quadratic filters that span the quadratic feature space of the unit. Because Q is a real symmetric matrix in equation (5), the right- and left-singular vector matrices are the same (both H). In the event that there are l inhibitory filters ($w_i<0$) in the model, we find and report the l eigenvectors corresponding to the l smallest eigenvalues of Λ. According to eigenvalue decomposition of Q, we have $Q=H\Lambda H^{\mathrm{T}}=\sum_{i=1}^{n}\lambda_i\mathbf{h}_i\mathbf{h}_i^T$. Replacing this into equation (4), we will reach the following formula for the GQM

$$\begin{aligned}\eta &= F\left(w_0\left(\mathbf{k}_0\cdot\mathbf{s}\right)+\mathbf{s}^T\left(\sum_{i=1}^{n}\lambda_i\mathbf{h}_i\mathbf{h}_i^T\right)\mathbf{s}\right)\\ &= F\left(w_0\left(\mathbf{k}_0\cdot\mathbf{s}\right)+\sum_{i=1}^{n}\lambda_i(\mathbf{h}_i\cdot\mathbf{s})^2\right).\end{aligned}$$

Since the linear filter in the GQM is not necessarily orthogonal to the quadratic filters, it may have a considerable correlation with the quadratic filters. To obtain an independent linear feature dimension from the quadratic feature space, we decompose the linear filter into $\mathbf{k}_0=\ \mathbf{h}_0+\mathbf{q}$, where $\mathbf{q}$ is the component that sits in the quadratic feature space spanned by $\{\mathbf{h}_1,\ldots,\mathbf{h}_n\}$ described as

$$\mathbf{q}=\sum_{i=1}^{n}\beta_i\mathbf{h}_i \qquad (6)$$

and $\mathbf{h}_0$ is the component that is perpendicular to $\mathbf{q}$. The coefficients describing $\mathbf{q}$ are

$$\beta_i = \mathbf{k}_0 \cdot \mathbf{h}_i \tag{7}$$

since $\{\mathbf{h}_1, \ldots, \mathbf{h}_m\}$ form an orthonormal set of basis vectors for the quadratic feature space. Consequently, $\mathbf{h}_0$ is obtained as

$$\mathbf{h}_0 = \mathbf{k}_0 - \sum_{i=1}^{n} \beta_i \mathbf{h}_i. \tag{8}$$

Replacing the expression above for the linear filter $\mathbf{k}_0$ in the GQM, one would get

$$\eta = F\left(w_0 (\mathbf{h}_0 \cdot \mathbf{s}) + \sum_{i=1}^{n} \lambda_i \left((\mathbf{h}_i \cdot \mathbf{s} - \mu_i)^2 - \mu_i^2 \right) \right) \tag{9}$$

where

$$\mu_i = -\frac{w_0 \beta_i}{2\lambda_i}. \tag{10}$$

This indicates that for any quadratic feature dimension that shows a non-zero correlation (inner product) with the linear feature dimension, there will be an offset (shift) in the corresponding feature-contrast response (FCR) nonlinearity. This shift provides a better fit to the actual RF nonlinearity of the unit.

2.2.3. Determining number of spatial filters

The number of spatial filters for each unit was determined as a hyper-parameter of the model using cross-validation by varying the number of filters and then evaluating the statistical significance of the variation with bootstrapping. We divided the neural data for each unit into three sets: training (64%), validation (16%), and test (20%). Specifically, the training data was used to estimate a version of the NIM and the GQM with one up to five filters. The validation data was bootstrapped (random re-sampling with replacement) to evaluate the performance of the models with different numbers of filters by calculating the log-likelihood of the filters on the remaining one-fifth of the data. This bootstrapping performance evaluation was repeated 500 times, which created a distribution for the log-likelihood for each model with different filter numbers. To determine the optimal combination of filters to use, the smallest number of filters (one filter) was used as a starting point. Then, a filter was added if the new filter significantly improved the log-likelihood of the model (Z-score > 2) on the validation test. Filters were incremented until there was no significant improvement in the log-likelihood. Furthermore, the combination of filters was compared against a control model, which had the filters replaced with noise (WGN with unit variance). If the optimum model was not significantly better than the noise model, then the RF could not be estimated for that unit. We have used the approach of choosing the simplest and most compact model (smallest number of significant filters) because there are more issues with models of higher complexity; there is inevitably a compromise between complexity and accuracy (Hastie *et al* 2009).

2.3. Uncovered feature spaces

The RF filters uncovered using models will span a space that is termed the neuron's feature space. The uncovered feature space is equal to the column space ($\mathcal{H}$) of the matrix collecting all the model identified RF filters as its columns $H = [\mathbf{h}_1 \cdots \mathbf{h}_m]$, where m denotes the number of RF filters. The feature space, by definition, contains all possible linear combinations of the RF filters:

$$\mathcal{H} = \left\{ \sum_{k=1}^{m} a_k \mathbf{h}_k \right\}, \forall a_k \in \mathbb{R}. \tag{11}$$

2.4. Dihedral angle between feature spaces

The identified RF filters by the NIM or the GQM for each unit result in a feature space for that unit. The feature space is in general a subspace of the input image space, i.e. the space representing all combination of input images, and correspond to hyperplanes in the input image space. A comparison of uncovered feature spaces using each model can thus be achieved by finding the angle between these intersecting hyperplanes, the so called dihedral angle. For this purpose, '*subspace.m*' MATLAB© function is employed, which returns the angle between two hyperplanes embedded in a higher dimensional space.

2.5. Traversing between the NIM and GQM feature spaces

Assuming that K and H are, respectively, the matrices collecting the NIM and GQM RF filters as their columns, the matrix that transforms a point in the NIM feature space to its corresponding point in the GQM feature space is given by $H^{\dagger}K$, where $H^{\dagger}$ is the Moore–Penrose inverse, also called pseudo-inverse, of H. Conversely, the matrix transforming a point in the GQM feature space into the NIM feature space is given by $K^{\dagger}H$, where $K^{\dagger}$ indicates the pseudo-inverse of K. The pseudo-inverse of a matrix is computed using its compact singular value decomposition (SVD). This can be achieved using '*pinv.m*' MATLAB© function. In general, the number of RF filters uncovered by the NIM and GQM are not necessarily the same due to the difference in their structural rigidity. Further, even if the number of RF filters is the same between the NIM and GQM, the corresponding feature spaces might not be the same. Therefore, the above transformation between the feature spaces is not completely lossless in most cases. However, if the feature spaces are close, i.e. the angle

between intersecting hyperplanes representing feature spaces is small, this transformation can be considered approximately reversible.

2.6. Statistical analysis

Unless stated otherwise, all results are shown as the mean ± standard error of the mean. Significance tests involving the mean were done with pairwise student's *t*-tests. All statistical analyses were performed using MATLAB©.

2.6.1. Control for the dihedral angle between feature spaces

We performed a statistical significance test to determine how the measured dihedral angle between the NIM and GQM feature spaces was different to that obtained by pure chance. We did a Monte Carlo simulation for each unit in which we replaced the uncovered NIM and GQM RF filters with random vectors whose elements were drawn independently from an identical uniform distribution in the image space. Since we only cared about the direction of filters in the image space, we did not apply any normalization on the randomly generated filters. The simulation was repeated for 10 000 trials. In each trial we measured the dihedral angle between the feature spaces formed by randomly generated NIM and GQM RF filters. We noticed that the empirical distribution of the dihedral angle between the randomly sampled feature spaces over all trials was nearly *gamma* distributed. Thus, we fitted a *gamma* probability distribution function to the empirical distribution of the dihedral angle between randomly generated feature spaces for each unit in the population. The fitting was done parametrically using '*fitdist.m*' MATLAB© function. Next, for each unit we found the likelihood of the dihedral angle between its NIM and GQM feature spaces being extracted from the fitted probability distribution of random filter directions.

2.6.2. FCR function comparison

The FCR function of a unit is the function that maps outputs of RF filters of a unit (i.e. feature-contrast) to the unit's response. Mathematically, the FCR function is defined as $r = \Gamma(c_1, c_2, \ldots, c_m)$ where $c_i = \mathbf{k}_i \cdot \mathbf{s}$ is the feature-contrast of the ith RF filter $\mathbf{k}_i$. We compared FCR functions that emerged from the NIM and GQM fits for each unit by calculating the pointwise correlation coefficient between the two FCR functions on a common representation of their shared feature subspace. That is, the FCR functions of each model were calculated on a grid of points on this common subspace, and the correlation coefficient was calculated between these values. The common subspace also allowed a visual inspection of the relationship between the two FCR functions (e.g. figure 7). This shared feature subspace and the mapping of each FCR function is described as follows.

In doing so, we first found an orthogonalized basis set for the NIM feature space by performing compact SVD. This is because, as stated previously, the NIM does not impose any orthogonality constraint on the estimated RF filters, and consequently, the uncovered RF filters may not necessarily form an orthogonal basis set, although they are almost always independent. Considering $K = [\mathbf{k}_1 \cdots \mathbf{k}_m]$ represents the matrix collecting all the m identified NIM RF filters as its columns, compact SVD of this matrix yields

$$K = U\Sigma V^{\mathrm{T}}. \tag{12}$$

The ith column in U indicates the left-singular vector corresponding to the ith largest singular value in Σ. In the compact SVD, the number of columns in U will be equal to its rank (i.e. number of independent columns), which in this case is equal to the number of RF filters m because they are independent. Also, Σ will be a $m \times m$ diagonal matrix that includes only non-zero singular values.

In this case, the m left-singular vectors in U form an orthonormal basis set for the NIM feature space. Note that U has a rank, or number of independent columns, that is equal to m. On the other hand, since the original NIM RF filters are normalized (they all have unit norm), the non-zero singular values will be 1, thus making Σ to be identity matrix. Taken together, we will have

$$K = UV^{\mathrm{T}}. \tag{13}$$

Because V is orthonormal, its inverse is given by V^{T}, which yields

$$U = KV. \tag{14}$$

This means that V is the matrix that transforms a given point in the orthogonalized NIM feature space to the original NIM feature space. The inverse transform is given by V^{T}.

