Linking normative models of natural tasks to descriptive models of neural response

Priyank Jaini
Cheriton School of Computer Science,
Waterloo, Ontario, Canada
Department of Psychology, University of Pennsylvania,
Philadelphia, PA, USA

Johannes Burge
Department of Psychology,
University of Pennsylvania, Philadelphia, PA, USA
Neuroscience Graduate Group,
University of Pennsylvania, Philadelphia, PA, USA
Bioengineering Graduate Group,
University of Pennsylvania, Philadelphia, PA, USA

Understanding how nervous systems exploit task-relevant properties of sensory stimuli to perform natural tasks is fundamental to the study of perceptual systems. However, there are few formal methods for determining which stimulus properties are most useful for a given natural task. As a consequence, it is difficult to develop principled models for how to compute task-relevant latent variables from natural signals, and it is difficult to evaluate descriptive models fit to neural response. Accuracy maximization analysis (AMA) is a recently developed Bayesian method for finding the optimal task-specific filters (receptive fields). Here, we introduce AMA–Gauss, a new faster form of AMA that incorporates the assumption that the class-conditional filter responses are Gaussian distributed. Then, we use AMA–Gauss to show that its assumptions are justified for two fundamental visual tasks: retinal speed estimation and binocular disparity estimation. Next, we show that AMA–Gauss has striking formal similarities to popular quadratic models of neural response: the energy model and the generalized quadratic model (GQM). Together, these developments deepen our understanding of why the energy model of neural response have proven useful, improve our ability to evaluate results from subunit model fits to neural data, and should help accelerate psychophysics and neuroscience research with natural stimuli.

Introduction

Perceptual systems capture and process sensory stimuli to obtain information about behaviorally relevant properties of the environment. Characterizing the features of sensory stimuli and the processing rules that nervous systems use is central to the study of perceptual systems. Most sensory stimuli are high-dimensional, but only a small set of stimulus features are relevant for any particular task. Thus, perceptual and neural processing in particular tasks is driven by sets of features that occupy a lower dimensional space (i.e., can be described more compactly) than the stimuli. These considerations have motivated perception and neuroscience researchers to develop methods for dimensionality reduction that characterize the statistical properties of proximal stimuli, that describe the responses of neurons to those stimuli, and that specify how those responses could be decoded (Bell & Sejnowski, 1997; Cook & Forzani, 2009; Cook, Forzani, & Yao, 2010; Hotelling, 1933; Lewicki, 2002; McFarland, Cui, & Butts, 2013; Olshausen & Field, 1996; Pagan, Simoncelli, & Rust, 2016; Park, Archer, Priebe, & Pillow, 2013; Ruderman & Bialek, 1994; Rust, Schwartz, Movshon, & Simoncelli, 2005; Schwartz, Pillow, Rust, & Simoncelli, 2006; Tipping & Bishop, 1999; Vintch, Movshon, & Simoncelli, 2015). However, many of these methods are task-independent; that is, they do not explicitly consider the sensory, perceptual, or behavioral tasks for which the encoded information will be used. Empirical studies in psychophysics and neuroscience often focus on the behavioral limits and neurophysiological underpinnings of performance in specific tasks. Thus, there is a partial disconnect between task-independent theories of encoding and common methodological practices in psychophysics, and sensory and systems neuroscience.

Citation: Jaini, P., & Burge, J. (2017). Linking normative models of natural tasks to descriptive models of neural response. *Journal of Vision*, *17*(12):16, 1–26, doi:10.1167/17.12.16.

doi: 10.1167/17.12.16 Received April 6, 2017; published October 25, 2017 ISSN 1534-7362

Task-specific normative models prescribe how best to perform a particular task. Task-specific normative models are useful because they provide principled hypotheses about (1) the stimulus features that nervous systems should encode and (2) the processing rules that nervous systems should use to decode the encoded information. Many normative models in widespread use are not directed at specific tasks. Methods for fitting neural response cannot generally be interpreted with respect to specific tasks. Accuracy maximization analysis (AMA) is a Bayesian method for finding the stimulus features that are most useful for specific tasks (Burge & Jaini, 2017; Geisler, Najemnik, & Ing, 2009). In conjunction with carefully calibrated natural stimulus databases, AMA has contributed to the development of normative models of several fundamental tasks in early- and mid-level vision (Burge & Geisler, 2011; Burge & Geisler, 2012; Burge & Geisler, 2014; Burge & Geisler, 2015), by determining the encoding filters (receptive fields) that support optimal performance in each task. These task-specific normative models have, in turn, predicted major aspects of primate neurophysiology and human psychophysical performance with natural and artificial stimuli (Burge & Geisler, 2014, 2015).

The primary theoretical contribution of this manuscript is to establish formal links between normative models of specific tasks and popular descriptive models of neural response (Figure 1). To do so, we first develop a new form of AMA called AMA–Gauss, which incorporates the assumption that the latent-variable-conditioned filter responses are Gaussian distributed. Then, we use AMA–Gauss to find the filters (receptive fields) and pooling rules that are optimal with natural stimuli for two fundamental tasks: estimating the speed of retinal image motion and estimating binocular disparity (Burge & Geisler, 2014, 2015). For these two tasks, we find that the critical assumption of AMA–Gauss is justified: the optimal filter responses to natural stimuli, conditioned on the latent variable (i.e., speed or disparity), are indeed Gaussian distributed. Then, we show that this empirical finding provides a normative explanation for why neurons that select for motion and disparity have been productively modeled with energy-model-like (i.e., quadratic) computations (Cumming & DeAngelis, 2001; DeAngelis, 2000; Ohzawa, 1998; Ohzawa, DeAngelis, & Freeman, 1990; Ohzawa, DeAngelis, & Freeman, 1997). Finally, we recognize and make explicit the formal similarities between AMA–Gauss and the generalized quadratic model (GQM) (Park et al., 2013; Wu, Park, & Pillow, 2015), a recently developed method for neural systems identification. These advances may help bridge the gap between empirical studies of psychophysical and neurophysiological tasks, methods for neural systems identification, and task-specific normative modeling (Figure 1).

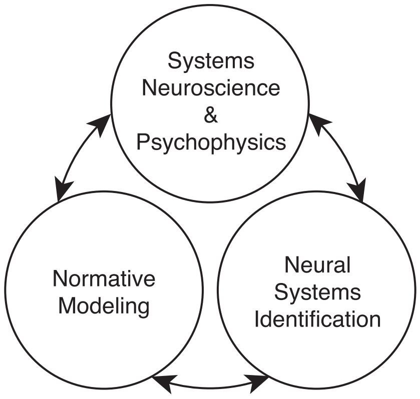
Systems Neuroscience & Psychophysics
Normative Modeling
Neural Systems Identification

Figure 1. Linking scientific subfields. Perception science benefits when links are drawn between psychophysical and neuroscience studies of particular tasks, task-agnostic statistical procedures that fit models to data, and task-specific normative methods that determine which models are best. The current work develops formal links between the energy model for describing neural response, the generalized quadratic model (GQM) for fitting neural response, and AMA–Gauss for determining the neural response properties that best serve a particular task.

In addition to these theoretical contributions, the development of AMA-Gauss represents a technical advance. The major drawback of AMA is its computational expense. Its compute-time for filter learning is quadratic in the number of stimuli in the training set, rendering the method impractical for large-scale problems without specialized computing resources. We demonstrate, both analytically and empirically, that AMA-Gauss reduces compute-time for filter learning from quadratic to linear. Thus, for tasks for which the critical assumption of AMA-Gauss is justified, AMA-Gauss can be of great practical benefit.

Background

Energy model

Energy models have been remarkably influential in visual neuroscience. The standard energy model posits two Gabor-shaped subunit receptive fields, the responses of which are squared and then summed (Figure 2A). These computations yield decreased sensitivity to the local position of stimulus features (i.e., spatial phase) and increased sensitivity to the task-relevant latent variable. Energy models have been widely used to describe the computations of neurons involved in coding retinal image motion and binocular disparity (Adelson & Bergen, 1985; Cumming & DeAngelis, 2001; DeAngelis, 2000). However, the motion–energy and disparity–energy computations are primarily de-

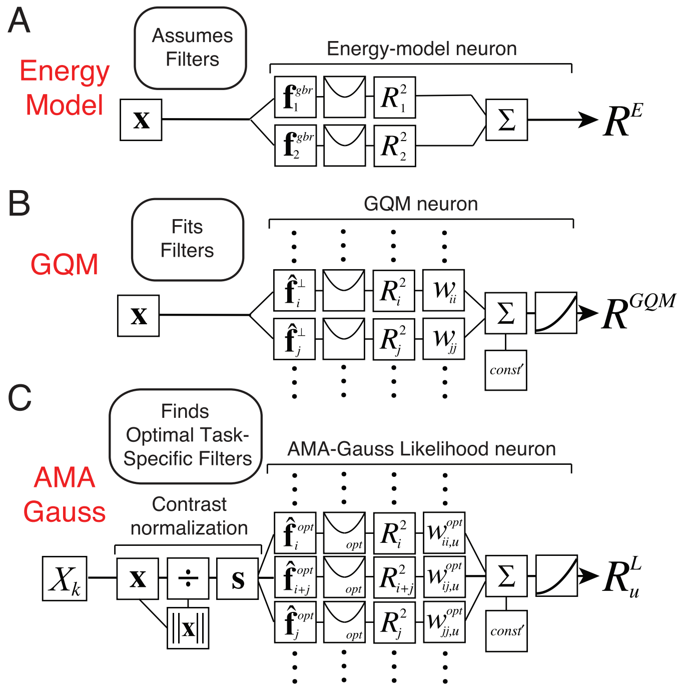

Figure 2. Computations of an energy model neuron, a GQM model neuron, and an AMA–Gauss likelihood neuron. All three have quadratic computations at their core. The energy model and the GQM describe the computations that neurons perform. AMA–Gauss prescribes the computations that neurons should perform to optimize performance in a specific task. (A) The standard energy model assumes two Gabor-shaped orthogonal subunit filters (receptive fields) $\mathbf{f}^{gbr}$ to account for a neuron's response. The response of an energy model neuron R^E is obtained by adding the squared responses of the filters. (B) The GQM fits multiple arbitrarily-shaped orthogonal subunit receptive fields $\mathbf{f}^{\perp}$ that best account for a neuron's response. The response of a GQM model neuron R^{GQM} is obtained by pooling the squared (and linear, not shown) responses of the subunit filters via a weighted sum, and passing the sum through an output nonlinearity. (C) AMA–Gauss finds the optimal subunit filters $\mathbf{f}^{opt}$ and quadratic pooling rules for a specific task. Unlike standard forms of the energy model and the GQM, AMA–Gauss incorporates contrast normalization and finds subunit filters that are not necessarily orthogonal. The response R_u^L of an AMA–Gauss likelihood neuron represents the likelihood of latent variable X_u. The likelihood is obtained by pooling the squared (and linear, not shown) subunit filter responses, indexed by i and j, via a weighted sum (Equation 20).

scriptive models of a neuron's response properties. The energy model does not make explicit how neural responses should be decoded into estimates.

Under what circumstances would energy-model-like computations be optimal? Energy-model-like computations are optimal if quadratic pooling is necessary for determining the likelihood of the task-relevant latent variable. We show in the following material that for retinal speed and binocular disparity estimation, two tasks classically associated with the energy model, quadratic pooling is indeed necessary to optimally decode the task-relevant latent variable (see Results). Therefore energy-model-like computations are optimal for these tasks with natural stimuli. AMA–Gauss is specifically designed to find the receptive fields and pooling rules that optimize performance under these conditions. It is thus likely to help accelerate the development of normative models of other tasks for which the energy model has provided a useful description.

Generalized quadratic model (GQM)

The standard energy model assumes that the responses of certain neurons can be accounted for by two Gabor-shaped subunit receptive fields. Real neurons are not constrained to have only two subunit receptive fields, nor are their shapes constrained to be Gabor-shaped. The generalized quadratic model (GQM) fits multiple arbitrarily-shaped subunit filters and quadratic pooling rules that best account for a neuron's response (Figure 2B; Park et al., 2013). The GQM is a specific example of a large class of models designed for neural systems identification, collectively known as "subunit models." The spike-triggered average (STA), spike-triggered covariance (STC), and the generalized linear model (GLM) are popular examples of this class of models. The goal of these models is to provide a computational level description of a neuron's computations that can predict its responses to arbitrary stimuli.

Unfortunately, a tight description of a neuron's computations does not necessarily provide insight about how (or whether) that neuron and its computations subserve a specific task; after a subunit model has been fit, the purpose of the neuron's computations is often unclear. Thus, although methods for neural systems identification are essential for determining what the components of nervous systems do, they are unlikely to determine why they do what they do. One way to address this issue is to develop normative frameworks (1) that determine the computations that are optimal for particular tasks and (2) that share the same or similar functional forms as popular methods for describing neural response.