Second, we sampled the orthogonalized NIM feature space uniformly using Cartesian coordinates. The number of sampled points depends on the dimensionality of the feature space. We sampled 25, 625, 6859, 50 625 and 161 051 points, respectively, for 1- to 5-dimensional feature spaces at equal distances within the range of feature-contrasts of NIM input functions. This range is determined by 95th percentile of the feature-contrast of WGN stimuli. Using the transformation matrix V, we transformed the sampled points from the orthogonalized feature space to the original NIM feature space. After this step, we excluded points that resulted in feature-contrasts outside the range found for NIM input functions. We

then found the FCR for these sampled points in the original NIM feature space.

Third, using the transformation between the NIM and the GQM feature spaces, we transformed the sampled points in the original NIM feature space to the GQM feature space and we found their FCRs. This allowed the FCRs predicted by the NIM and the GQM to be directly compared to each other, both visually and using a correlation coefficient.

2.6.3. Model performance comparison

After estimating the best possible models for the GQM and the NIM using training set and tuning the hyper-parameters using validation set, we used a held-out test set to compare the model performance of the GQM and the NIM. This was done by bootstrapping (sampling with replacement) from the test set and computing the predictive ability of the model measured using log-likelihood of the GQM and NIM. Later, a two-sample t-test was done on the two distributions to compare the statistical significance of the difference in the model log-likelihoods.

2.6.4. Measuring significance of the linear filter in GQM

The linear filter in the GQM is not necessarily orthogonal to the quadratic filters. As a result, it can show non-zero correlation with the quadratic RF filters, which results in a shift in the quadratic input functions for those filter dimensions (see section 2.2.2). Further, the linear filter may completely be sitting in the quadratic feature space, which indicates that the unit has no filter dimension on which the unit shows linear response dependencies. To determine the statistical significance of the linear residual component (the part that is left after subtracting the quadratic dependencies) and the shift in the quadratic input functions in the GQM, we performed as follows.

First, as mentioned in section 2.2.2, we decomposed the linear filter into $\mathbf{k}_0 = \mathbf{h}_0 + \mathbf{q}$, where $\mathbf{q}$ is the component that sits in the quadratic feature space whereas $\mathbf{h}_0$ is the component that is perpendicular to the quadratic feature space, termed residual linear filter. Then, we formed three GQMs based on the above filters:

(i) The original GQM with $\mathbf{k}_0$ as its linear filter.
(ii) The original GQM with $\mathbf{h}_0$ as its linear filter, termed GQM_{res}.
(iii) The original GQM with $\mathbf{q}$ as its linear filter, termed GQM_{proj}.

We then compared the performance of the above models against the hold-out test set as described in the previous section. The residual linear filter was found significant if and only if GQM outperformed GQM_{proj}. This is because the only difference between these two models is the residual linear filter, which GQM_{proj} lacks. Therefore, if GQM performs better it is due to the inclusion of the residual linear filter. The offset in the quadratic input functions was found significant if and only if GQM outperformed GQM_{res}. This is because the only difference between these two models is the quadratic projection of the linear filter, which GQM_{res} lacks. Therefore, if GQM outperforms the GQM_{res} it indicates that including the quadratic component of the linear filter is crucial for the model to perform better.

3. Results

We recorded from 342 isolated SUs in V1 from 10 anaesthetized adult cats that were responsive to WGN and whose RFs could be uncovered. The RFs were estimated using two methods: (**a**) the NIM, which employs a nonparametric representation for general input functions; and (**b**) the GQM, which is the same as the NIM but imposes the input functions to be either linear or quadratic, as shown in figure 1. For all RFs characterized using the NIM and the GQM, only purely excitatory or mixed excitatory/inhibitory filters were found, despite our characterization techniques allowing identification of purely inhibitory filters. The optimum combination and number of filters for both models were determined as hyperparameters of the model using a statistical significance test as described in the methods.

Figures 2 and 3 show five example units with their spatial RF characterized with the NIM (figures 2 and 3; left column) and the GQM (figures 2 and 3; right column). Within each example, their spatial filters are shown above, and the corresponding input functions are plotted below. The contrast polarity sensitivity of the spatial filters is represented as ON (red) and OFF (blue) with the horizontal and vertical axes denoting space. The input functions presented describe how the unit pools across the output of spatial filters, termed feature-contrast, which indicates the overall feature selectivity and invariance of the unit (Almasi *et al* 2020). For example, *Unit 1* in figure 2 had a single filter that is nearly identical for both NIM and GQM. The corresponding input function is also similar, and is monotonically increasing, resulting in responses of opposing polarity to opposite feature-contrasts. The input function is identified to be linear by the GQM, while the function estimated by the NIM exhibits signs of saturation, therefore is deviating from a perfect linear function. Units with a single linear filter typically correspond to simple units with distinct ON and OFF regions in the whole RF.

In contrast, both the NIM and the GQM characterized *Unit 2* with quadrature pair filters, which consist of two filters with similar orientation and spatial frequency preferences but 90° spatial phase differences. For this unit the filters are also largely indistinguishable between models. Further, both models have quadratic-shaped input functions that respond

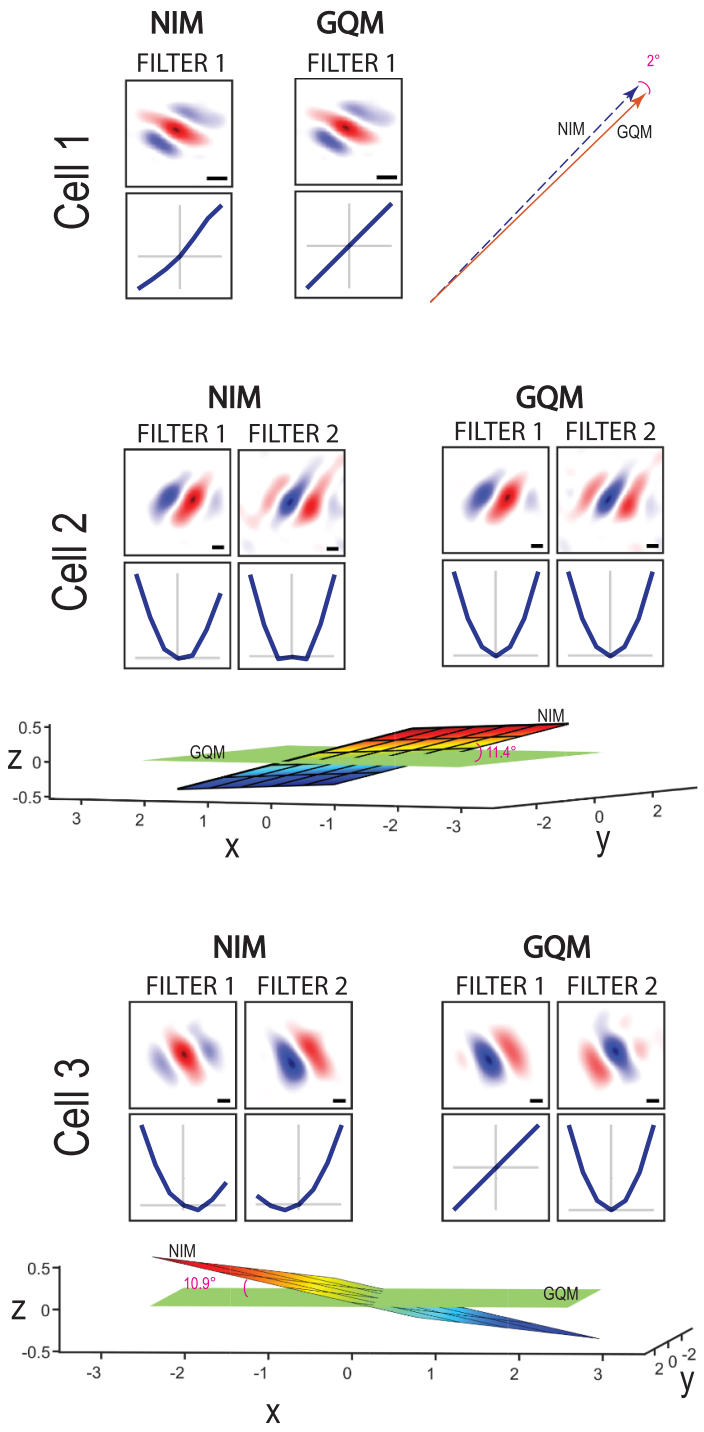

Figure 2. Three example V1 units to showcase the GQM and NIM fits. For *Unit 1*, the model fits are given in the right panel. The left panel visualizes the NIM and GQM feature space, as 1D vectors, in relation to each other. For *Units 2 & 3*, the model fits are given in the upper panels. The RF filters are presented above, and their corresponding input functions are given below. The 2D planes in the bottom panels depict the relation between the GQM and NIM feature spaces in a 3D subspace. The angle between vectors or planes representing feature spaces denotes the dihedral angle (see the main text). The spiking nonlinearity of the fits are not shown as it has a parametric form of log-exponential function. The black bars superimposed on RF filters indicate 1° of the visual field.

equally to opposing contrast polarities of the feature. This characterization is reminiscent of the classical definition of a complex cell in accordance with the energy model (Adelson and Bergen 1985), which describes fine selectivity to orientation and spatial frequency, but invariance to the spatial phase of the input stimulus.

Unit 3 is an example that does not fit the classical simple or complex cell dichotomy. The GQM defined the RF of *Unit 3* as comprising two filters, associated with one linear and one quadratic input function. While the NIM also identified two filters for this unit, they differed from the GQM filters; further the input functions also differed, possessing

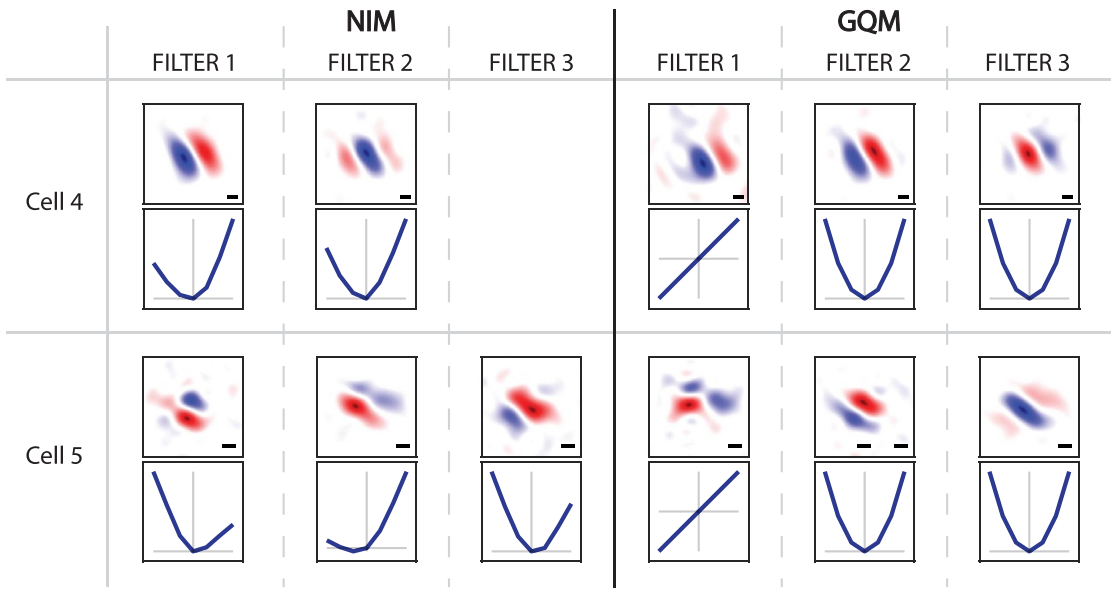

Figure 3. Two additional V1 example units that resulted in model fits with more than two RF filters. For *Unit 4* (top row) the NIM has found two RF filters while the GQM fitted the same unit with three RF filters. For *Unit 5* (bottom row) the two model fits matched in the number of RF filters, but there are apparent differences between the estimated input functions in the NIM and the linear/quadratic types included in the GQM. The black bars superimposed on RF filters indicate 1° of the visual field.

approximately threshold-linear input functions that are responding mostly to only one polarity of the feature. This indicates that the unit only shows partial invariance to the spatial phase of the stimulus (see also Almasi *et al* 2020).

Unit 4 is an example wherein the RFs characterized by the two models do not match in the number of filters. With the GQM, the unit was characterized with three filters, one linear and two quadratic, whereas with NIM only two filters with quadratic but unequal selectivity for opposite polarities of feature-contrast were sufficient to describe the RF of the same unit. As we will explain in section 3.2, the extra filter characterized by the GQM may be a compensation for the mismatch in RF nonlinearities from linear/quadratic types, and may not represent a true increase in the dimensionality of the feature space to which the unit responds.

Lastly, *Unit 5* is an example wherein both the NIM and the GQM characterized the unit with three filters. However, the input functions between the two models were very different.

The main difference between the NIM and the GQM is the constraints on their input functions, and the examples in figures 2 and 3 show the range of RFs characterized by the two models.

3.1. Defining the feature space of a cell model

While the last three examples in figures 2 and 3 appear to show significant differences between the NIM and GQM characterizations, both in terms of filters and nonlinearities, more care is required in understanding the structure of the models before this can be concluded. For instance, although the spatial structure of the filters might be different as seen, some characteristics of the RF filters, such as peak orientation and peak spatial frequency, are the same between the two models. More generally the filters define a space of features to which the unit responds. Rather than compare the models using just the filters, it is important to compare the feature spaces of the two models.

The feature space is determined by the first stage of processing in both models, which is linear filtering of the input image using RF filters. Mathematically, the input image is described using a vector whose dimension is equal to the number of pixels. The act of filtering the input image using the d RF filters is then equivalent to projecting the input image into a much lower dimensional vector space spanned by the RF filters (of dimension $\leqslant d$), which we will refer to as the unit's feature space. The coordinates of a point in this space are given by the outputs of each filter; these correspond to the contrast levels in image of the spatial features that match each RF filter. Thus, the feature space of a unit contains all possible linear weighted sums (i.e. contrast levels) of features corresponding to its RF filters. The uncovered feature space of each unit indicates the set of features in the input image to which that unit shows sensitivity in its response. Therefore, to compare the feature sensitivity of RFs of the same unit uncovered using different models, we should compare the resultant feature spaces spanned by the uncovered RF filters of each model.

Intuitively and generally, the feature space corresponds to either a one-dimensional (1D) line, 2D plane, or higher dimensional hyperplane in cases of more than two filters. The comparison between uncovered feature spaces of the two models can be

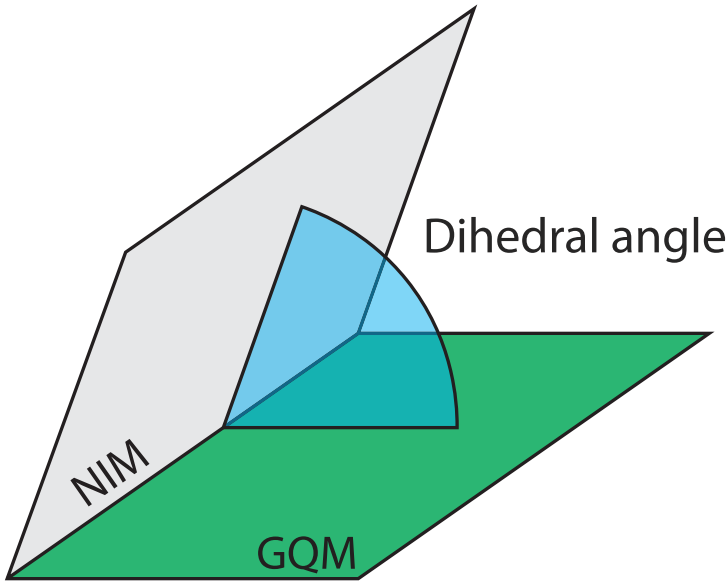

Figure 4. Illustration of the dihedral angle between intersecting planes. The similarity between the identified NIM and GQM feature spaces is measured using the dihedral angle between hyperplanes corresponding to the feature spaces. This angle can be easily generalized to hyperplanes in higher than 3D spaces.

done geometrically and by measuring how the corresponding planes (or hyper-planes) relate to each other in the higher dimensional input image space. We determine the similarity between the NIM and the GQM feature spaces using the dihedral angle between the planes (or hyper-planes) representing feature spaces, which is the angle between two intersecting planes (figure 4). For example, the NIM and the GQM feature spaces of *Unit 1* in figure 2 correspond to a line since both models contain only a single RF filter. The blue and orange vectors in figure 2 (right-most panel) for *Unit 1* indicate the feature spaces for the NIM and the GQM, respectively. As expected from the high similarity between the RF filters uncovered by the NIM and GQM, the lines indicating feature spaces almost align with only a 2° angle between them.

For *Units 2 & 3* (figure 2), because the estimated RFs of the two models both have two filters, the NIM and the GQM feature spaces correspond to 2D planes similar to the illustration given in figure 4. The relation between the feature spaces for *Units 2 & 3* are visualized below the RFs in figure 2 (bottom panels). In this representation, the GQM feature space corresponds to the x–y plane (green). The NIM feature spaces are also 2D planes, however, they deviate from the GQM (x–y) plane in the z axis due to the differences between the RF filters of the two models. An increase or decrease in the elevation (z-axis) of the NIM feature spaces are color coded. For *Unit 2*, as expected based on the high similarity between the uncovered RF filters of the two models, there is a small dihedral angle of 11.4° between the planes constituting feature spaces. The similarity between the uncovered RF filters of the NIM and the GQM is less evident for *Unit 3*. However, when comparing their corresponding feature spaces in figure 2 (bottom panel), the similarity is again very high, with only 10.9° dihedral angle between the planes describing feature spaces. Therefore, comparing feature spaces alleviates any need for comparing individual RF filters between models and provides a more holistic comparison. Population statistics comparing feature spaces for all V1 units will be given in section 3.3.

3.2. Feature space defined by linear and quadratic RF filters of the GQM

Using the GQM fits, 72% of our V1 units are described using a single linear spatial filter, 21% as a combination of linear and quadratic filters, and 7% with only quadratic filters (one or more). The uncovered linear filter in the GQM is equivalent to the filter that is obtained using the STA analysis (Park and Pillow 2011a).

In the GQM fitting procedure, there is no constraint imposed on the linear filter dimension in relation to the quadratic filters, so the linear filter is not necessarily orthogonal to the quadratic feature space, and in principle may even be identical to a quadratic filter. Consequently, the dimensionality of the GQM feature space may be less than the total number of filters. For units whose GQM-uncovered RFs are described using a combination of linear and quadratic filters, the identified linear filter dimension with respect to the quadratic feature space will in general fall into one of the three categories illustrated in figures 5(a)–(c). (i) The linear filter dimension is completely orthogonal to the quadratic feature space (figure 5(a)). In this case, the overall dimensionality of the GQM feature space comprises the dimension