AMA–Gauss is a normative method that is designed to find the filters (receptive fields) and quadratic pooling rules that are optimal for specific sensory-perceptual tasks (Figure 2C; see Methods). AMA–Gauss has a functional form that is closely related to the energy model and the GQM, but it has a different aim. Rather than describing what a neuron does, it prescribes what neurons should do. In fact, given a hypothesis about the function of a particular neuron, AMA–Gauss can predict the subunit filters and pooling rules that will be recovered by the GQM. The development of closely

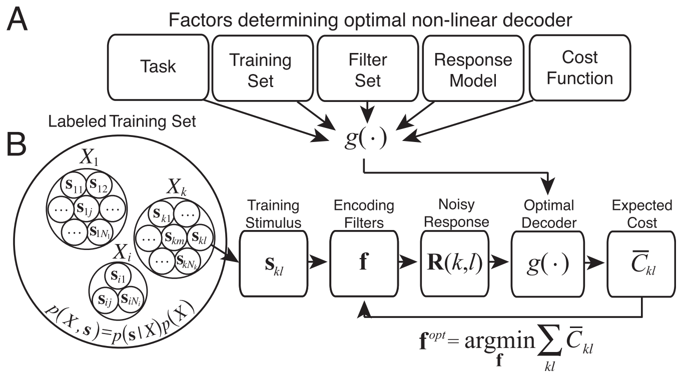
A
Factors determining optimal non-linear decoder
Task
Training Set
Filter Set
Response Model
Cost Function
$g(\cdot)$
B
Labeled Training Set
X_1
$\mathbf{s}_{11}$ $\mathbf{s}_{12}$ $\cdots$ $\mathbf{s}_{1j}$ $\cdots$ $\cdots$ $\mathbf{s}_{1N_1}$
X_k
$\mathbf{s}_{k1}$ $\cdots$ $\cdots$ $\mathbf{s}_{km}$ $\mathbf{s}_{kl}$ $\cdots$ $\mathbf{s}_{kN_k}$
X_i
$\mathbf{s}_{i1}$ $\mathbf{s}_{ij}$ $\mathbf{s}_{iN_i}$
$p(X,\mathbf{s})=p(\mathbf{s}|X)p(X)$
Training Stimulus
$\mathbf{s}_{kl}$
Encoding Filters
$\mathbf{f}$
Noisy Response
$\mathbf{R}(k,l)$
Optimal Decoder
$g(\cdot)$
Expected Cost
$\bar{C}_{kl}$
$\mathbf{f}^{opt} = \underset{\mathbf{f}}{\operatorname{argmin}} \sum_{kl} \bar{C}_{kl}$

Figure 3. Accuracy maximization analysis. (A) Factors determining the optimal decoder used by AMA during filter learning. (B) Steps for finding optimal task-specific filters via AMA.

related normative models and methods for neural systems identification is likely to enhance our ability to interpret fits to neural data and accelerate progress in psychophysical and neuroscientific research.

Methods

This section formally develops AMA–Gauss. To provide context for this technical contribution, we first review the setup and main equations for AMA (Geisler et al., 2009). Then, we derive the main equations for AMA–Gauss, provide a geometric intuition for how it works, and discuss practices for best use. Readers who are more interested in the scientific implications, and less interested in the mathematical formalisms, can skip ahead to Results.

Accuracy maximization analysis

The goal of AMA is to find the filters (receptive fields) that extract the most useful stimulus features for a particular task. Consistent with real biological systems, AMA places no constraints on the orthogonality of its filters, and its filter responses are corrupted by noise. AMA searches for the optimal filters with a closed form expression for the cost that relies on a Bayes optimal decoder. The filters are constrained to have unit magnitude ($||\mathbf{f}|| = 1.0$). The expression for the cost requires the specification of five factors (see Figure 3A). These factors are (1) a well-defined task (i.e., a latent variable to estimate from high-dimensional stimuli), (2) a labeled training set of stimuli, (3) a set of encoding filters, (4) a response noise model, (5) and a cost function (Figure 3A). The training set specifies the joint probability distribution $\mathcal{P}(X, \mathbf{s})$ between the latent variable X and the stimuli $\mathbf{s}$ (Figure 3B) and implicitly defines the prior $\mathcal{P}(X) = \sum_s \mathcal{P}(X, \mathbf{s})$ over the latent variable (see Discussion). If the training set is representative, results will generalize well to stimuli outside the training set.

For any particular filter set, the matched Bayes optimal decoder provides the cost by computing the posterior probability over the latent variable $\mathcal{P}(X|\mathbf{R})$, reading out the optimal estimate from the posterior, and then assigning a cost to the error. The steps for finding the optimal task-specific filters are: (1) select a particular stimulus $\mathbf{s}_{kl}$ from the labeled training set, (2) obtain a set of noisy filter responses $\mathbf{R}(k, l)$ from a particular (possibly nonoptimal) set of filters, (3) use the optimal non-linear decoder g(.) to obtain the optimal estimate $\hat{X}^{opt}$ and its expected cost $\bar{C}_{kl}$, (4) repeat for each stimulus and compute the average cost across all stimuli in the training set (5) update the filters to reduce the cost, and (6) repeat until the average cost is minimized. The optimal task-specific filters $\mathbf{f}^{opt}$ are those that minimize the cost (Figure 3B).

Bayes optimal decoder and filter response model

The Bayes optimal decoder gives a closed form expression for the cost for any filter or set of filters, given the training stimuli. The posterior probability of latent variable X_u given the noisy filter responses $\mathbf{R}(k, l)$ to stimulus $\mathbf{s}_{kl}$ is given by Bayes' rule

$$\mathcal{P}(X_u|\mathbf{R}(k,l)) = \frac{\mathcal{P}(\mathbf{R}(k,l)|X_u)\mathcal{P}(X_u)}{\sum_{i=1}^{N_{lvl}} \mathcal{P}(\mathbf{R}(k,l)|X_i)\mathcal{P}(X_i)} \quad (1)$$

where N_{lvl} is the number of latent variable level, and l indexes the stimuli having latent variable value X_k. The conditional distribution of noisy responses given the latent variable is

$$\mathcal{P}(\mathbf{R}|X_u) = \sum_{v=1}^{N_u} \mathcal{P}(\mathbf{R}|\mathbf{s}_{uv})\mathcal{P}(\mathbf{s}_{uv}|X_u) \quad (2)$$

where N_u is the number of stimuli having latent variable level X_u, and v indexes training stimuli having that latent variable value. Conveniently, $\mathcal{P}(\mathbf{s}_{uv}|X_u)$ and $\mathcal{P}(X_u)$ are determined by the training set; $\mathcal{P}(\mathbf{s}_{uv}|X_u) = \frac{1}{N_u}$ is the probability of particular stimulus v with latent variable X_u given that there are N_u such stimuli, and $\mathcal{P}(X_u) = \frac{N_u}{N}$ is the fraction of all stimuli having latent variable X_u. Therefore, Equation 1 reduces to

$$\mathcal{P}(X_u|\mathbf{R}(k,l)) = \frac{\sum_{v=1}^{N_u} \mathcal{P}(\mathbf{R}(k,l)|\mathbf{s}_{uv})}{\sum_{i=1}^{N_{lvl}} \sum_{J=1}^{N_i} \mathcal{P}\big(\mathbf{R}(k,l)|\mathbf{s}_{ij}\big)} \quad (3)$$

Equation 3 indicates that the posterior probability is given by the sum of the within-level stimulus likelihoods, normalized by the sum of all stimulus likelihoods.

Our aim is to understand task-specific information processing in biological systems. Thus, the response noise model should be consistent with the properties of biological encoders. AMA uses scaled additive (e.g., Poisson-like) Gaussian noise, a broadly used model of neural noise in early visual cortex (Geisler & Albrecht, 1997). Equations 4–7 define the response model, and specify the distribution of noisy filter responses $\mathcal{P}(\mathbf{R}|\mathbf{s}_{uv})$ to each stimulus. For an individual filter $\mathbf{f}_t$ from set of filters $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_q]$ (where q is the number of filters), the mean response r_t, noisy response R_t, and noise variance σ_t^2 to stimulus $\mathbf{s}_{uv}$ having latent variable value X_u are

$$r_{uv,t} = \mathbf{f}_t^T \mathbf{s}_{uv} \quad (4)$$

$$R_{uv,t} = r_{uv,t} + \eta_t \quad (5)$$

$$\eta_t \sim \mathcal{N}\left(0, \sigma_{uv,t}^2\right) \quad (6)$$

$$\sigma_{uv,t}^2 = \alpha|r_{uv,t}| + \sigma_0^2 \quad (7)$$

where η is a noise sample, α is the Fano factor, and σ_0^2 is baseline noise variance. The proximal stimulus $\mathbf{s} = \frac{\mathbf{x}-\bar{\mathbf{x}}}{||\mathbf{x}-\bar{\mathbf{x}}||}$ is contrast-normalized consistent with standard models (Albrecht & Geisler, 1991; Heeger, 1992), where $\mathbf{x}$ is an (possibly noise corrupted) intensity stimulus. If q filters are considered simultaneously, the response distributions $\mathcal{P}(\mathbf{R}|\mathbf{s})$, and the variables in Equations 4–7 become q-dimensional: mean response vector $\mathbf{r} = [r_1, r_2, ..., r_q]$, noisy response vector $\mathbf{R} = [R_1, R_2, ..., R_q]$, and response noise covariance matrix Λ.

The posterior probability distribution over the latent variable given the noisy filter responses to any stimulus in the training set is fully specified by Equations 3–7. The next step is to define the cost associated with a noisy response to an individual stimulus. The cost is given by

$$C_{kl} = \sum_X \left[\gamma\left(\hat{X}_{kl}^{opt}, X\right)\mathcal{P}(X|\mathbf{R}(k,l))\right] \quad (8)$$

where $\gamma(.)$ is an arbitrary cost function and $\hat{X}_{kl}^{opt}$ is the optimal estimate associated with noisy response $\mathbf{R}(k, l)$. The overall cost for a set of filters is the expected cost for each stimulus averaged over all stimuli

$$\bar{C} = \frac{1}{N}\sum_{k,l}^{N} E_{\mathbf{R}(k,l)}[C_{kl}] \quad (9)$$

The goal of AMA is to obtain the filters $\mathbf{f}$ that minimize the overall cost

$$\mathbf{f}^{opt} = \underset{\mathbf{f}}{\text{argmin}}\ \bar{C} \quad (10)$$

where $\mathbf{f}^{opt}$ are the optimal filters.

A single evaluation of the posterior probability distribution (Equation 3) for each stimulus in the training set requires $\mathcal{O}(N^2 N_{lvl})$ operations where N is the total number of stimuli and N_{lvl} is the number of latent variable levels in the training set. As noted earlier, this compute time makes AMA impractical for large scale problems without specialized computing resources.

There are at least two methods for achieving significant computational savings in optimization problems: employing models with strong parametric assumptions, and employing stochastic gradient descent routines. Both methods have drawbacks. Models with strong parametric assumptions are only appropriate for cases in which the assumptions approximately hold. Stochastic gradient descent routines are noisy and may not converge to the optimum filters. We have previously developed AMA–SGD, a stochastic gradient descent routine for AMA (Burge & Jaini, 2017). Here, we develop AMA–Gauss, a model with strong parametric assumptions.

AMA–Gauss

In this section, we first introduce AMA–Gauss and highlight its advantages over AMA. Subsequently, we provide expressions for AMA–Gauss likelihood function, L_2 and L_0 cost functions, and their gradients. We believe this is a valuable step toward making AMA a more practical tool in vision research.

AMA–Gauss: Class-conditional Gaussian assumption

AMA–Gauss is a version of AMA that makes the parametric assumption that the filter responses are Gaussian distributed when they are conditioned on a particular value of the latent variable

$$\mathcal{P}(\mathbf{R}|X_u) = \mathcal{N}(\mathbf{R}; \mu_u, \Sigma_u) \quad (11)$$

where $\mathbf{R}$ are responses to stimuli having latent variable level X_u,

$$\mu_u = \frac{1}{N_u}\sum_{v=1}^{N_u} \mathbf{f}^T\mathbf{s}_{uv} = \mathbf{f}^T\mathbf{s}_u \quad (12)$$

is the class-conditional mean of the noisy filter responses and

$$\Sigma_u = \frac{1}{N_u}\left[\sum_{v=1}^{N_u}\left(\left(\mathbf{f}^T\mathbf{s}_{uv} - \mathbf{f}^T\mathbf{s}_u\right)\left(\mathbf{f}^T\mathbf{s}_{uv} - \mathbf{f}^T\mathbf{s}_u\right)^T\right)\right] + \Lambda \quad (13)$$

is the class-conditional covariance of the noisy filter responses. The first term in Equation 13 is the class-conditional covariance of the expected filter responses. The second term in Equation 13, Λ, is the covariance matrix of the filter response noise $\eta \sim \mathcal{N}(\mathbf{0}, \Lambda)$. There are two major reasons for making the Gaussian assumption. First, if the response distributions are Gaussian, then AMA–Gauss will return the same filters as AMA while simultaneously providing huge savings in compute time. Second, the assumption is justified for at least two fundamental visual tasks in early vision (see Results; Burge & Geisler, 2014, 2015). With time, we speculate that similar statements will be justified for other sensory-perceptual tasks.