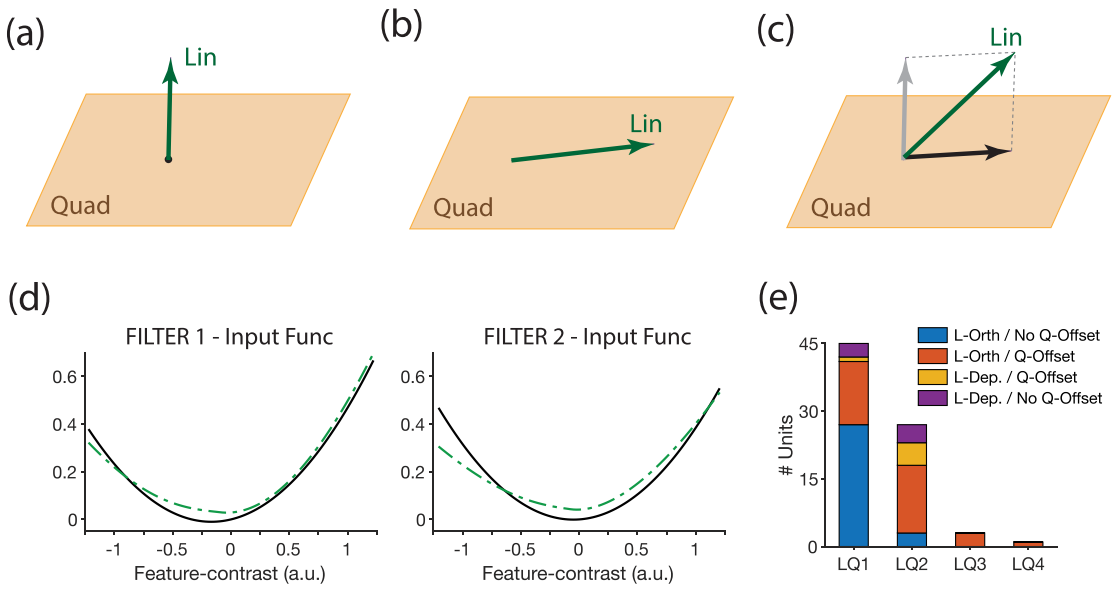

Figure 5. (a)–(c) Relationship between the linear and quadratic filters in the RF uncovered using the GQM. (d) Input functions of *Unit 4* (figure 3) estimated using the NIM (dashed green curves) are compared with the offset quadratic input functions of the GQM (solid black curves), due to the present of a dependent linear filter in the GQM fit. Input functions are plotted separately for each RF filter, and are applied on the filters' outputs, i.e. feature-contrasts. (e) Bar graph gives the distribution of different GQM fits according to the combination of linear (*L*) and quadratic filters (*Q*). The subsequent number after letter '*Q*' denotes the number of quadratic filters in the fit. The distribution is further split into three fit categories depending on the relationship between the linear filter and the quadratic feature space of the fit: (*i*) the linear filter was orthogonal to the quadratic feature space (blue), (*ii*) the linear filter had a component orthogonal to and a component in the quadratic feature space (orange), (*iii*) the linear filter was found to be sitting in the quadratic feature space or was found to be entirely non-significant (yellow).

of the quadratic feature space plus an additional linear dimension. (ii) The linear filter resides completely in the quadratic feature space and is thus a linear combination of the quadratic RF filters (figure 5(b)). In this case, the overall dimensionality of the GQM feature space is equal to the dimensionality of the quadratic feature space. (iii) The linear filter dimension is neither perpendicular to the quadratic feature space, nor does it sit within the quadratic feature space (figure 5(c)). In this case, the linear filter is decomposed into two components; one that is orthogonal to the quadratic feature space (light-colored vector; figure 5(c)), and another component that resides within the quadratic feature space, which is referred to as the linear-projected component (dark-colored vector; figure 5(c)). Like the first category, the dimensionality of the GQM feature space in this case is equal to the dimensionality of the quadratic feature space plus an additional linear residual dimension. The impact of a non-zero linear-projected filter component (figures 5(b) and (c)) is to impose an offset from zero feature-contrast in the unit's response function across any quadratic filter dimension that shows non-zero correlation with the linear-projected component (see methods). In other words, in this case the linear-projected component is a consequence of the nonlinear dependency of a unit's response on a filter dimension, manifested as a linear filter to correct for approximating the response nonlinearity by a set of quadratic functions. Figure 5(d) demonstrates this notion for *Unit 4* (in figure 3) in which the black curves indicate the quadratic input functions offset from zero due to a linear-projected filter dimension. For this unit, as seen, there is a good match between the offset quadratic functions and the input functions identified using the NIM (dashed green curves; figure 5(d)).

Further interrogation of the GQM fits on our V1 data separated the fits with a combination of linear and quadratic RF filters into the three categories described above (figure 5(e)). This explains the inconsistencies in RF estimation and the number of uncovered RF filters by the GQM and the NIM, which is presented in figure 6(a).

3.3. Feature space comparison between models

As seen in the model fit examples in figures 2 and 3, the RFs estimated by the NIM and the GQM may differ in the number of filters for the same unit. Figure 6(a) presents a summary of the population of our V1 data ($n = 342$) based on the number of identified filters using the NIM (black bars) and the GQM (white bars). Both models had three quarters of the population estimated as single-feature RFs (1 filter: NIM = 75%, GQM = 77%) and a quarter as multi-featured (>1 filter: NIM = 25%, GQM = 23%). It should be noted that in reporting the number of uncovered RF filters using GQM, we have corrected

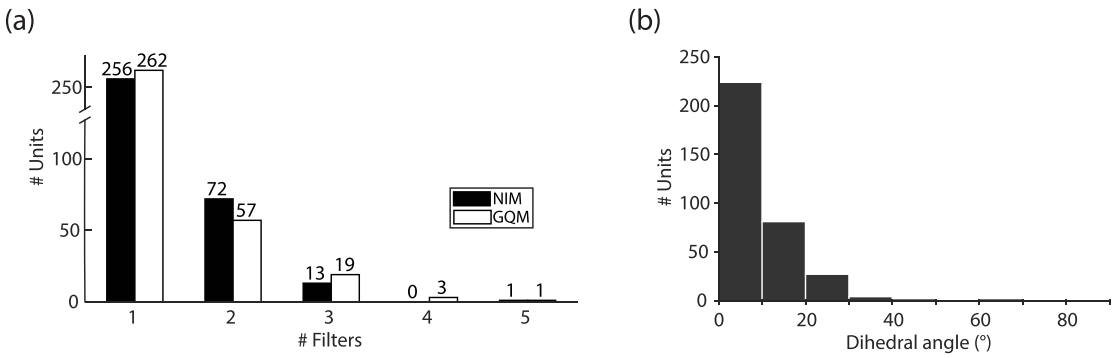

Figure 6. (a) Bar graph presents the distribution of units according to the number of RF filters in their NIM (black bars) and GQM (white bars) fits across our entire V1 unit population. (b) Histogram gives the distribution of the dihedral angle between the identified NIM and GQM feature spaces for our V1 unit population.

for the dependent and non-significant linear filter dimensions. That is, the linear filter is only counted if it contains a significant independent feature dimension from the quadratic filters, as described and discussed in the previous section.

Single- and multi-featured RFs have been closely related to simple and complex cells, respectively (Touryan *et al* 2005). Altogether, most units resulted in an equal number of filters from both model fits (91%, $n = 310/342$). Out of the units producing non-equal number of filters, 14 had the NIM producing more filters while 18 had the GQM producing more filters. This shows that the NIM and the GQM estimated the same number of filters in the great majority of V1 units.

Despite the similarities in the number of filters produced by the NIM and the GQM for each unit, the space spanned by the filters (i.e. space that contains all linear combinations of the RF filters) could be different. As explained in section 3.1, the similarity (or difference) between the NIM and the GQM feature spaces is measured using the dihedral angle between the planes (or hyper-planes) representing feature spaces. The dihedral angle between spaces that are completely overlapped (i.e. parallel) is 0°, while a dihedral angle of 90° indicates no overlap between the spaces, meaning they are orthogonal to each other. A complete overlap between feature spaces also includes cases where one feature space is a proper subspace of the other. Figure 6(b) shows the distribution of the dihedral angle for all units in our V1 data. Nearly all units had an angle of less than 30° (97%, $n = 332/342$). This shows that, in terms of the dihedral angle between feature spaces, there is minimal difference between the RF filters produced by the NIM and the GQM. For all units, the difference in dihedral angle between the feature spaces was found to be significantly different from that found for randomly generated feature spaces with similar dimensionalities (p-value $< 6 \times 10^{-4}$; see methods).

Due to the strong similarity between the uncovered NIM and GQM feature spaces for most units, the mapping between the NIM and GQM RF filters can be given by a linear invertible transformation for these units.

3.4. RF nonlinearities

In the previous section we saw that, irrespective of some differences in the RF filters, the uncovered feature spaces using the NIM and the GQM are very similar to each other. The major difference between the NIM and the GQM fits, however, lies in the input functions imposed on the filter dimensions. The NIM searches for filter dimensions that show maximal additive separability without any constraints on the input nonlinearities (Almasi *et al* 2020), while the GQM finds filter dimensions upon which the unit's response exhibits maximal linear or quadratic dependencies. Here, additive separability means that a spike rate that is in principle a completely general function of m different feature dimensions can be written as a function of a sum of separate 1D nonlinearities applied to each dimension: $\eta = \tilde{F}(\mathbf{k}_1 \cdot \mathbf{s}, \mathbf{k}_2 \cdot \mathbf{s}, \ldots, \mathbf{k}_m \cdot \mathbf{s}) = F(\sum_{i=1}^{m} g_i(\mathbf{k}_i \cdot \mathbf{s}))$, as per equation (1).

Figures 7(a) and (b) present 2D FCR functions, respectively, for *Units 3 & 4* whose model fits were previously discussed (see figures 2 and 3). For each unit, its FCR function is the function, $\tilde{F}$, mapping the RF filters' outputs, i.e. feature contrasts $c_i = \mathbf{k}_i \cdot \mathbf{s}$, to the unit's responses, η. For these units it is given as a 2D histogram and is plotted against the unit's feature space. The FCR histograms are computed by binning the empirical responses of each unit to each stimulus in terms of the pair of feature contrasts present in the image to the two filters. The grayscale indicates spike rate. The horizontal and vertical axes of the feature space correspond to the GQM-uncovered filters (linear/quadratic filters). The green- and orange-colored arrows denote the NIM-uncovered filter dimensions with respect to the GQM feature space. The magenta and cyan contour plots superimposed on the feature

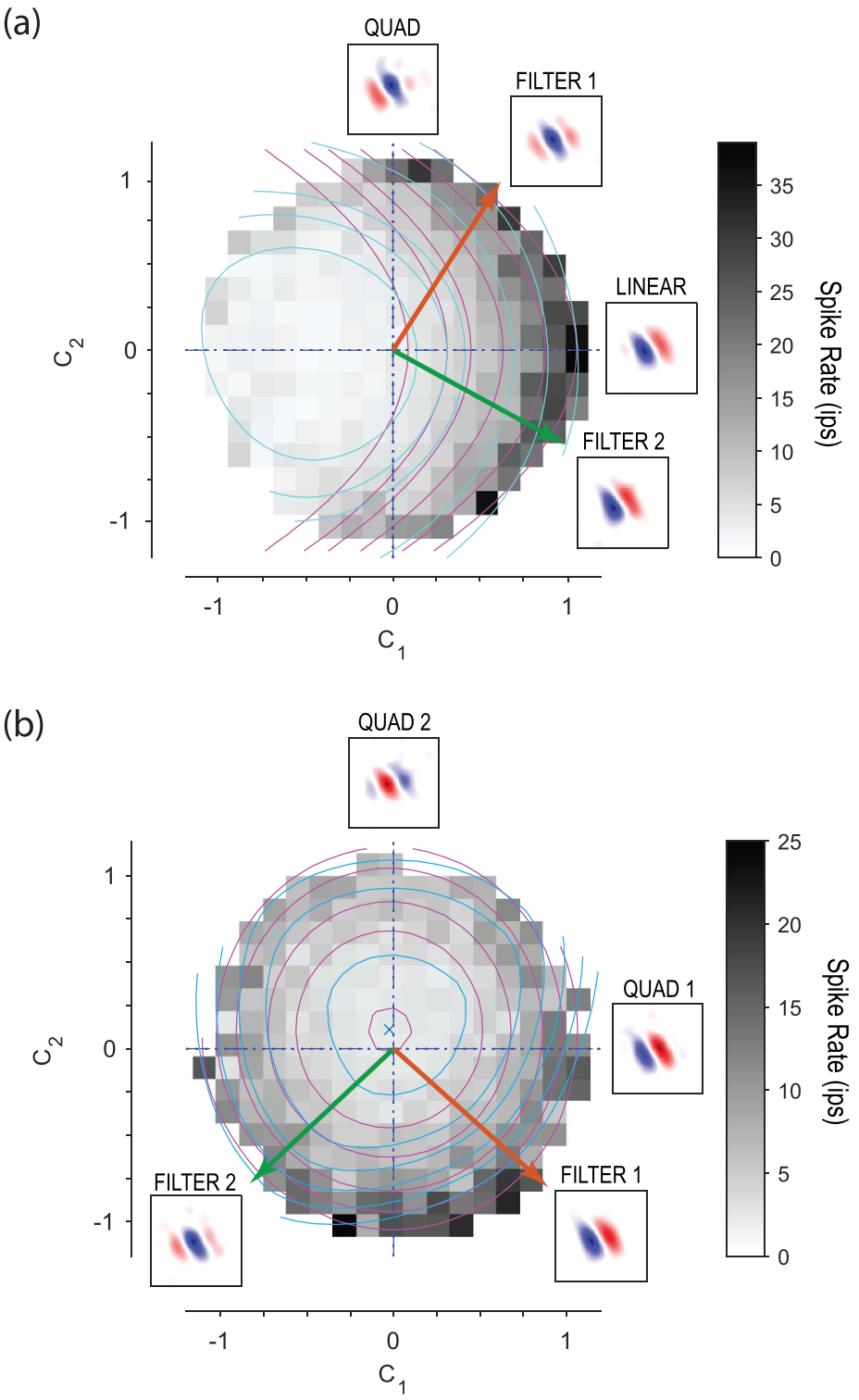

Figure 7. Comparison of the feature-contrast response (FCR) functions that emerge from model predictions over the unit feature space. (a) FCR function of *Unit 3* whose model fits were presented previously in figure 2, (b) FCR function of *Unit 4* whose model fits were presented in figure 3. For each unit, the FCR is demonstrated as a function of feature-contrasts of the presented RF filters. The grey pixels show the empirically calculated mean responses of each unit across the WGN stimuli in the unit's feature space. Color-bars denote the spike rate intensity for each unit. Only for visualization purposes here, the NIM feature space is projected onto the GQM feature space, hence, both are represented in a 2D subspace. The NIM (FILTER 1 & 2) and the GQM (LINEAR & QUAD) RF filters are shown as vectors within this subspace that indicate relevant feature dimensions for each fit. The superimposed magenta and cyan iso-response contours show equal spike rates that are predicted using the GQM and the NIM fits, respectively. The '×' symbol in (b) indicates the shift in the minimum of the FCR function predicted by the GQM due to the offset quadratic input functions as a consequence of the dependent linear filter. C_1 and C_2 denote feature-contrast (arbitrary units) of the linear and quadratic RF filters in (a), and feature-contrasts of the first and the second quadratic filters in (b). Feature-contrasts can be mapped between the feature spaces using the transformation introduced in the methods.

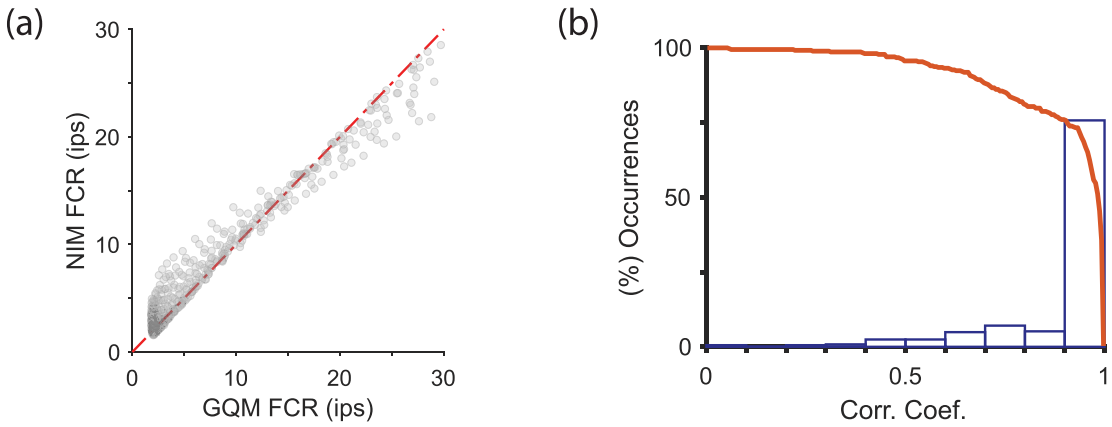

Figure 8. Quantitative comparison between FCR functions predicted by model fits for the same unit. (a) Scatter plot shows FCRs predicted by the GQM (horizontal axis) versus FCRs predicted by the NIM (vertical axis) for points sampled uniformly from the feature space of *Unit 3*. FCRs are given in impulses per second (ips). Red dashed line indicates identity. (b) Blue bar graph gives the distribution of the correlation coefficient between predicted FCRs using the NIM and the GQM fits for the same unit, across the entire V1 population. The ordinate of the orange curve gives the proportion of units in our V1 population which showed a correlation coefficient between their predicted NIM and GQM FCRs that was not less than the given value on the abscissa.

space indicate the different response level sets predicted using the GQM and the NIM, respectively. We can see that, although the response level sets predicted using the NIM better matched the empirical response distributions, the two models reasonably match in predicting the response patterns qualitatively. For *Unit 3* (figure 7(a)), the GQM and the NIM filters are rotated versions of each other (i.e. the NIM filter directions are approximately orthogonal and so can be interpreted as a rotation of the GQM axes defined by the GQM filters). Note that although for this unit the NIM RF filters are almost orthogonal to each other, this is not necessarily the case. Also, it is worth noticing that the linear and quadratic input functions of the GQM cannot capture the increase in the empirical response that occurs when C_1 becomes increasingly negative (the orange contours do not accurately track the greyscale histogram for $C_1 < 0$). In contrast, the input functions estimated by the NIM can successfully predict this pattern in the response (notice the closed cyan contour plot).

For *Unit 4* (figure 7(b)), the GQM and the NIM RF filters are similarly rotated versions of each other. The empirical FCR function is lopsided; the response is stronger on the negative side of C_2 than the positive side. This skewness in the FCR function is captured using a dependent linear filter in the GQM fit, which causes a shift in the minimum of the FCR function from the zero feature-contrast that is shown using the 'x' symbol. The NIM, however, can easily capture this response behavior using input functions that have unequal selectivity for the opposite feature-contrasts. Hence, in this case, the NIM provides a more compact representation of the neuron using less RF filters.

Direct comparison of the predicted FCR functions between models can be challenging as these functions are imposed on feature spaces, and feature spaces uncovered by models are not perfectly identical. However, we can compare the predicted FCR functions on a shared feature space. This shared feature space that we have considered is the NIM feature space projected onto the GQM feature space (see methods). Provided the dihedral angle between the two features spaces is small or insignificant, this is a very good approximation.

To compare FCR functions across the shared feature space, we start by sampling a uniform grid of points in the shared feature space within the feature-contrast range covered by the WGN stimuli. This results in the sampling of the feature space similar to the ones presented in figures 7(a) and (b). This approach amounts to sampling the feature space in Cartesian coordinates. Each grid point in these coordinates corresponds to a linearly combined set of RF filters whose combination is determined by the coordinates of that point. For each point in this space we calculate the value of the FCR function for the NIM and GQM and compare them. For *Unit 3* presented in figure 7(a), the correlation between predicted FCRs for NIM versus GQM is given in figure 8(a). For this unit, as we see, there is a very high correlation between the models in predicted FCRs, with a correlation coefficient (r) that is equal to 0.98.

We performed this systematic comparison between the predicted FCRs of the two models for all units in our V1 data. Figure 8(b) presents the results of this comparison by showing the distribution of correlation coefficients for all units using the blue bar graph. The orange superimposed graph shows the corresponding reversed cumulative distribution: i.e. at each given correlation coefficient indicates the percentage of V1 data that resulted in a correlation