Under the AMA–Gauss assumption, the posterior probability (Equation 1) of latent variable X_u is

$$\mathcal{P}(X_u|\mathbf{R}(k,l)) = \frac{\mathcal{N}(\mathbf{R}(k,l);\mu_u,\Sigma_u)}{\sum_{i=1}^{N_{lvl}}\mathcal{N}(\mathbf{R}(k,l);\mu_i,\Sigma_i)} \quad (14)$$

where N_{lvl} is the number of latent variable levels. The AMA–Gauss posterior (Equation 14), has a simpler form than the AMA posterior (Equation 3). Hence, whereas a single evaluation of the AMA posterior probability distribution requires $\mathcal{O}(N^2 N_{lvl})$ operations (Equation 3), the AMA–Gauss posterior requires only $\mathcal{O}(NN_{lvl})$ operations where N is the number of stimuli in the training set (see Results). This reduction in compute-time substantially improves the practicality of AMA when the Gaussian assumption is justified. Even if the Gaussian assumption is not justified, AMA–Gauss is guaranteed to make the best possible use of first- and second-order conditional response statistics, and could thus provide a decent initialization at low computational cost.

AMA–Gauss: Derivations of the likelihood function, costs, and gradients

Analytic solutions for the optimal filters under AMA–Gauss (and AMA) are not available in closed form. Here, we provide expressions for the AMA–Gauss likelihood function, L_2 cost, L_0 cost, and their gradients.

The maximum likelihood AMA–Gauss encoding filters $\mathbf{f}_{\mathcal{L}}$ are those that simultaneously maximize the likelihood of the correct latent variable X_k across all stimuli in the training set. Stimuli having latent variable value X_k are indexed by l, and the i^{th} stimulus in the training set is denoted (k_i, l_i). The likelihood function of the AMA–Gauss filters is

$$\mathcal{L}(\mathbf{f}) = \prod_{i=1}^{N}\left[(2\pi)^{-\frac{d}{2}}|\Sigma_{k_i}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{R}(k_i,l_i)-\mu_{k_i}\right)^T\Sigma_{k_i}^{-1}\left(\mathbf{R}(k_i,l_i)-\mu_{k_i}\right)\right]\right] \quad (15)$$

The maximum likelihood filters can be determined by maximizing the likelihood function or, equivalently, minimizing the negative log-likelihood function

$$\mathbf{f}_{\mathcal{L}} = \underset{\mathbf{f}}{\operatorname{argmin}}\,[-\log\mathcal{L}(\mathbf{f})]$$

In practice, the expected negative log-likelihood is easier to minimize. Complete derivations of the likelihood function, the expected log-likelihood function, and closed form expressions for the associated gradients are provided in Appendix A. These expressions can be used to estimate the maximum-likelihood filters via gradient descent.

Next, we derive the AMA–Gauss cost for two popular cost functions for which the minimum mean squared error (MMSE) estimate and maximum a posteriori (MAP) are optimal: the L_2 and L_0 cost. The cost function specifies the penalty assigned to different types of error. For the L_2 (i.e. squared error) cost function, the expected cost for each stimulus $\mathbf{s}_{kl}$ (Equation 9) is

$$\bar{C}_{kl} = E_{\mathbf{R}(k,l)}\left[\left(\hat{X}_{kl}^{opt} - X_k\right)^2\right] \quad (16)$$

where the optimal estimate $\hat{X}_{kl}^{opt} = \sum_{u=1}^{N_{lvl}} X_u\mathcal{P}(X_u|\mathbf{R}(k,l))$ is the mean of the posterior.

For the L_0 (i.e., 0,1) cost function, the expected cost across all stimuli is closely related to the KL-divergence of the observed posterior and an idealized posterior with all its mass at the correct latent variable X_k; in both cases, cost is determined only by the posterior probability mass at the correct level of the latent variable (Burge & Jaini, 2017; Geisler et al., 2009). Here, the expected KL-divergence per stimulus is equal to the negative log-posterior probability at the correct level (Geisler et al., 2009)

$$\bar{C}_{kl} = E_{\mathbf{R}(k,l)}[-\log\mathcal{P}(X_k|\mathbf{R}(k,l))] \quad (17)$$

In a slight abuse of terminology, we refer to this divergence as the L_0 or KL-divergence cost.

The gradient of the total expected cost across all stimuli can be evaluated by calculating the gradient of the cost for each stimulus $\nabla_{\mathbf{f}}\bar{C}_{kl}$ (see Equation 9). Hence, the gradient of the total expected cost is

$$\nabla_{\mathbf{f}}\bar{C} = \frac{1}{N}\sum_{k,l}^{N}\nabla_{\mathbf{f}}\bar{C}_{kl} \quad (18)$$

The gradient of the cost for each stimulus can be evaluated by calculating the gradient of the posterior probability. Complete derivations of the cost and the gradient of the cost for the L_2 and L_0 cost functions are given in Appendix B and Appendix C, respectively.

Cost is minimized when responses to stimuli having different latent variable values overlap as little as possible. The cost functions (i.e., max-likelihood, L_0 cost, L_2 cost) exert pressure on the filters to produce class-conditional response distributions that are as different as possible given the constraints imposed by the stimuli. Hence the optimal filters will (1) maximize

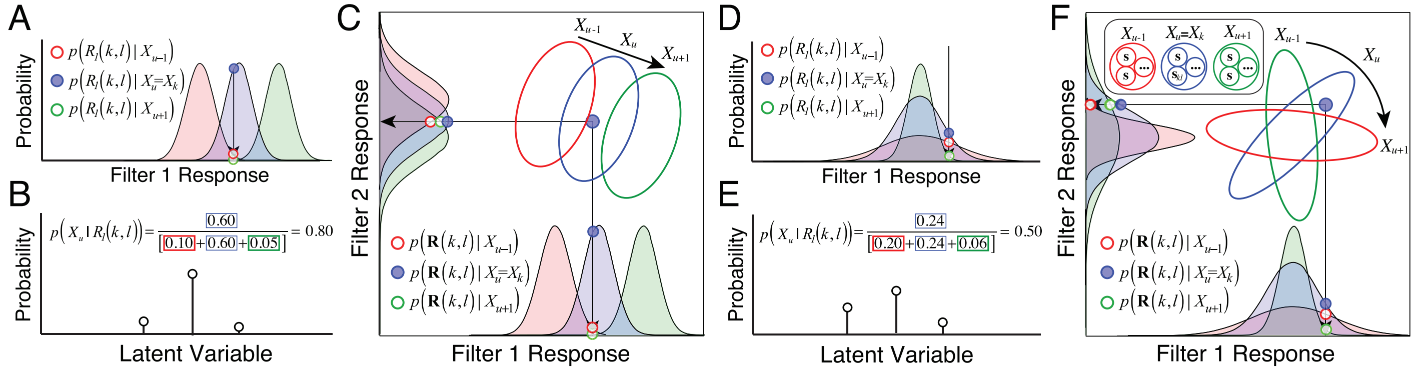

Figure 4. Relationship between conditional filter response distributions, likelihood, and posterior probability. Two hypothetical cases are considered, each with three latent variable values. (A) One-dimensional (i.e., single filter) Gaussian conditional response distributions, when information about the latent variable is carried only by the class-conditional mean; distribution means, but not variances, change with the latent variable. The blue distribution represents the response distribution to stimuli having latent variable value X_k. The red and green distributions represent response distributions to stimuli having different latent variables values$X_u \neq X_k$. The blue dot represents the likelihood $\mathcal{L}(X_u; R_1(k,l)) = \mathcal{N}(R_1(k,l); \mu_u, \Sigma_u)$ that observed noisy filter response $R_1(k, l)$ to stimulus $\mathbf{s}_{k,l}$ was elicited by a stimulus having latent variable level $X_u = X_k$. Red and green dots represent the likelihoods that the response was elicited by a stimulus having latent variable $X_u \neq X_k$ (i.e., by a stimulus having the incorrect latent variable value). (B) Posterior probability over the latent variable given the noisy observed response in (A). The posterior probability of the correct latent variable value (in this case, X_k) is given by the likelihood of the correct latent variable value normalized by the sum of all likelihoods. Colored boxes surrounding entries in the inset equation indicate the likelihood of each latent variable. (C) Two-dimensional (i.e., two-filter) Gaussian response distributions. Each ellipse represents the joint filter responses to all stimuli having the same latent variable value. The second filter improves decoding performance by selecting for useful stimulus features that the first filter does not. The black dot near the center of the blue ellipse represents an observed noisy joint response $\mathbf{R}(k, l)$ to stimulus $\mathbf{s}_{k,l}$. The likelihood $\mathcal{L}(X_u; \mathbf{R}(k,l)) = \mathcal{N}(\mathbf{R}(k,l); \mu_u, \Sigma_u)$ that the observed response was elicited by a stimulus having latent variable value X_u is obtained by evaluating the joint Gaussian at the noisy response; in this case, the product of the likelihoods represented by the blue dots on the single filter response distributions. (D–F) Same as A–C, but where information about the latent variable is carried by the class-conditional covariance instead of the mean; ellipse orientation, but not location, changes with the latent variable. AMA–Gauss finds the filters yielding conditional response distributions that are as different from each other as possible, given stimulus constraints.

the differences between the class-conditional means or covariances and (2) minimize the generalized variance for each class-conditional response distribution. (Generalized variance is a measure of overall scatter, represents the squared volume of the ellipse, and is given by the determinant of the covariance matrix.)

AMA–Gauss: Geometric intuition

Figure 4 provides a geometric intuition for the relationship between the filter response distributions, the likelihood, and the posterior probability distribution for two simple hypothetical cases. Both cases have three latent variable values. In one case, the information about the latent variable is carried by the class-conditional mean (Figure 4A–C). In the other case, the information about the latent variable is carried by the class-conditional covariance (Figure 4D–F). In all cases, the class-conditional responses to stimuli having the same latent variable value are Gaussian distributed. With a single filter, the response distributions are one-dimensional (Figure 4A, D). For any observed noisy response R, the likelihood of a particular level of the latent variable X_u is found by evaluating its response distribution at the observed response (blue dot; Figure 4A, D). The posterior probability of latent variable X_u is obtained by normalizing with the sum of the likelihoods (blue, red, and green dots; Figure 4B, E). With two filters, the response distributions are two-dimensional (red, blue, and green ellipses with corresponding marginals; Figure 4C, F). The second filter will increase the posterior probability mass at the correct value of the latent variable (not shown) because the second filter selects for useful stimulus features that the first filter does not. These hypothetical cases illustrate why cost is minimized when mean or covariance differences are maximized between classes and generalized variance is minimized within classes. The filters that make the response distributions as different as possible make it as easy as possible to decode the latent variable.

AMA–Gauss: Best practices

The AMA–Gauss method developed here does not automatically determine the number of stimuli to train on, or the number of task-specific filters to learn; these choices are left to the user.

To obtain representative results (i.e., to minimize sampling error) the training set must be of sufficient size. AMA–Gauss uses the sample mean and covariance to approximate the Gaussian distributions of filter responses conditional on each value of the latent variable (Equation 11). Training sets with at least 250 stimuli per level tend to give representative results.

To extract all task-relevant information from each stimulus a sufficient number of receptive fields must be learned. In general, the best practice is to learn filters until the change in the value of the total cost is negligible (Geisler et al., 2009). The current paper aims to demonstrate the properties and usefulness of AMA–Gauss rather than determine the best number of filters; for clarity, we show only four filters for each task (see Results). Previous work has shown that, for the two tasks considered here, eight filters are required to capture nearly all task-relevant information (Burge & Geisler, 2014, 2015). The results presented in this paper hold for all eight filters, but we show only four for ease of presentation.

Results

Retinal speed estimation and binocular disparity estimation are canonical visual tasks. Accurate and precise estimation of retinal image motion is critical for the accurate estimation of object motion and self-motion through the environment. Accurate and precise estimation of binocular disparity is critical for the accurate estimation of depth and the control of fixational eye movements. Although of fundamental importance for mobile seeing organisms, both tasks are difficult in natural conditions because of the enormous variability and complexity in natural images.