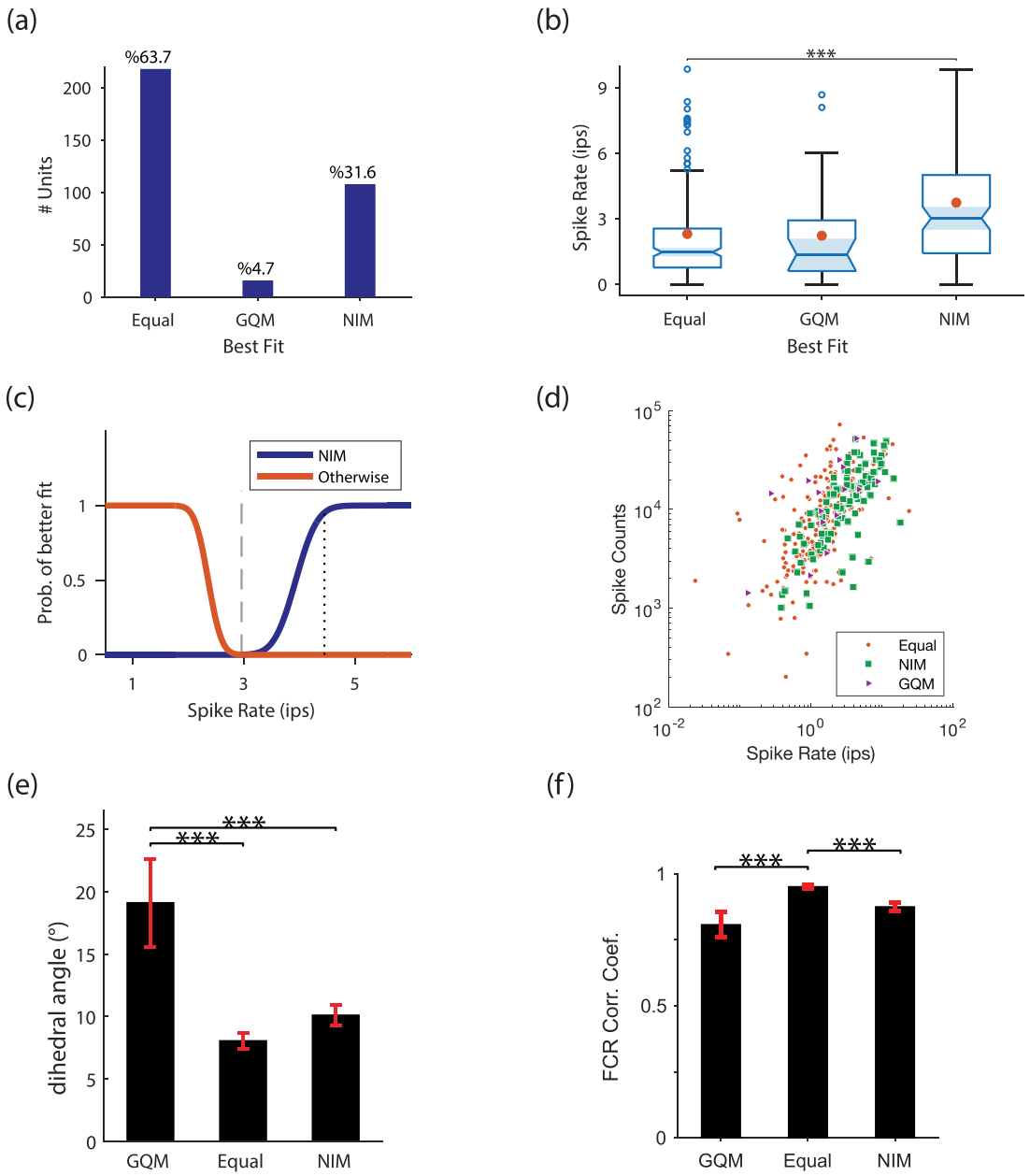

Figure 9. Quantitative model performance comparison. (a) The bar graph gives distribution of the units for which the NIM outperformed the GQM (NIM), the GQM outperformed the NIM (GQM), or both models gave statistically similar fits (Equal). (b) The boxplot presents distributions of unit spike rates according to the goodness of model fits. (c) The dark-blue graph shows the increase in the likelihood of the NIM to give a better fit for a unit as a function of the spike rate of the unit. The opposite scenario is given by the orange curve. (d) The scatter plot shows spike rate versus the number of collected spikes for each unit in our V1 population, separated according to the superior model fit. (e) Bar graphs and error bars show means and standard errors of the mean for the dihedral angle between identified NIM and GQM feature spaces, separated according to the better model fit, across our V1 population. (f) Distributions of the correlation coefficient between predicted FCRs using the NIM and the GQM, presented in a similar way to (e). (***: p-value < 0.01).

coefficient between NIM and GQM FCRs that was equal or greater than that given correlation coefficient. According to this graph, for instance, 79% of V1 units showed a correlation coefficient of 0.9 or higher between their predicted NIM and GQM FCRs. This indicates a strong similarity between the overall predicted FCR functions of the NIM and the GQM for most V1 units, despite differences between input functions of the two models.

3.5. Model performance comparison

We compared the predictive ability of the NIM and the GQM by measuring their log-likelihood on a held-out WGN dataset. The result of such a comparison is shown in figure 9(a), according to which for 32% ($n = 108$) of units in our V1 population the NIM significantly outperforms the GQM. In contrast, for only 5% ($n = 16$) of units did the GQM results give a significantly better fit to the data. For the remaining

63% ($n = 218$) of units, NIM and GQM gave statistically similar predictions to the data. This shows NIM was either better or equal to GQM at estimating RFs in 95% of units.

We suspected that the units resulting in superior fits to the NIM resulted from an ability of the NIM to improve the fit in cases with access to more spiking data beyond a point at which the GQM performance began to saturate. The spike rates of units for which NIM outperformed GQM were indeed significantly higher than units with equal performance (t-test; p-value $= 10^{-5}$) (figure 9(b)). In other words, the probability of getting a better fit to data using the NIM increased as a function of spike rate (averaged across all stimuli presented to the neuron). Our statistical analysis shows that for a spike rate of ∼3 ips, the probability of getting better fits to neural data with the NIM exceeds the probability of getting better or comparable fits with the GQM (figure 9(c); dashed-gray line). Further, for spike rates higher than 4.4 ips, the probability of getting a better fit to data with the NIM than the GQM goes beyond 0.95 (figure 9(c); black-dotted line). Although RF characterization depends on spike rate, figure 9(d) shows that, unsurprisingly, spike rate is also correlated with the number of collected spikes.

For the 5% of units where the GQM surpassed the NIM in predictive ability, we found a significant difference between the distribution of dihedral angles between feature spaces compared to units for which the NIM performed better than or equal to the GQM (figure 9(e); p-value < 0.01). According to figure 9(e), for units in which the GQM outperformed the NIM, the difference between the feature spaces tends to be significantly higher than for the rest of the units. Therefore, for units in which the GQM performed better, the most distinguishing factor between the model fits seems to be the estimate of the unit's feature space. In this case, the inferior performance of the NIM is most likely due to its struggle in finding an accurate estimate of the true feature space of the unit, which may be due to a lack of spikes. The NIM's struggle to find an accurate estimate of the feature space for such units is most likely attributed to its complexity compared to the GQM, which demands more neural data for fitting (see discussion).

We also studied model performance against the differences between RF nonlinearities captured using the model predicted FCR functions. For this, we looked at the distribution of the correlation coefficient between predicted FCRs of NIM and GQM against model performance (figure 9(e)). We noticed that for units for which NIM and GQM performed equally, the similarity between the predicted FCRs using the two models is significantly higher than when the NIM or GQM outperformed the other (p-value $= 6.5 \times 10^{-6}$).

4. Discussion

In this study, we compared two of the most robust RF characterization techniques based on maximum likelihood estimation: the GQM and the NIM. The NIM gives a general representation for a neuron using a cascaded structure of filters and nonlinear functions.

The most prominent distinction between the GQM and the NIM lies in the type of input functions imposed on the filters' outputs; in the GQM, input functions are always of linear or quadratic types, while in the NIM they can be of any form, which is determined by the neural data. A second difference between the two models is in the representation of the RF filters. The GQM does not provide a unique set of RF filters for the neuron. Any rotation (with scaling) of the quadratic filters tends to provide the same description for the neuron. In fact, for the GQM to find a unique set of RF filters, one needs to impose additional constraints on the quadratic filters, e.g. orthogonality. In this case, SVD of the quadratic feature space leads to a unique set of quadratic filters given that the weights by which quadratic filters contribute to the neuronal response are not identical. In the NIM, however, maximizing the likelihood of the data under the assumption of additively separable input streams can find a unique set of RF filters (Almasi *et al* 2020).

RF estimation comparisons between the GQM and the NIM of V1 units have been done before when McFarland *et al* (2013) first introduced the NIM framework. However, their comparisons were only for two recorded monkey V1 neurons while ours comprise 342 neurons from cat V1, allowing us to compare the two models with greater statistical power. It is also important to note that the NIM framework used in this study differs from the original model introduced by McFarland *et al* (2013) in the estimation procedure. Originally, a sequential optimization over the set of parameters was proposed to estimate the model, which is often in the order of filters, input functions, and spiking nonlinearity. However, the objective function to solve the optimization problem over the whole set of parameter space is highly nonlinear and not necessarily convex. Therefore, sequentially optimizing a set of parameters (e.g. filters) while fixing all other parameters in the model is highly susceptible of getting trapped in local optima, thereby not finding the optimal combination of model parameters to give the optimal solution. We rectified this issue by performing a simultaneous optimization over the set of all model parameters, and

we do an exhaustive search of the parameter space to maximize the chance of finding the global optimum (see Almasi *et al* 2020).

4.1. Biological plausibility and functional implications

One major strength of the NIM framework over the GQM is that it allows nonlinear input functions that may depart from quadratic functions. This can allow more biologically plausible input functions that have a threshold below which response is suppressed, rather than quadratics that give excitation for both positive and negative inputs. This can be informative to provide better insights about the functional neural circuity and the underlying architecture of sensory processing (Kaardal *et al* 2013). Through this, the NIM can potentially reveal presynaptic structures to neurons, mediated by thresholded input functions that represent synaptic inputs that integrate additively. Figure 10 presents such an example in which the NIM has found two non-oriented, but slightly spatially displaced filters with alike input functions, whereas the RF obtained using the GQM returns a single filter with a relatively elongated spatial structure, which is seemingly the superposition of the NIM filters. Note here, that the models behave similarly because there is also a hidden spiking-nonlinearity that follows the summation of the output function(s) in each model. This explains why responses for negative values of the feature contrast (negative x-axis of input function) are suppressed in figure 10(b), just as they are explicitly by each input function in figure 10(a). In this case, the two non-oriented filters could reasonably be suggested to be LGN afferents feeding into this cortical cell. On the other hand, this type of interpretation must be treated cautiously, as most units in this study had only one or two filters (and at most 5 filters) demonstrating that model clearly does not capture the many thousands of synaptic inputs into a typical cortical V1 cell. It is likely that many filters represent the aggregate input of many presynaptic input neurons for both NIM and GQM.