The plan for the results section is as follows. First, we use AMA–Gauss[1] to find the receptive fields that are optimal for estimating speed and disparity from local patches of natural images. Second, we compare AMA–Gauss and AMA and show that both methods (1) learn the same filters and (2) converge to the same cost for both tasks. Third, we verify that AMA–Gauss achieves the expected reductions in compute-time: filter-learning with AMA–Gauss is linear whereas AMA is quadratic in the number of stimuli in the training set. Fourth, we show that the class-conditional filter responses are approximately Gaussian, thereby justifying the Gaussian assumption for these tasks. Fifth, we show how contrast normalization contributes to the Gaussianity of the class-conditional responses. Sixth, we explain how the filter response distributions determine the likelihood functions and optimal pooling rules. Seventh, we explain how these results provide a normative explanation for why energy-model-like computations describe the response properties of neurons involved in these tasks. Eighth, and last, we establish the formal relationship between AMA–Gauss and the GQM, a recently developed method for neural systems identification.

For each task, we obtained an existing labeled training set of natural photographic stimuli consisting of approximately 10,000 randomly sampled stimuli. All stimuli subtended 1° of visual angle. Perspective projection, physiological optics, and the wavelength sensitivity, spatial sampling, and temporal integration functions of the foveal cones were accurately modeled. Input noise was added to each stimulus with a noise level just high enough to mask retinal image detail that would be undetectable by the human visual system (Williams, 1985). Both training sets had flat prior probability distributions $\mathcal{P}(X)$ over the latent variable (see Discussion). The training set for speed estimation consisted of 10,500 stimuli [10,500 stimuli = 500 stimuli/level × 21 levels; (Burge & Geisler, 2015)]. Retinal speeds ranged from $-8°$/s to $+8°$/s; negative and positive speeds correspond with leftward and rightward drifting movies. Each stimulus had a duration of 250 ms. The training set for disparity estimation consisted of 7,600 stimuli [7,600 stimuli = 400 stimuli/level × 19 levels; (Burge & Geisler, 2014)]. Binocular disparities ranged from -16.875 arcmin to $+16.875$ arcmin; negative and positive disparities correspond to uncrossed and crossed disparities. (Note that although these training sets have a discrete number of latent variable values, AMA filters can be learned with discrete- or with real-valued latent variables.) For extensive additional details on these training sets and for ideal observer performance in these tasks, please see Burge and Geisler, 2014, 2015. One important limitation of these datasets is that all motion signals were rigid and that all disparity signals were planar. Future work will examine the impact of nonrigid motion (e.g., looming) and local depth variation (e.g., occlusion) on performance (see Discussion).

Before processing, retinal image stimuli for both tasks were vertically averaged under a raised cosine window (0.5° at half-height). Vertically oriented linear receptive fields respond identically to the original and vertically averaged stimuli, and canonical receptive fields for both tasks are vertically oriented (Burge & Geisler, 2014, 2015). Thus, the vertically averaged signals represent the signals available to the orientation column that would be most useful to the task. Future work will examine the impact of off-vertical image features on performance.

Next, we used AMA–Gauss to find the optimal filters for both tasks. The results presented as follows were obtained using the L_0 cost function and constant, additive, independent filter response noise. In general, we have found that the optimal filters are quite robust to the choice of cost function when trained with natural

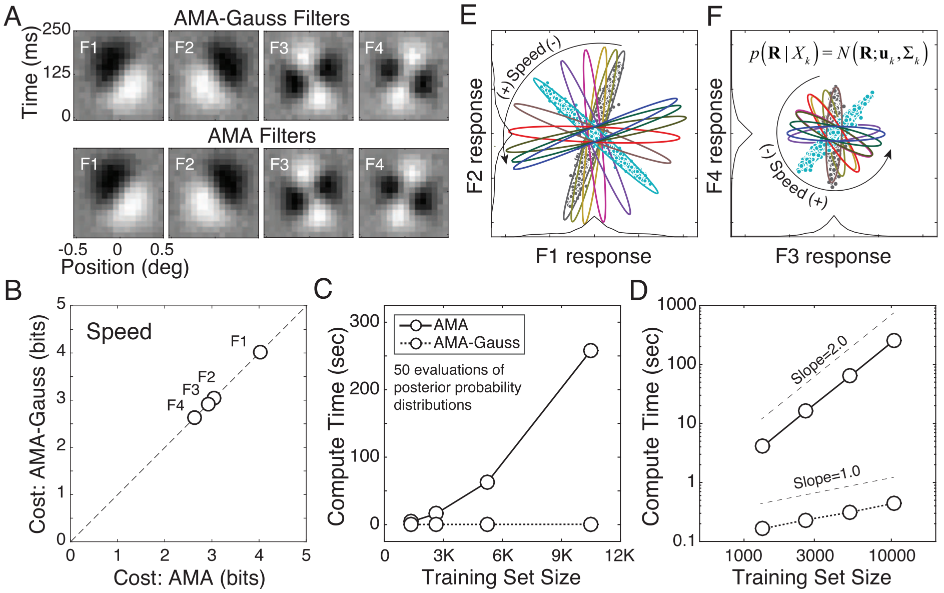

Figure 5. Speed estimation task: filters, cost, compute time, and class-conditional response distributions. (A) AMA–Gauss and AMA filters for estimating speed (−8 to +8°/s) from natural image movies are nearly identical; $\rho > 0.96$ for all filters. (B) The cost for all the filters for both the models is identical. (C) Compute time for 50 evaluations of the posterior probability distribution is linear with AMA–Gauss, and quadratic with the full AMA model, in the training set size. (D) Same data as in (C) but on log–log axes. (E), (F) Joint filter responses, conditioned on each level of the latent variable, are approximately Gaussian. Different colors indicate different speeds. Individual symbols represent responses to individual stimuli. Thin black curves show that the filter response distributions, marginalized over the latent variable $\mathcal{P}(R) = \sum_{u=1}^{N_{lvl}} \mathcal{P}(R|X_u)\mathcal{P}(X_u)$, are heavier-tailed than Gaussians (see Results and Discussion).

stimuli (Burge & Jaini, 2017). Figure 5 shows results for the retinal speed estimation task. Figure 6 shows results for the disparity estimation task. AMA and AMA–Gauss learn nearly identical encoding filters (Figure 5A and 6A; $\rho > 0.96$) and exhibit nearly identical estimation costs (Figure 5B and 6B); note, however, that these filter and performance similarities are not guaranteed (see Appendix D). AMA–Gauss also dramatically reduces compute time (Figures 5C, D and 6C, D). With AMA, the time required to learn filters increases quadratically with their number of stimuli in the training set. With AMA–Gauss, filter learning time increases linearly with the number of stimuli. Finally, the class-conditional filter responses are approximately Gaussian (Figures 5E, F and 6E, F), indicating that the Gaussian assumption is justified for both tasks. Quadratic computations are therefore required to determine the likelihood of a particular value of the latent variable. The posterior probability distribution over the latent variable $\mathcal{P}(X|\mathbf{R})$ can be obtained from the likelihoods by straightforward application of Bayes' rule.

Response normalization, response Gaussianity, and decoding performance

Contrast varies significantly in natural stimuli. How does contrast normalization affect the filter responses? For the class of problems considered here (e.g., retinal speed estimation, binocular disparity estimation, and other energy-model-related tasks), neurophysiologically plausible contrast normalization (Albrecht & Geisler, 1991; Heeger, 1992) must be built into the filter response model (Equation 4) for the class-conditional filter responses $\mathcal{P}(\mathbf{R}|X_u)$ to be Gaussian distributed. [Note that some standard models of normalization are computationally equivalent (Albrecht & Geisler, 1991; Heeger, 1992), but that other more specialized forms of normalization are not (Carandini & Heeger, 2012; Coen-Cagli, Dayan, & Schwartz, 2012; Gao & Vasconcelos, 2009).) In AMA–Gauss, the input stimulus $\mathbf{s}$ is a contrast normalized ($||\mathbf{s}|| \leq 1.0$) version of a (possibly noise-corrupted) intensity stimulus $\mathbf{x}$ with mean intensity $\bar{\mathbf{x}}$. Luminance normalization converts the intensity stimulus to a contrast stimulus $\mathbf{c} = \frac{\mathbf{x}-\bar{\mathbf{x}}}{\bar{\mathbf{x}}}$ by subtracting off and dividing by the mean. Contrast normalization converts the contrast stimulus to a contrast normalized signal with unit magnitude (or less) $\mathbf{s} = \frac{\mathbf{c}}{\sqrt{nc_{50}^2 + \sum_i \mathbf{c}_i^2}}$ where c_{50} is an additive constant and n is the dimensionality of (e.g., number of pixels defining) each stimulus. Here, we assumed that the value of the additive constant is $c_{50} = 0.0$. The effect of the value of c_{50} has been studied previously (Burge & Geisler, 2014).

To examine the effect of contrast normalization on the class-conditional filter response distributions, we computed the filter responses to the same stimuli with and without contrast normalization. With contrast

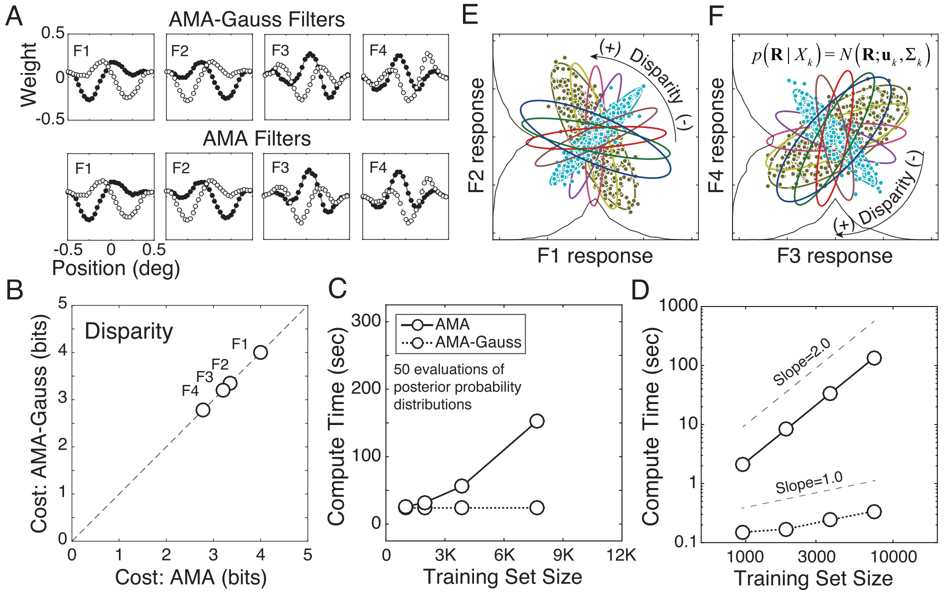

Figure 6. Disparity estimation task: filters, cost, compute time, and class-conditional response distributions. (A) AMA–Gauss and AMA filters for estimating disparity from natural stereo-images (−15 to +15 arcmin). (B–F) Caption format same as Figure 5B–F.

normalization, filter response distributions are approximately Gaussian (Figure 7A, B, E, F). Without contrast normalization, filter response distributions have tails much heavier than Gaussian (Figure 7C, D, G, H). (Note that AMA–Gauss learns very similar filters with and without contrast normalization. Normalization does not change which stimulus features should be selected; it changes only how the selected features are represented.) Thus, biologically realistic normalization helps Gaussianize the conditional response distributions. Related results have been reported by other groups (Lyu & Simoncelli, 2009; Wang, Bovik, Sheikh, & Simoncelli, 2004).

Contrast normalization not only Gaussianizes the response distributions; it also improves performance. If response distributions are heavy-tailed and have strong peaks at zero, then the Gaussian assumption is violated and attempts to decode the latent variable from those responses suffer. Contrast normalization reduces the peak at zero, thereby reducing decoding difficulty. Figure 8 compares decoding cost in the speed and disparity tasks with and without contrast normalization and shows that failing to normalize harms performance. Thus, contrast normalization improves task performance by decreasing kurtosis and increasing response Gaussianity.

Subunit response models (e.g., the standard energy model, the GQM, and other LN models) are widely used to describe and fit neurons. They do not generally incorporate normalization (Adelson & Bergen, 1985; Park et al., 2013; Rust et al., 2005; Vintch et al., 2015). This fact is unsurprising. Many laboratory experiments use high-contrast white noise stimuli to map neural receptive fields (Jones & Palmer, 1987a, 1987b). Linear subunit receptive field responses to Gaussian noise are guaranteed to be Gaussian, so the lack of contrast normalization does not hurt performance in common laboratory conditions. With natural signals, the failure to normalize can hurt performance. Perhaps this is one reason why subunit models tend to generalize poorly to natural stimuli (but see Eickenberg, Rowekamp, Kouh, & Sharpee, 2012). It may be useful to incorporate response normalization in future instantiations of these models.

Data-constrained likelihood functions

The class-conditional response distributions fully determine the likelihood function over the latent variable for any joint filter response **R** to an arbitrary stimulus. When the class-conditional response distributions are Gaussian, as they are here, the log-likelihood of latent variable value X_u is quadratic in the encoding filter responses

$$\log\mathcal{L}(X_u;\mathbf{R}) = \log\mathcal{P}(\mathbf{R}|X_u)$$
$$= -\frac{1}{2}(\mathbf{R}-\mu_u)^T\Sigma_u^{-1}(\mathbf{R}-\mu_u)+\zeta_u \quad (19)$$

where $\zeta_u = -\frac{1}{2}\log|2\pi\Sigma_u|$. (Note that the likelihood function over the latent variable (Equation 19) is distinct from likelihood function over the AMA–Gauss filters; Equation 15.) Carrying out the matrix multiplication shows that the log-likelihood can be re-expressed as the weighted sum of squared, sum-squared, and linear filter responses

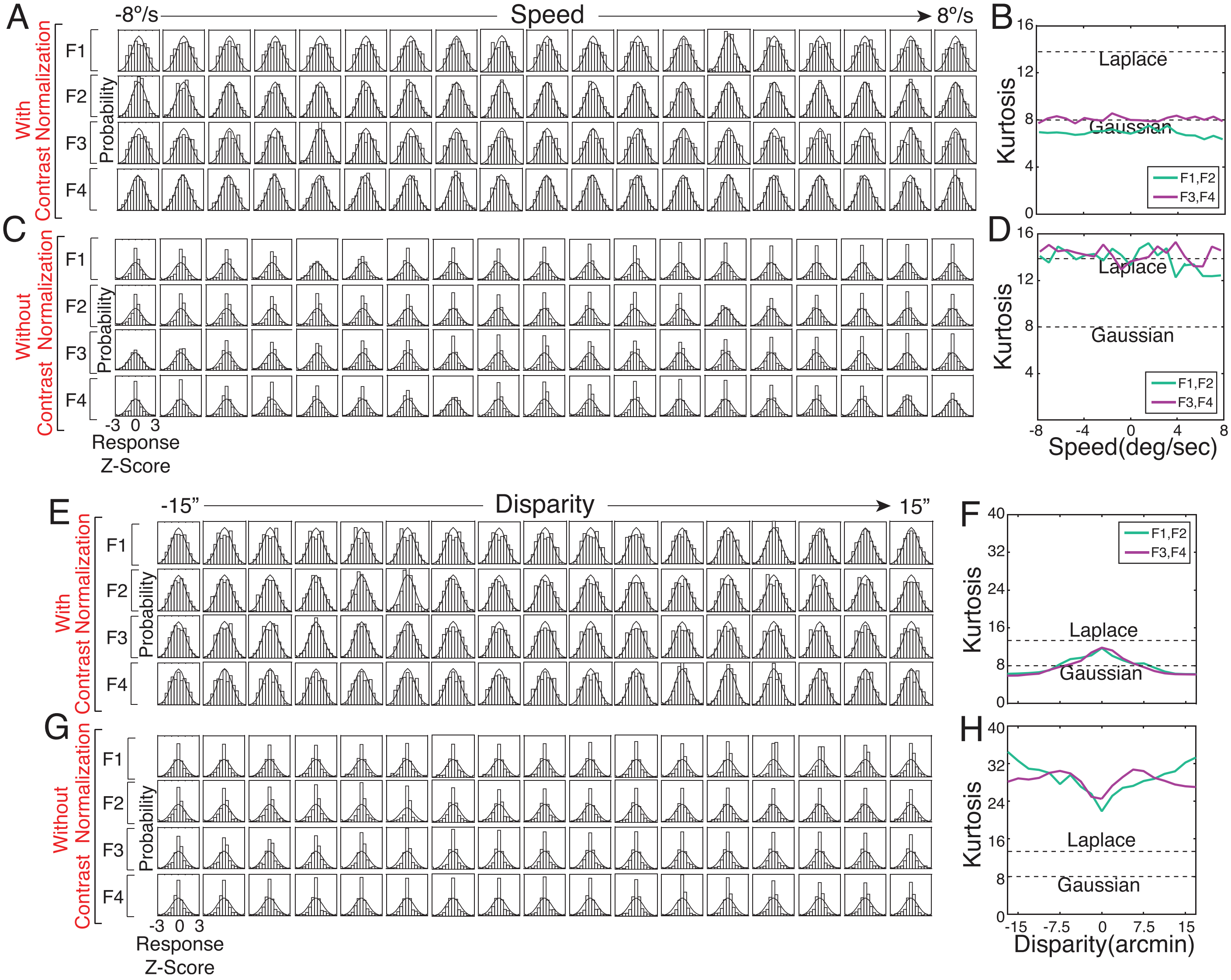

Figure 7. Filter responses with and without contrast normalization. (A) Class-conditional filter response distributions $\mathcal{P}(R_t|X_u)$ to contrast-normalized stimuli for each individual filter and each level of the latent variable in the speed estimation task. For visualization, responses are transformed to *Z*-scores by subtracting off the mean and dividing by the standard deviation. Gaussian probability density is overlaid for purposes of comparison. (B) Kurtosis of the two-dimensional conditional response distributions from filters 1 and 2 (violet; also see Figure 5E) and filters 3 and 4 (green; also see Figure 5F) for all levels of the latent variable. A two-dimensional Gaussian has a kurtosis of 8.0. Kurtosis was estimated by fitting a multidimensional generalized Gaussian via maximum likelihood methods. (C, D) Same as A, B but without contrast normalization. (E–H) Same as (A–D), but for the task of disparity estimation.

$$\log[\mathcal{P}(\mathbf{R}|X_u)] = \sum_{i=1}^{q} \mathbf{w}_{i,u} R_i + \sum_{ii=1}^{q} \mathbf{w}_{ii,u} R_i^2 + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \mathbf{w}_{ij,u} \left(R_i + R_j\right)^2 + \zeta'_u \tag{20}$$

where q is the number of filters and where the weights are functions of the class-conditional mean and covariance for each value X_u of the latent variable (Burge & Geisler, 2014, 2015). Specifically,

$$\mathbf{w}_{i,u} = \Sigma_u^{-1} \mu_u \tag{21}$$

$$\mathbf{w}_{ii,u} = -diag\left(\Sigma_u^{-1}\right) + 0.5\Sigma_u^{-1}\mathbf{1} \tag{22}$$

$$\mathbf{w}_{ij,u} = -0.5\Sigma_{ij,u}^{-1}, \forall ij \text{ where } j > i \tag{23}$$

$$\zeta'_u = -0.5\mu_u^T \Sigma_u^{-1} \mu_u + \zeta_u \tag{24}$$

where $diag(.)$ is a function that returns the diagonal of a matrix and $\mathbf{1}$ is a column vector of ones. [Note that in these equations i and j index different filters (see Figure

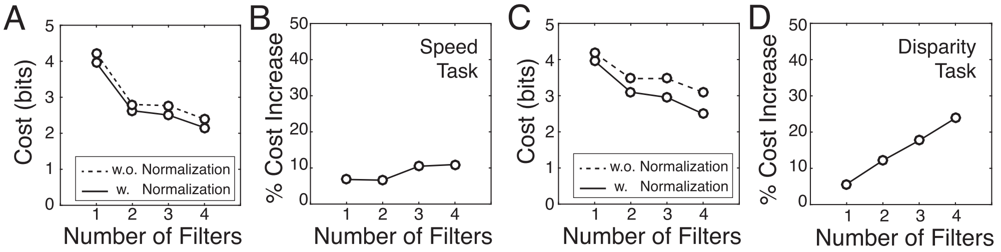

Figure 8. Decoding performance with and without contrast normalization. (A) Contrast normalization decreases decoding cost for the speed estimation task. (B) Percentage increase in cost without contrast normalization. (C, D) Same as (A, B) but for the disparity estimation task. The same result holds for different cost functions (e.g., squared error) and larger number of filters. If eight filters are used in the disparity task, failing to contrast normalize can decrease performance by ~40%.

2), not different latent variables and stimuli, as they do elsewhere in this manuscript.] These equations (Equations 20–24) indicate that the log-likelihood of latent variable value X_u is obtained by pooling the squared (and linear) responses of each receptive field with weights determined by the mean $\boldsymbol{\mu}_u$ and covariance Σ_u) of the subunit responses to stimuli with latent variable X_u.

In the speed and disparity estimation tasks, nearly all of the information about the latent variable is carried by the class-conditional covariance; the covariance of the filter responses to natural stimuli changes significantly with changes in the latent variable (see Figures 4D–F, 5E, F, and 6E, F). Thus, the weights on the squared and the sum-squared filter responses change dramatically with the value of the latent variable (Figure 9). In the speed estimation task, for example, the weights $\mathbf{w}_{34}(X)$ on the sum-squared response of filter 3 and filter 4 peak at 0°/s (see Figure 9A). This peak results from the fact that the filter 3 and filter 4 response covariance is highest at 0°/s (see Figure 5F; Equation 23). In contrast, very little information is carried by the class-conditional means; the mean filter responses to natural stimuli are always approximately zero. Hence, the weights on the linear subunit responses are approximately zero (see Equation 21, Figure 4A–C).

The filter response distributions determine the computations (i.e., quadratic pooling rules and weights) required to compute the likelihood of different latent variable values. If these computations (Equation 20–24) are paired with an exponential output nonlinearity and implemented in a neuron, the neuron's response $R_u^L = \mathcal{L}(X_u; \mathbf{R})$ would represent the likelihood that a stimulus having a particular value X_u of the latent variable elicited the observed filter responses $\mathbf{R}$. This latent variable value X_u would be the preferred stimulus of the likelihood neuron. We refer to this hypothetical neuron as an AMA–Gauss likelihood neuron (see Equation 20).

Four example, likelihood functions are shown in Figure 10A, one for each of four stimuli having a true speed of −4°/s. Figure 10B shows four likelihood functions for stimuli having a true speed of 0°/s. Figure 10C, D show likelihood functions for stimuli having −15 arcmin and 0 arcmin of binocular disparity, respectively. These plots show the likelihood functions, but they are not the standard way of assessing the response properties of neurons in cortex.

The response properties of neurons in cortex are most commonly assessed by their tuning curves. Likelihood neuron tuning curves are obtained by first computing the mean likelihood neuron response across all natural stimuli having latent variable value X_k

$$\bar{R}_u^L(X_k) = \frac{1}{N_k}\sum_{l=1}^{N_k} \mathcal{L}(X_u; \mathbf{R}(k,l)) \quad (25)$$

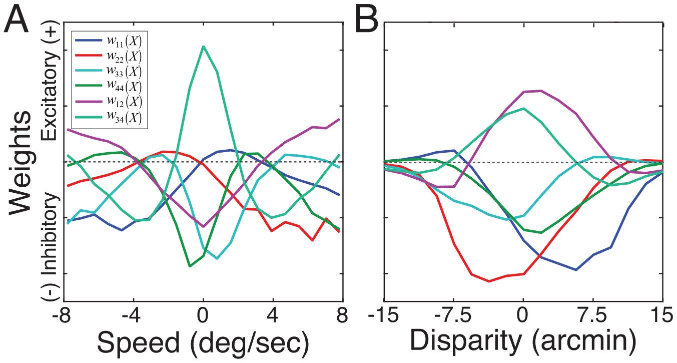

Figure 9. Quadratic pooling weights for computing the likelihood. Weights on squared and sum-squared filter responses ($\mathbf{w}_{ii}(X)$ and $\mathbf{w}_{ij}(X)$) as a function of the latent variable. Weights on the linear filter responses are all approximately zero and are not shown. (A) Weights for speed estimation task. (B) Weights for disparity estimation task. Weights on squared responses are at maximum magnitude when the variance of the corresponding filter responses is at minimum. Weights on sum-squared responses are at maximum magnitude for latent variables yielding maximum response covariance (see Figures 5E, F and 6E, F).