Regarding the biological plausibility of the NIM compared to the GQM, one may question the permissibility of double-sided input functions in the NIM in the present study, which allows functions such as quadratics. However, these functions can form because of integrating different functions. In fact, any filter with a double-sided input function can be considered as two separate parallel streams which have the same filter but with opposite polarities and input functions that are threshold-linear and selective to only one polarity of the feature (figure 11). In the NIM framework this issue can be dealt with by confining the input functions to be monotone increasing. However, in our NIM framework we do not have this constraint and allowed the input functions to take on any form, including double-sided nonlinearities like squaring functions. This is encouraged by reducing the total number of parameters in the model by removing duplicate filters, which also avoids any redundancy in the model representation. Hence, we permit double-sided input functions in the NIM to facilitating the optimization process and reducing the chance of overfitting.

4.2. Quantitative performance

It is worth emphasizing that this study does not aim to determine whether the GQM and the NIM are better than each other. Rather it aims to show the disparity in their RF characterization. We compared the performance of the two models by calculating their log-likelihood on a hold-out WGN test set. For a majority of units (63%) the two models resulted in a statistically comparable predictive ability on the test set. For a small minority (5%) of units, the GQM gave a better performance than the NIM. The GQM is considered a subclass of the NIM, in which input functions are bound to linear or quadratic types. Hence, given that the GQM is a subclass of the NIM, in theory, it is expected to not give a better fit to neural data than the NIM. However, in practice this can happen. The NIM provides a more complex representation for a unit, with a greater number of parameters, than the GQM. Fitting a more complex model like the NIM to neural data can be accompanied with some issues. First, estimating the greater number of parameters of the NIM requires more neural data compared to the GQM. Neural data is often scarce and long duration recording may not be as physiologically stable as shorter duration recordings. Second, the risk of model overfit to data is higher in a more complex model like the NIM than the GQM, hence the fitting procedure demands certain measures to reduce the chance of overfitting. Indeed, our analysis shows that in the presence of more spiking data, better model fits are obtained using the NIM than the GQM (figures 9(b) and (c)). Nonetheless, figure 9(d) shows that even with total recorded spikes as low as 1000, the NIM is able to provide a good estimate of the RF. This is different to conventional RF characterization techniques, such as spike-triggered covariance, in which good RF estimation depends heavily on the number of collected spikes (Schwartz *et al* 2006).

4.3. Comparison to other approaches for estimating RF models

A variety of other models and methods have been developed to fit RF models to the spiking-data of neural responses to visual stimuli. These include approaches to fitting forms of LN–LN cascade models, similar in form to GQM or NIM. One of these is spike triggered. It was developed primarily in the context of fitting models of retinal ganglion cell responses to photoreceptor/bipolar cell subunit inputs, but it can be applied to V1 as well (Shah *et al* 2020). While the model filters are allowed to be general, the input

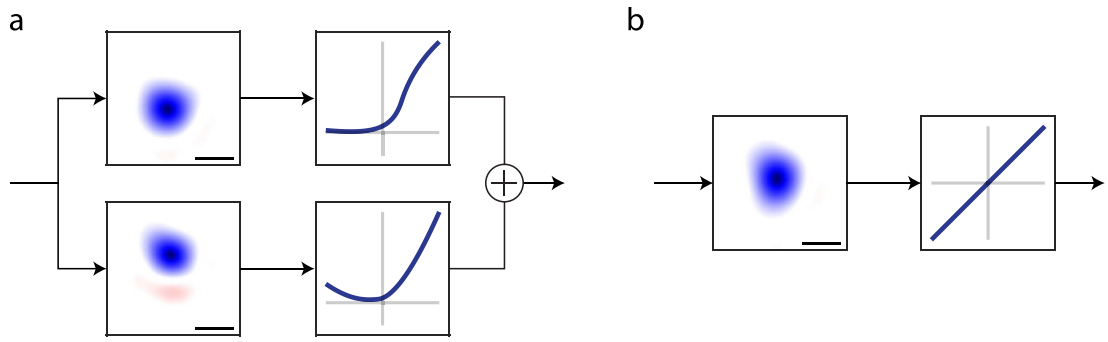

Figure 10. Comparison of the biological plausibility of the NIM and the GQM fits for an example V1 unit. The NIM and the GQM fits for this example unit are given in (a) and (b), respectively.

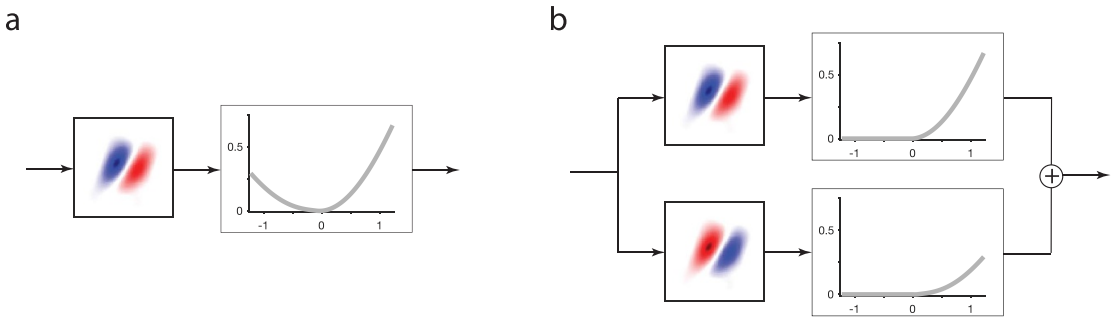

Figure 11. Decomposition of a double-sided input function in the NIM into two parallel streams with the same filter, but opposite polarities, with one-sided input functions.

functions for each filter are restricted to be exponentials, and the spiking nonlinearity has a particular saturating parametric form. Thus, similar to the GQM, this results in a model with a more restrictive family of nonlinearities applied to the underlying feature subspace of the cell compared to NIM. An advantage of assuming exponential forms for the input functions is that an efficient clustering algorithm can be used on a convex upper bound of the (negative) log-likelihood to robustly estimate the filters in the case where WGN stimuli are used (nonlinearities are estimated in a second stage). This is in contrast to the typically slower gradient descent method used here to estimate filters for NIM and GQM. Further clustering may lead to more accurate filter estimates, given limited data. The optimization of the number of filters is via a cross-validation procedure, similar to the process used here. In principle, the different method for filter estimation might result in a different feature subspace to either the NIM or GQM. However, if both algorithms find good approximations to the global optima, then differences between spike-triggered clustering and NIM may be small and may be similar in nature to the differences between GQM and NIM reported here.

Another approach that has been used to fit RF models with multiple subunit filters is spike-triggered nonnegative matrix factorization (STNMF). Like spike-triggered clustering, it was developed in the context of fitting models of retinal ganglion cell responses (Liu *et al* 2017). STNMF does not align formally with an LN–LN cascade. Instead, it obtains a set of subunit filters by first finding a set of non-negative basis functions (with the same dimensionality as the stimulus) that are weighted to reconstruct the set of spike-triggered stimuli. Both the basis functions and the weights are optimized to reconstruct these stimuli. The basis functions are then interpreted as the spatial filters of the model subunits and additional steps are taken to fit the nonlinearities of an LN–LN cascade. This involves assuming a threshold linear input function corresponding to each filter and fitting a spiking nonlinearity similar to the log-exponential function used here for NIM and GQM. The non-negativity constraint on the filters of STNMF would preclude obtaining filters with both negative and positive subfields that were commonly obtained with both GQM and NIM in the present study. Here, these subfields of opposite polarity are responsible for effectively capturing the combined effect of excitatory and inhibitory interactions between LGN neurons responding to ON (bright) and OFF (dark) parts of the image stimulus. Applying STNMF to data from such V1 neurons may result in separating different subfields into different subunit filters. In principle, this

could still result in the same feature space as NIM or GQM.

Methods for fitting RF models, such as NIM, GQM, spike triggered clustering and STNMF, can be improved by using an appropriate basis for representing RF filters. The basis reduces the dimensionality of the representation, and therefore the number of free parameters in the model. This can improve several aspects of performance including predictive power, computational time for fitting and the quantity of data required for fitting. Here we used a Fourier basis to represent filters that was chosen based on the RF size and prior knowledge about the typical spatial frequency bandwidth of V1 neurons. A more systematic approach based on a Fourier representation has demonstrated good improvements in computation efficiency and data requirements (Aoi and Pillow 2017). Another recent approach has used a natural cubic spline basis to represent filters, together with a systematic approach to determine the number of basis functions using cross-validation (Huang *et al* 2021). It was shown that this often led to good improvements in model predictions, particularly with relatively small data sets.

Another approach for improving model predictions, while reducing the quantity of data required for fitting, is to use some form of regularization to include information about prior expectations concerning the structure of RF filters, such as their sparsity (in terms of non-zero weights), spatial smoothness, or local nature. The NIM, spike-triggered clustering and STNMF all include an option for L1-norm regularization of sparse RFs. The degree of regularization can be optimized with cross-validation. In the case of the simpler general linear model, methods for automatic smoothness determination (Sahani and Linden 2002) and automatic locality determination (Park and Pillow 2011b) that allow a maximum *a posteriori* estimate have been devised. Here, we did not use any form of regularization for the filters since the use of the Fourier basis for representing the filters enforced an appropriate degree of spatial smoothing.

A further way of reducing the number of free parameters in RF models with LN–LN cascades is to assume a convolutional architecture, whereby the initial layer contains multiple copies of the same filter, systematically shifted in space relative to each other (Vintch *et al* 2015, Wu *et al* 2015). This can be justified in terms of the retinotopic organization of the visual pathway, whereby similar processing occurs across neighboring locations in the representation of the visual field. This simplification allows a more complex, nonlinear model to be well-fit with just limited data, compared to e.g. spike-triggered covariance. The convolutional architecture of the model leads to feature spaces that are much higher dimensionally than any of the LN–LN cascade models previously mentioned. Nonetheless, the model often remains readily interpretable because, typically, only a small number of filters are identified, and their convolutional action can be readily conceptualized. However, the convolutional architecture, while plausible, is a strong assumption that may lack accuracy in some cases.