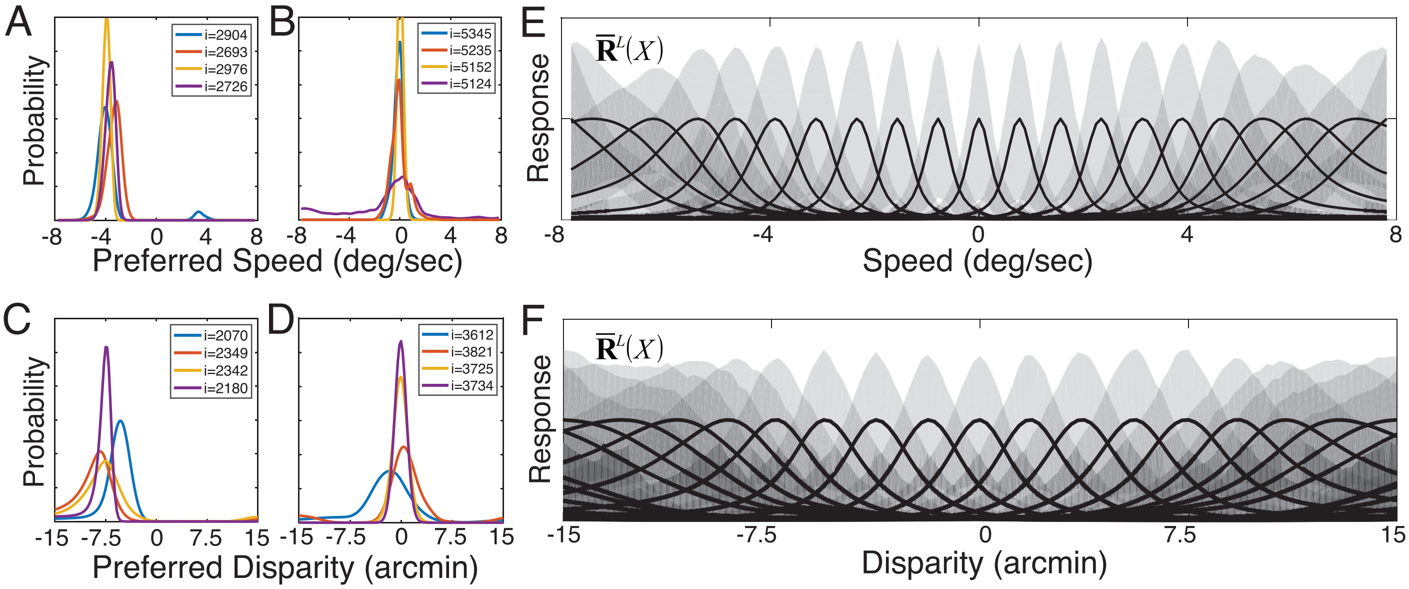

Figure 10. Likelihood functions for speed and disparity tasks. (A) Likelihood functions for four randomly chosen natural image movies having true speeds of 4°/s. Each likelihood function represents the population response of the set of likelihood neurons, arranged by their preferred speeds. To ease the visual comparison, the likelihood functions are normalized to a constant volume by the sum of the likelihoods. (B) Same as (A), but for movies with a true speed of 0°/s. (C, D) Same as (A, B) but for stereo-images with −7.5 arcmin and 0.0 arcmin of disparity, respectively. (E) Tuning curves of speed-tuned likelihood neurons. For speeds sufficiently different from zero, tuning curves are approximately log-Gaussian and increase in width with speed. For near-zero speeds, tuning curves are approximately Gaussian. Each curve represents the mean response (i.e., tuning curve) of a likelihood neuron having a different preferred speed, normalized to a common maximum. Gray areas indicate 68% confidence intervals. (F) Tuning curves of disparity-tuned likelihood neurons.

and then repeating for all values of the latent variable. Tuning curves for a population of likelihood neurons $\overline{\mathbf{R}}^L(X)$ having a range of preferred speeds are shown in Figure 10E. The speed tuning curves are approximately Gaussian for preferred speeds near 0°/s and approximately log-Gaussian otherwise. Consistent with these results, neurons in middle temporal (MT) area have approximately log-Gaussian speed tuning curves, and have bandwidths that increase systematically with speed (Nover, Anderson, & DeAngelis, 2005). It is also interesting to note that while quadratic computations are required to optimally decode the latent variable directly from the filter responses (see Figure 5E, F), likelihood neuron responses are linearly separable in speed. Similar points can be made about the disparity likelihood neurons (Figure 10F). The computations reported here thus constitute a general recipe for how to construct selective, invariant neurons having an arbitrary preferred stimulus (latent variable) from the responses of a small, well-chosen set of receptive fields.

Linking AMA–Gauss and the energy model

Neural activity involved in many visual tasks has been productively modeled by energy-model-like (i.e., quadratic) computations (Cumming & DeAngelis, 2001; Emerson, Bergen, & Adelson, 1992; Peng & Shi, 2010). We have shown that in two classic tasks (retinal speed and binocular disparity estimation), the class-conditional filter response distributions to natural stimuli are approximately Gaussian distributed. In such cases, quadratic combinations of the filter responses are the optimal computations and yield the likelihood of a particular value of the latent variable (Equation 20). The weights are determined by the filter responses to natural stimuli (see Methods). Thus, if these computations were instantiated in a neuron, then its response would represent the likelihood of latent variable (Figure 2C). The current results therefore constitute a normative explanation for why energy-model-like computations account for response properties of neurons involved in these tasks.

Interestingly, in recent years, discrepancies have emerged between the properties of neurons in cortex and the energy models that are often used to describe them (Cumming & DeAngelis, 2001; Rust et al., 2005; Tanabe, Haefner, & Cumming, 2011). Many of these discrepancies are a natural consequence of the optimal computations for estimating disparity and motion from natural stimuli (Burge & Geisler, 2014, 2015). For example, the responses of motion- and disparity-selective neurons have both been found to depend on multiple excitatory and suppressive subunit receptive fields, rather than the two exclusively excitatory subunit receptive fields posited by the energy model. Multiple subunit receptive fields have increased potential to select task-relevant information from each stimulus. Excitatory and inhibitory weighting schemes are required to use the selected information optimally. The quadratic computations in Equation 20 specify exactly how to optimally weight and sum the responses from multiple receptive fields to achieve selectivity for particular latent variable values (also see Figure 9). These computations yield more selective, invariant tuning curves (and improved estimation performance) than those of the standard energy model (Burge & Geisler, 2014, 2015), and follow directly from the normative framework employed here.

Linking AMA–Gauss and the GQM: Connecting normative and response triggered analyses

In this section, we establish the formal similarities between AMA–Gauss and the generalized model (GQM), a recently developed subunit model for neural systems identification (Park et al., 2013; Wu et al., 2015). The goal of the GQM is to identify (fit) the subunit receptive fields that drive a neuron's response (Figure 2B). The goal of AMA–Gauss is to find the subunit receptive fields and quadratic pooling rules that are best for a particular task (Figure 2C). AMA can thus generate predictions about the subunit receptive fields that the GQM will recover from a neuron, under the hypothesis that the neuron computes the likelihood of a task-relevant latent variable.

The GQM assumes that a neuron's spiking or intracellular voltage response to a stimulus is given by

$$y \sim P(f(Q(\mathbf{x}))) \quad \text{where } Q(\mathbf{x}) = \mathbf{x}^T C\mathbf{x} + \mathbf{b}^T\mathbf{x} + a \tag{26}$$

where y is the neural response, $P(.)$ is the noise model, $f(.)$ is a nonlinearity, and $\lambda = f(Q(\mathbf{x}))$ is the mean response. In Park et al., 2013, the authors use maximum likelihood methods to recover the parameters of the model given a set of stimuli, the neuron's response to each stimulus, and a noise model. In AMA–Gauss, the log-likelihood of latent variable X_u is given by

$$l(X_u) = -\frac{1}{2}(\mathbf{R} - \mu_u)^T\Sigma_u^{-1}(\mathbf{R} - \mu_u) + \zeta_u \tag{27}$$

where $\boldsymbol{\mu}_u$ and Σ_u are the class-conditional response mean and covariance and ζ_u is a constant. The noisy filter response vector $\mathbf{R}$ is given by the projection of the stimulus onto the filters $\mathbf{f}$ plus noise (Equations 4, 5). Hence, Equation 27 can be rewritten as

$$l(X_u) = -\frac{1}{2}\Big(\mathbf{x}^T\mathbf{f}\Sigma_u^{-1}\mathbf{f}^T\mathbf{x} - 2(\mu_u^T\Sigma_u^{-1} - \eta_u^T\Sigma_u^{-1})\mathbf{f}^T\mathbf{x} + \mu_u^T\Sigma_u^{-1}\mu_u - \eta_u^T\Sigma_u^{-1}\eta_u + 2\eta_u^T\Sigma_u^{-1}\mu_u\Big) + \zeta_u \tag{28}$$

$$\text{or} \quad l(X_u) = \mathbf{x}^T C\mathbf{x} + \mathbf{b}^T\mathbf{x} + a$$

where $C = -\frac{1}{2}\mathbf{f}^T\Sigma_u^{-1}\mathbf{f}$ is a rank-q matrix where q is the number of filters, $\mathbf{b}^T = \mu_u^T\Sigma_u^{-1}\mathbf{f}^T - \eta_u^T\Sigma_u^{-1}\mathbf{f}^T$, and $a = -\frac{1}{2}\mu_u^T\Sigma_u^{-1}\mu_u + \frac{1}{2}\eta_u^T\Sigma_u^{-1}\eta_u - \eta_u^T\Sigma_u^{-1}\mu_u + \zeta_u$. (Parameter values under the expected log-likelihood are provided in Appendix E). The parameters of the GQM are therefore simple functions of the AMA–Gauss encoding filters $\mathbf{f}$ and their responses to natural stimuli, conditional on latent variable X_u. Given a hypothesis about the functional purpose of a neuron's activity, AMA–Gauss could predict the parameters that the GQM would recover via response-triggered analyses.

The primary formal distinction between AMA–Gauss and the GQM is that AMA–Gauss explicitly models noise in the encoding filter responses, whereas the GQM models noise only after quadratic pooling of the filter responses; that is, the GQM implicitly assumes noiseless filter responses. When subunit responses are noiseless, all subunit receptive fields spanning the same subspace (i.e., all linear filter combinations) provide an equivalent encoding. When responses are noisy (as they are in all biological systems), the stimulus encodings provided by different filters spanning the same subspace are no longer equivalent (Burge & Jaini, 2017). Future work will examine whether this distinction between AMA and the GQM can be leveraged to overcome a limitation common to all standard subunit models, namely, that their descriptions of neurons are unique only up to the subspace spanned by the subunit receptive fields (but see Kaardal, Fitzgerald, Berry, & Sharpee, 2013).

Discussion

Accuracy maximization analysis (AMA) is a supervised Bayesian method for task-specific dimensionality reduction; it returns the encoding filters (receptive fields) that select the stimulus features that provide the most useful information about the task-relevant latent variable (Geisler et al., 2009). In conjunction with carefully collected databases of natural images and scenes and psychophysical experimental techniques, AMA has contributed to the development of ideal observers for several fundamental sensory-perceptual tasks in early- and mid-level vision (Geisler et al., 2009; Burge & Geisler, 2011, 2014, 2015). Unfortunately, AMA's compute time is high enough to render the method impractical for large problems without specialized computing resources.

We have developed AMA–Gauss, which makes the assumption that the class-conditional filter responses are Gaussian distributed and have shown that AMA–Gauss markedly reduces compute-time without compromising performance when the assumption is justified. We show that the assumption is justified for two fundamentally important visual tasks with natural stimuli (see Figure 5 and Figure 6; Burge & Geisler, 2014, 2015). These results provide a normative explanation for why energy model-like computations have proven useful in the study of motion and disparity estimation and discrimination. We speculate that the assumption will prove justified for other energy-model-related tasks in early vision (e.g., motion-in-depth estimation). AMA–Gauss also has the same formal structure as the generalized quadratic model (GQM) a recently developed method for neural systems identification, raising the possibility that a single framework could be used both to predict and estimate the properties of involved in particular tasks.

There are several important implications of these results. First, the optimal filters and the optimal pooling rules for decoding the latent variable are all determined by the properties of natural stimuli. If the training sets are representative of stimuli encountered in natural viewing, then the computations reported here should be optimal for the tasks of speed and disparity estimation. Second, at the right level of abstraction, the optimal solutions to these two different tasks share deep similarities, thereby raising the possibility that the same normative computational framework will apply to all energy-model related tasks.

Response distributions: Gaussian vs. heavy-tailed

The results reported here may appear to conflict with the widely reported finding that receptive field responses to natural images are highly non-Gaussian, with heavy tails sharp peaks at zero (Cadieu & Olshausen, 2012; Field, 1987; Olshausen & Field, 1997). There are two explanations for this apparent discrepancy. First, previous analyses generally have not incorporated contrast normalization. Second, previous analyses are generally unsupervised and therefore do not condition on relevant latent variables (e.g. motion; Cadieu & Olshausen, 2012). Note that even when contrast normalization is incorporated and the class-conditional responses are Gaussian, the filter responses, marginalized over the latent variable, tend to be heavy-tailed because the marginals are mixtures of Gaussians $\mathcal{P}(R_t) = \sum_u \mathcal{P}(R_t|X_u)\mathcal{P}(X_u)$ with different variances (see black curves in Figure 5E, F and Figure 6E, F). Therefore, our results are more similar to previous results than it may appear at first glance (Ruderman & Bialek, 1994). In general, heavy-tailed response distributions are yielded by response models that do not incorporate biologically plausible contrast normalization and response analyses that do not include latent variable conditionalization (Lyu & Simoncelli, 2009; Wang et al., 2004). Incorporating response normalization and latent variable conditionalization, as we have here, may help reveal statistical properties of receptive field responses to complex natural stimuli that have not yet been fully appreciated.

Likelihood functions: Data-constrained vs assumed

Evolution selects organisms because they perform certain critical sensory, perceptual, and behavioral tasks better than their evolutionary competitors. Certain features of sensory stimuli are more useful for some tasks than others. The stimulus features that are most useful to encode thus depend on the task-relevant latent variables that will be decoded from the stimuli. However, many models of neural encoding do not explicitly consider the tasks for which the encoded information will be decoded (Olshausen & Field, 1997; Simoncelli & Olshausen, 2001) and many task-specific models of neural decoding do not explicitly consider how sensory stimuli and neural encoders constrain the information available for decoding (Ernst & Banks, 2002; Ma, Beck, Latham, & Pouget, 2006).

The approach advanced here is an early attempt to address both issues simultaneously. By performing task-specific analyses using thousands of individual natural stimuli, learning the optimal filters, and characterizing the class-conditional responses to natural stimuli, we determined the likelihood functions that optimize performance in natural viewing. The likelihood functions that result from the filter response distributions are (on average) log-Gaussian in speed and disparity, with widths that increase with the value of the latent variable. In previous work with natural stimuli, we showed that the optimal receptive fields, response distributions, and resulting likelihood functions are robust to significant variation in the shape of the prior, cost function, and noise power (Burge & Jaini, 2017). It is reasonable to conclude that the task and the constraints imposed by natural stimuli are the most important determinants of the width and shape of the likelihood functions.

Some prominent theories of neural processing operate on the assumption that likelihood functions can take on arbitrary widths and shapes via flexible allocation of neural resources (Ganguli & Simoncelli, 2014; Girshick, Landy, & Simoncelli, 2011; Seung & Sompolinsky, 1993; Wei & Stocker, 2015). Some reports have gone further to suggest that, in the context of Bayesian efficient coding, the prior probability distribution over the latent variable is the primary factor determining the widths and shapes of the likelihood functions (Ganguli & Simoncelli, 2014; Wei & Stocker, 2015). These reports predict that if the prior probability distribution is flat, the likelihood functions will be symmetric and have widths that remain constant with changes in the value of the latent variable. These reports also predict that if the prior probability distribution is nonuniform (e.g., peaked at zero), the likelihood functions will be asymmetric with widths that change systematically with the latent variable.

In the tasks that we examined, we found that asymmetric likelihood functions optimize performance despite the fact that the training sets from which the optimal filters were learned had flat priors over the latent variable (see Results; Burge & Geisler, 2014, 2015; Burge & Jaini, 2017). These results appear at odds with the predictions of previous reports. However, these previous reports do not model the impact of natural stimulus variation. The implicit assumption is that task-irrelevant ("nuisance") stimulus variation can be ignored (Ganguli & Simoncelli, 2014; Wei & Stocker, 2015). If the goal is to understand optimal task-specific processing of natural signals, our results indicate that such variation cannot be ignored. Indeed, task-relevant and irreducible task-irrelevant natural stimulus variability are almost certainly the most important determinants of likelihood function shapes and widths.

In natural viewing, visual estimates are driven primarily by stimulus measurements (likelihood functions), not by prior distributions. If estimates were driven only by the prior, observers could not respond to spatial or temporal changes in the environment. A full account of task-specific perceptual processing and its underlying neurophysiology must therefore incorporate natural stimulus variability. Future studies on the efficient allocation of neural resources should verify that the likelihood functions used in modeling efforts can be constructed by nervous systems given the constraints imposed by natural stimuli.

Natural vs. artificial stimuli

The problem of estimating speed and disparity from natural images is different from the problem of estimating speed and disparity with artificial laboratory stimuli in at least one important respect. Variability among natural stimuli having the same latent variable level is substantially greater than variability amongst artificial stimuli commonly used in vision and visual neuroscience experiments. In motion experiments (Figure 11A), drifting Gabors and random-dot kinematograms are common artificial stimuli. In disparity experiments, phase-shifted binocular Gabors and random-dot stereograms are common artificial stimuli (Figure 11B). The statistical properties of these artificial stimuli are notably different than the statistical properties of natural stimuli. Gabors have Gaussian amplitude spectra and random-dot stereograms have delta auto-correlation functions. Natural stimuli have a rich variety of textures and shapes that cause significant variation in their 1/f amplitude spectra and auto-correlation functions.

To examine the impact of artificial stimuli on the class-conditional responses, we created artificial stimulus sets comprised of contrast-fixed, phase-randomized Gabors drifting at different speeds and having different amounts of disparity. For each task, the spatial frequency of the carrier was closely matched to the preferred spatial frequency of the first two optimal filters (1.0 cpd for speed, 1.5 cpd for disparity). Joint filter responses to these artificial stimuli are shown in Figure 11C, D; they are notably different than the filter responses to natural stimuli. Although the class-conditional responses to Gabors are approximately aligned with the major axis of the Gaussian characterizing responses to corresponding natural stimuli, the responses themselves are no longer Gaussian distributed, exhibiting ring-shaped structure instead. Thus, determining the optimal rules for processing natural stimuli by analyzing only artificial stimuli is likely to be a difficult enterprise.

These results suggest another conclusion that may be somewhat counterintuitive given the history of the field. The tradition in vision science has been to eliminate irrelevant stimulus variation from experimental protocols by using simple artificial stimuli. These stimuli are easy to characterize mathematically and manipulate parametrically. But artificial stimuli lack the richness and variability that visual systems evolved to process. Analyzing complex, variable natural stimuli may reveal simple (e.g., Gaussian) statistical structure that might otherwise be missed. We believe that the results presented here highlight the importance of conducting rigorous, well-controlled, task-focused computational and behavioral investigations with natural stimuli. These investigations complement classic studies with artificial stimuli, and provide a fuller picture of how visual systems function in natural circumstances.

Limitations and future directions

The results presented here represent the first in what we hope is a series of steps to link normative models for natural tasks and descriptive models of neural re-

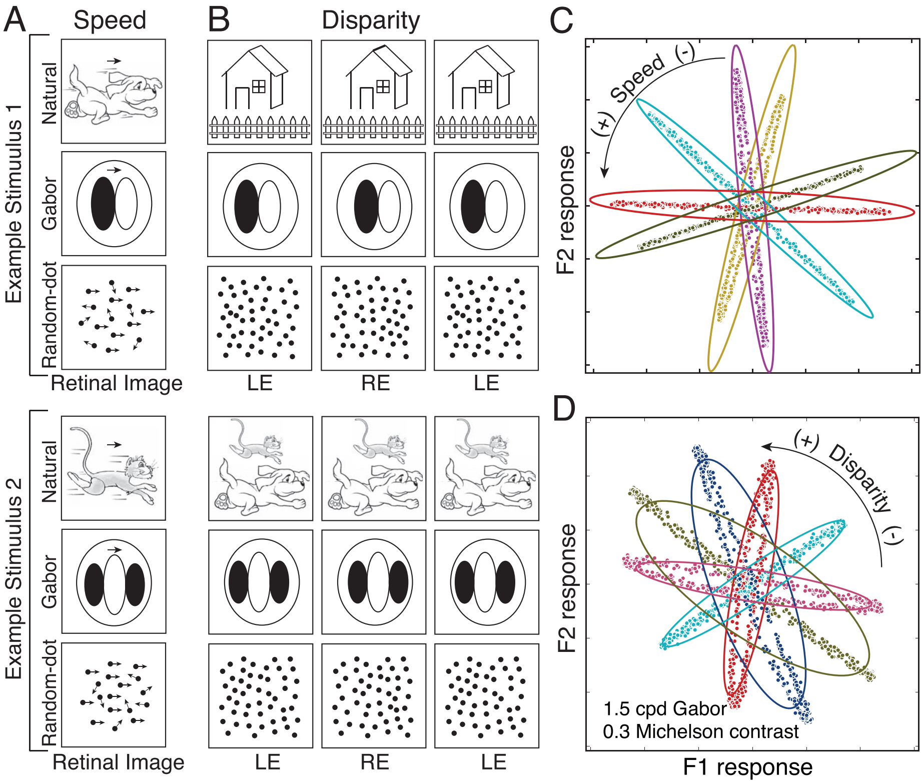

Figure 11. Natural stimuli, artificial stimuli, and class-conditional responses. Many different retinal images are consistent with a given value of the task-relevant latent variable. These differences cause within-class (task-irrelevant) stimulus variation. Within-class stimulus variation is greater for natural stimuli than for typical artificial stimuli used in laboratory experiments. (A) Stimuli for speed estimation experiments. Two different example stimuli are shown for each stimulus type: natural stimuli (represented by cartoon line drawings), Gabor stimuli, and random-dot stimuli. Both example stimuli for each stimulus type drift at exactly the same speed, but create different retinal images. Natural stimuli cause more within-class retinal stimulus variation than artificial stimuli. (B) Same as (A), but for disparity. (C) Speed task: class-conditional responses to contrast-fixed 1.0 cpd drifting Gabors with random phase (speed task). Colors indicate different speeds. Ellipses represent filter responses to natural stimuli having the same speeds. (D) Disparity task: Class-conditional responses to contrast-fixed 1.5 cpd binocular Gabors with random phase. Class-conditional responses no longer have Gaussian structure, and instead have ring structure.

sponse. However, although we believe that developing AMA-Gauss and demonstrating its links to methods for neural systems identification are useful advances, several limitations should be kept in mind. Here, we address the drawbacks of the natural stimulus sets, the general applicability of AMA–Gauss, and the importance of the links that we have drawn to descriptive models of neural response.

The natural image sets used in this manuscript had natural contrast distributions and photographic textures, but they lacked natural depth structure. All motion signals were rigid and all disparity signals were planar. Future work will examine the impact of nonrigid motion (e.g., looming) and local depth variation (e.g., occlusion) on performance. We have recently collected a dataset of stereo images that addresses this limitation (Burge, McCann, & Geisler, 2016). Each stereo image has co-registered distance data from which ground truth disparity patterns can be computed. Pilot analyses suggest that the results presented in the current manuscript hold for natural stereo images with local depth variation. We suspect, but we are not yet well-positioned to show, that the same will be true of motion signals having natural depth variation.

AMA–Gauss is the appropriate normative framework for understanding energy-model-related tasks, but the general usefulness of AMA–Gauss is unknown. AMA–Gauss makes the best possible use of the first- and second-order filter response statistics, but it is blind to higher-order response statistics that may exist in natural motion (Nitzany & Victor, 2014) and natural disparity signals (see Appendix D). To increase generality, one could develop a further variant of AMA that incorporates rectification into the response model. This modification would confer the ability, at least in principle, to pick up on potentially useful higher-order motion and disparity cues, and provide a normative model that complements other methods for neural systems identification (McFarland et al., 2013).

Conclusion

In this paper, we develop AMA–Gauss, a new form of AMA that incorporates the assumption that the class-conditional filter responses are Gaussian distributed. We use AMA–Gauss to establish links between task-specific normative models of speed and disparity estimation and the motion- and disparity-energy models, two popular descriptive models of neurons that are selective for those quantities. Our results suggest that energy-model-like (i.e., quadratic) computations are optimal for these tasks in natural scenes. We also establish the formal similarities between AMA–Gauss and the generalized quadratic model (GQM), a recently developed model for neural systems identification. The developments presented here forge links between normative task-specific modeling and powerful statistical tools for describing neural response, and demonstrate the importance of analyzing natural signals in perception and neuroscience research.

Keywords: normative model, neural systems identification, accuracy maximization analysis, energy model, generalized quadratic model, contrast normalization, natural scene statistics, quadratic computations, speed, disparity

Acknowledgments

P. J. was supported by the Cheriton Graduate Scholarship and startup funds from the University of Pennsylvania. J. B. was supported by National Institutes of Health Grant EY011747 and startup funds from the University of Pennsylvania.

Commercial relationships: none.

Corresponding author: Johannes Burge.
Email: jburge@sas.upenn.edu.
Address: Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA.

Footnote

[1] AMA–Gauss software (MATLAB) is available at https://www.github.com/BurgeLab/AMA.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*(2), 284–299.