Using a different paradigm for fitting, deep convolutional neural networks (DCNNs) have been used to predict responses of neurons in middle (V4) and later (infero-temporal) stages of the ventral visual pathway that is implicated in object recognition. In general, this involves further stages of processing beyond a two-stage LN–LN cascade, and is thus extremely challenging to fit due the complexity and number of parameters. Leveraging the success of DCNNs in machine learning for object classification tasks, this approach has demonstrated impressive results in predicting neural responses in these higher areas. The approach uses a DCNN trained for object recognition using supervised learning on a classified image database, such as ImageNet. Neural responses to visual stimuli are then predicted based on a simpler, often linear model, fit to the activity in an appropriate layer of the DCNN when presented with those stimuli (Yamins *et al* 2014, Kriegeskorte 2015, Yamins and DiCarlo 2016, Zhuang *et al* 2021). A limitation with this approach is that such multilayer neural networks remain difficult to interpret in terms of understanding the processing they perform. In contrast, a recent approach that incorporates divisive normalization in the RF model, allows for fitting more complex models that are nonetheless more interpretable (Burg *et al* 2021). Divisive normalization is a well-established principle of nonlinear computation in the visual system that results in a variety of types of contextual processing, including beyond the classical RF (Heeger 1992, Carandini and Heeger 2012, Sawada and Petrov 2017). However, both DCNNs and divisive normalization models are considerably more complex to interpret than the LN–LN cascade models considered here, including the NIM and GQM.

5. Conclusion

According to our quantitative comparison, the NIM provides a representation of neural RFs that is better than or equal to the GQM in 95% of cases. Furthermore, in line with the qualitative comparisons in the current study, the NIM has the ability of providing biological interpretations for data by potentially teasing apart tentative synaptic inputs to neurons. However, all the advantages of the NIM depend somewhat on spike frequency. The NIM requires more spiking data or a higher spiking rate due to its more complex structure. Therefore, in the presence of sufficient neural data, the NIM is undoubtedly the preferred model choice. In contrast, although it often cannot capture subtleties of the FCR function, the GQM is capable of describing the neural

data using a model that is as good as NIM in 67% of cases. Our analysis showed that, for a majority of units, there is a strong similarity between the identified feature spaces and their predicted FCR functions using the NIM and the GQM. Thus, if one wishes to only study the feature space of a unit but does not care about the RF nonlinearities and the nonlinear integration of the features, the GQM provides a computationally affordable choice.

Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This work was supported by the Australian Research Council Centre of Excellence for Integrative Brain Function (CE140100007), the National Health and Medical Research Council (GNT1106390), and Lions Club of Victoria.

ORCID iD

Hamish Meffin https://orcid.org/0000-0003-4307-6841

References

Adelson E H and Bergen J R 1985 Spatiotemporal energy models for the perception of motion *J. Opt. Soc. Am.* A **2** 284–99

Almasi A, Meffin H, Cloherty S L, Wong Y, Yunzab M and Ibbotson M R 2020 Mechanisms of feature selectivity and invariance in primary visual cortex *Cereb. Cortex* **30** 5067–87

Almasi A, Sun S H, Yunzab M, Jung Y J, Meffin H and Ibbotson M R 2022 How stimulus statistics affect the receptive fields of cells in primary visual cortex *J. Neurosci.* **42** 045003

Aoi M C and Pillow J W 2017 Scalable Bayesian inference for high-dimensional neural receptive fields *bioRxiv* p 212217

Bellman R 1957 *Dynamic programming* (Princeton University Press)

Berkes P and Wiskott L 2005 Slow feature analysis yields a rich repertoire of complex cell properties *J. Vis.* **5** 579–602

Burg M F, Cadena S A, Denfield G H, Walker E Y, Tolias A S, Bethge M and Ecker A S 2021 Learning divisive normalization in primary visual cortex *PLOS Comput. Biol.* **17** e1009028

Carandini M 2006 What simple and complex cells compute *J. Physiol.* **577** 463–6

Carandini M and Heeger D J 2012 Normalization as a canonical neural computation *Nat. Rev. Neurosci.* **13** 51–62

Chichilnisky E J 2001 A simple white noise analysis of neuronal light responses *Netw.: Comput. Neural Syst.* **12** 199–213

DeAngelis G C, Ohzawa I and Freeman R D 1993 Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation *J. Neurophysiol.* **69** 1118–35

Einhäuser W, Kayser C, König P and Körding K P 2002 Learning the invariance properties of complex cells from their responses to natural stimuli *Eur. J. Neurosci.* **15** 475–86

Emerson R C, Citron M C, Vaughn W J and Klein S A 1987 Nonlinear directionally selective subunits in complex cells of cat striate cortex *J. Neurophysiol.* **58** 33–65

Fitzgerald J D, Rowekamp R J, Sincich L C and Sharpee T O 2011 Second order dimensionality reduction using minimum and maximum mutual information models PLoS *Comput. Biol.* **7** e1002249

Hansard M and Horaud R 2011 A differential model of the complex cell *Neural Comput.* **23** 2324–57

Hastie T, Tibshirani R and Friedman J 2009 The elements of statistical learning: data mining, inference, and prediction 2nd edn (Springer-Verlag) (available at: www.springer.com/gp/book/9780387848570) (Accessed 18 November 2019)

Heeger D J 1992 Half-squaring in responses of cat striate cells *Vis. Neurosci.* **9** 427–43

Huang Z, Ran Y, Oesterle J, Euler T and Berens P 2021 Estimating smooth and sparse neural receptive fields with a flexible spline basis (arXiv:2108.07537)

Hubel D H and Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex *J. Physiol.* **160** 106–54

Kaardal J, Fitzgerald J D, Berry MJ II and Sharpee T O 2013 Identifying functional bases for multidimensional neural computations *Neural Comput.* **25** 1870–90

Körding K P, Kayser C, Einhäuser W and König P 2004 How are complex cell properties adapted to the statistics of natural stimuli? *J. Neurophysiol.* **91** 206–12

Kriegeskorte N 2015 Deep neural networks: a new framework for modeling biological vision and brain information processing *Annu. Rev. Vis. Sci.* **1** 417–46

Lian Y, Almasi A, Grayden D B, Kameneva T, Burkitt A N and Meffin H 2021 Learning receptive field properties of complex cells in V1 *PLOS Comput. Biol.* **17** e1007957

Lindeberg T 2020 Provably scale-covariant continuous hierarchical networks based on scale-normalized differential expressions coupled in cascade *J. Math. Imaging Vis.* **62** 120–48

Liu J K, Schreyer H M, Onken A, Rozenblit F, Khani M H, Krishnamoorthy V, Panzeri S and Gollisch T 2017 Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization *Nat. Commun.* **8** 149

McFarland J M, Cui Y and Butts D A 2013 Inferring nonlinear neuronal computation based on physiologically plausible inputs *PLOS Comput. Biol.* **9** e1003143

Merolla P and Boahn K 2004 A recurrent model of orientation maps with simple and complex cells *Advances in Neural Information Processing Systems (NIPS 2004)* pp 995–1002

Movshon J A, Thompson I D and Tolhurst D J 1978 Receptive field organization of complex cells in the cat's striate cortex *J. Physiol.* **283** 79–99

Pachitariu M, Steinmetz N, Kadir S, Carandini M and Harris K D 2016 Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels (available at: http://biorxiv.org/lookup/doi/10.1101/061481) (Accessed 16 July 2018)

Park I M, Archer E W, Priebe N and Pillow J W 2013 Spectral methods for neural characterization using generalized quadratic models *Advances in Neural Information Processing Systems* pp 2454–62

Park I M and Pillow J W 2011a Bayesian spike-triggered covariance analysis *Advances in Neural Information Processing Systems* vol 24, ed J Shawe-Taylor, R S Zemel, P L Bartlett, F Pereira and K Q Weinberger (Curran Associates, Inc.) pp 1692–700 (available at: http://papers.nips.cc/paper/4411-bayesian-spike-triggered-covariance-analysis.pdf) (Accessed 16 December 2018)

Park M and Pillow J W 2011b Receptive field inference with localized priors PLoS *Comput. Biol.* **7** e1002219

Rossant C *et al* 2016 Spike sorting for large, dense electrode arrays *Nat. Neurosci.* **19** 634–41

Rust N C, Schwartz O, Movshon J A and Simoncelli E P 2005 Spatiotemporal elements of macaque V1 receptive fields *Neuron* **46** 945–56

Sahani M and Linden J 2002 Evidence optimization techniques for estimating stimulus-response functions *Advances in Neural Information Processing Systems* vol **15**

Sawada T and Petrov A A 2017 The divisive normalization model of V1 neurons: a comprehensive comparison of physiological data and model predictions *J. Neurophysiol.* **118** 3051–91

Schwartz O, Pillow J W, Rust N C and Simoncelli E P 2006 Spike-triggered neural characterization *J. Vis.* **6** 13

Serre T and Riesenhuber M 2004 Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex *Technical Report AI Memo* 2004-017 (MIT Computer Science and Artificial Intelligence Laboratory)

Shah N P, Brackbill N, Rhoades C, Kling A, Goetz G, Litke A M, Sher A, Simoncelli E P and Chichilnisky E J 2020 Inference of nonlinear receptive field subunits with spike-triggered clustering *eLife* **9** e45743

Sun S H, Almasi A, Yunzab M, Zehra S, Hicks D G, Kameneva T, Ibbotson M R and Meffin H 2021 Analysis of extracellular spike waveforms and associated receptive fields of neurons in cat primary visual cortex *J. Physiol.* **599** 2211–38

Touryan J, Felsen G and Dan Y 2005 Spatial structure of complex cell receptive fields measured with natural images *Neuron* **45** 781–91

Touryan J, Lau B and Dan Y 2002 Isolation of relevant visual features from random stimuli for cortical complex cells *J. Neurosci.* **22** 10811–8

Vintch B, Movshon J A and Simoncelli E P 2015 A convolutional subunit model for neuronal responses in macaque V1 *J. Neurosci.* **35** 14829–41

Wu A, Park I M and Pillow J W 2015 Convolutional spike-triggered covariance analysis for neural subunit models *Advances in Neural Information Processing Systems* pp **28**

Yamins D L and DiCarlo J J 2016 Using goal-driven deep learning models to understand sensory cortex *Nat. Neurosci.* **19** 356–65

Yamins D L, Hong H, Cadieu C F, Solomon E A, Seibert D and DiCarlo J J 2014 Performance-optimized hierarchical models predict neural responses in higher visual cortex *Proc. Natl Acad. Sci.* **111** 8619–24

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M C, DiCarlo J J and Yamins D L 2021 Unsupervised neural network models of the ventral visual stream *Proc. Natl Acad. Sci.* **118** e2014196118