Albrecht, D. G., & Geisler, W. S. (1991). Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Visual Neuroscience*, *7*(6), 531–546.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Burge, J., & Geisler, W. S. (2011). Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences, USA*, *108*(40), 16849–16854.

Burge, J., & Geisler, W. S. (2012). Optimal defocus estimates from individual images for autofocusing a digital camera. In *Proceedings of the IS&T/SPIE 47th annual meeting*. Proceedings of SPIE.

Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision*, *14*(2):1, 1–18, doi:10.1167/14.2.1. [PubMed] [Article]

Burge, J., & Geisler, W. S. (2015). Optimal speed estimation in natural image movies predicts human performance. *Nature Communications*, *6*, 7900.

Burge, J., & Jaini, P. (2017). Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS Computational Biology*, *13*(2), e1005281.

Burge, J., McCann, B. C., & Geisler, W. S. (2016). Estimating 3d tilt from local image cues in natural scenes. *Journal of Vision*, *16*(13):2, 1–25, doi:10.1167/16.13.2. [PubMed] [Article]

Cadieu, C. F., & Olshausen, B. A. (2012). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation*, *24*(4), 827–866.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62.

Coen-Cagli, R., Dayan, P., & Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Computational Biology*, *8*(3), e1002405.

Cook, R., Forzani, L., & Yao, A. (2010). Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression. *Statistica Sinica*, *20*(1), 235–238.

Cook, R. D., & Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the*

American Statistical Association, *104*(485), 197–208.

Cumming, B., & DeAngelis, G. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, *24*(1), 203–238.

DeAngelis, G. C. (2000). Seeing in three dimensions: the neurophysiology of stereopsis. *Trends in Cognitive Sciences*, *4*(3), 80–90.

Eickenberg, M., Rowekamp, R. J., Kouh, M., & Sharpee, T. O. (2012). Characterizing responses of translation-invariant neurons to natural stimuli: maximally informative invariant dimensions. *Neural Computation*, *24*(9), 2384–2421.

Emerson, R. C., Bergen, J. R., & Adelson, E. H. (1992). Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research*, *32*(2), 203–218.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and Image Science*, *4*(12), 2379–2394.

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural computation*, *26*(10), 2103–2134.

Gao, D., & Vasconcelos, N. (2009). Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, *21*(1), 239–271.

Geisler, W. S., & Albrecht, D. G. (1997). Visual cortex neurons in monkeys and cats: Detection, discrimination, and identification. *Visual Neuroscience*, *14*(5), 897–919.

Geisler, W. S., Najemnik, J., & Ing, A. D. (2009). Optimal stimulus encoders for natural tasks. *Journal of Vision*, *9*(13):17, 1–16, doi:10.1167/9.13.17. [PubMed] [Article]

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(02), 181–197.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychol*ogy, *24*(6), 417–441.

Jones, J. P., & Palmer, L. A. (1987a). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1233–1258.

Jones, J. P., & Palmer, L. A. (1987b). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1187–1211.

Kaardal, J., Fitzgerald, J. D., Berry, M. J., & Sharpee, T. O. (2013). Identifying functional bases for multidimensional neural computations. *Neural Computation*, *25*(7), 1870–1890.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Lyu, S., & Simoncelli, E. P. (2009). Modeling multi-scale sub-bands of photographic images with fields of Gaussian scale mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 693–706.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

McFarland, J. M., Cui, Y., & Butts, D. A. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Computational Biology*, *9*(7), e1003143.

Nitzany, E. I., & Victor, J. D. (2014). The statistics of local motion signals in naturalistic movies. *Journal of Vision*, *14*(4):10, 1–15, doi:10.1167/14.4.10. [PubMed] [Article]

Nover, H., Anderson, C. H., & DeAngelis, G. C. (2005). A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *The Journal of Neuroscience*, *25*(43), 10049–10060.

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: The disparity energy model. *Current Opinion in Neurobiology*, *8*(4), 509–515.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, *249*(4972), 1037–1041.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *Journal of Neurophysiology*, *77*(6), 2879–2909.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, *37*(23), 3311–3325.

Pagan, M., Simoncelli, E. P., & Rust, N. C. (2016). Neural quadratic discriminant analysis: Nonlinear decoding with V1-like computation. *Neural Computation*, *29*, 2291–2319.

Park, I. M., Archer, E. W., Priebe, N., & Pillow, J. W. (2013). Spectral methods for neural characterization using generalized quadratic models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 2454–2462). Red Hook, NY: Curran Associates.

Peng, Q., & Shi, B. E. (2010). The changing disparity energy model. *Vision Research*, *50*(2), 181–192.

Petersen, K. B., & Pedersen, M. S. (2008). The matrix cookbook. *Technical University of Denmark*, *7*, 15.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, *73*(6), 814–817.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, *46*(6), 945–956.

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, *6*(4):13, 484–507, doi:10.1167/6.4.13. [PubMed] [Article]

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences, USA*, *90*(22), 10749–10753.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

Tanabe, S., Haefner, R. M., & Cumming, B. G. (2011). Suppressive mechanisms in monkey V1 help to solve the stereo correspondence problem. *Journal of Neuroscience*, *31*(22), 8295–8305.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611–622.

Vintch, B., Movshon, J. A., & Simoncelli, E. P. (2015). A convolutional subunit model for neuronal responses in macaque v1. *The Journal of Neuroscience*, *35*(44), 14829–14841.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509–1517.

Williams, D. R. (1985). Visibility of interference fringes near the resolution limit. *Journal of the Optical Society of America A*, *2*(7), 1087–1093.

Wu, A., Park, I. M., & Pillow, J. W. (2015). Convolutional spike-triggered covariance analysis for neural subunit models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 793–801). Red Hook, NY: Curran Associates.

Appendix A: Gradient of the likelihood function

In any given training set having N stimuli, each stimulus is associated with some category k and an associated stimulus from that category l. Let us denote this pair (k, l) for the i^{th} sample point with (k_i, l_i). Then assuming that the response distribution conditioned on the classes is Gaussian, the likelihood function can be written as

$$\mathcal{L}(\mathbf{f}) = \prod_{i=1}^{N}(2\pi)^{-\frac{d}{2}}|\Sigma_{k_i}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\mathbf{R}(k_i,l_i)-\mu_{k_i}\right)^T\Sigma_{k_i}^{-1}\left(\mathbf{R}(k_i,l_i)-\mu_{k_i}\right)\right]$$

Substituting the expression for the noisy responses (Equation 5) and defining $l(\mathbf{f}) = \log\mathcal{L}(\mathbf{f})$ yields the log-likelihood function of the AMA–Gauss filters

$$l(\mathbf{f}) = \zeta_u - \frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{f}^T B_{k_i}\mathbf{f}+\Lambda| + \left(\mathbf{f}^T\mathbf{s}_{k_il_i}-\mathbf{f}^T\mathbf{s}_{k_i}+\eta\right)^T\left(\mathbf{f}^T B_{k_i}\mathbf{f}+\Lambda\right)^{-1}\left(\mathbf{f}^T\mathbf{s}_{k_il_i}-\mathbf{f}^T\mathbf{s}_{k_i}+\eta\right)$$

where $\mathbf{s}_{k_i} = \frac{1}{N_{k_i}}\sum_{m_i=1}^{N_{k_i}} \mathbf{s}_{k_i m_i}$ and $B_{k_i} = \frac{1}{N_{k_i}}\sum_{m_i=1}^{N_{k_i}} (\mathbf{s}_{k_i,m_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i,m_i} - \mathbf{s}_{k_i})^T$ are the class-conditional stimulus mean and covariance matrix, respectively, and $\zeta_u = -\frac{1}{2}\log|2\pi\Sigma_u|$ is a constant.

Rearranging to segregate terms that do not depend on noise samples

$$l(\mathbf{f}) = \zeta_u - \frac{1}{2}\sum_{i=1}^{N} \big[\log|\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda| + (\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T \mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i}) \\ + \eta^T\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i}) + (\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\eta + \eta^T\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\eta\big] \quad (29)$$

where $\mathbf{f}^T B\mathbf{f} + \Lambda$ is a symmetric matrix. Recognizing that each term in Equation 29 is a scalar, and rewriting using the properties that $Tr(a) = a$, $Tr(\mathbf{AB}) = Tr(\mathbf{BA})$ and $Tr(\mathbf{A}) = Tr(\mathbf{A}^T)$ yields

$$l(\mathbf{f}) = \zeta_u - \frac{1}{2}\sum_{i=1}^{N}\Bigg[\log|\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda| + Tr\Big(\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\Big) \\ + 2Tr\Big((\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\eta\Big) + Tr\Big(\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\eta\eta^T\Big)\Bigg] \quad (30)$$

To determine the gradient of the log-likelihood$\nabla_{\mathbf{f}} l(\mathbf{f})$, we derive the gradient of each term in Equation 30 separately as follows. Before doing so, we state some standard matrix results that will be used in the derivation (Petersen, Pedersen et al., 2008).

$$\frac{\partial \log(det(\mathbf{X}))}{\partial \mathbf{X}} = \big(\mathbf{X}^T\big)^{-1} \quad (31)$$

$$\frac{\partial}{\partial \mathbf{X}} Tr\Big(\big(A + \mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{B}\mathbf{X}\Big) = -2\mathbf{C}\mathbf{X}\big(A + \mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{B}\mathbf{X}\big(A + \mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1} + 2\mathbf{B}\mathbf{X}\big(A + \mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1} \quad (32)$$

$$\frac{\partial}{\partial \mathbf{X}} Tr\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{A} = -\mathbf{C}\mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\big(\mathbf{A} + \mathbf{A}^T\big)\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1} \quad (33)$$

The gradient of the first term in Equation 30 is obtained by using Equation 31 and the chain rule of differentiation

$$\begin{aligned} \nabla_{\mathbf{f}} \log|\overbrace{\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda}^{\mathbf{Y}}| &= \frac{\partial \log|\mathbf{Y}|}{\partial \mathbf{Y}} \frac{\partial\left(\overbrace{\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda}^{\mathbf{Y}}\right)}{\partial \mathbf{f}} \\ \nabla_{\mathbf{f}} \log|\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda| &= 2B_{k_i}\mathbf{f}\big(\mathbf{f}B_{k_i}\mathbf{f}^T + \Lambda\big)^{-1} \quad (34) \end{aligned}$$

The gradient of the second term in Equation 30 is obtained using Equation 32

$$\nabla_{\mathbf{f}} Tr\Big(\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\Big) \\ = 2(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1} - 2B_{k_i}\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1} \\ (35)$$

The gradient of the third term is obtained using Equation 33 and the chain rule of differentiation

$$\nabla_{\mathbf{f}} 2Tr\Big((\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\eta\Big) = 2(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})\eta^T\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1} \\ + 2B_{k_i}\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\Big(\eta(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f} + \mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})\eta^T\Big)\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1} \\ (36)$$

The gradient of the fourth term is similarly obtained using Equation 33

$$\nabla_{\mathbf{f}} Tr\Big(\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\eta\eta^T\Big) = -4B_{k_i}\mathbf{f}\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1}\big(\eta\eta^T\big)\big(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\big)^{-1} \quad (37)$$

The full gradient of the AMA–Gauss filter log-likelihood $l(\mathbf{f})$ stated in Equation 30 can therefore be found by combining Equations 34–37.

The gradient of the expected log-likelihood follows directly from the gradient of the log-likelihood. The response noise $\eta \sim \mathcal{N}(\mathbf{0}, \Lambda)$ is normally distributed (Equation 6); therefore, $E_\eta[\eta^T(\mathbf{f}^T B_{k_i}\mathbf{f})^{-1}\eta] = Tr((\mathbf{f}^T B_{k_i}\mathbf{f})^{-1}\Lambda)$. Substituting into Equation 30 yields the expected log-likelihood of the AMA–Gauss filters

$$E_\eta[l(\mathbf{f})] = \zeta_u - \frac{1}{2}\sum_{i=1}^{N}\left[\log|\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda| - (\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i}) - Tr\left(\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1}\Lambda\right)\right] \tag{38}$$

The gradient of the expected log-likelihood, using Equations 34, 35, and 37, is given by

$$\begin{aligned}\nabla_{\mathbf{f}} E_\eta[l(\mathbf{f})] = -\sum_{i=1}^{N}\Big[&B_{k_i}\mathbf{f}\left(\mathbf{f}B_{k_i}\mathbf{f}^T + \Lambda\right)^{-1} - B_{k_i}\mathbf{f}\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1}\mathbf{f}^T(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1} \\ &+ (\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})(\mathbf{s}_{k_i l_i} - \mathbf{s}_{k_i})^T\mathbf{f}\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1} - \frac{1}{2}B_{k_i}\mathbf{f}\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1}\left(\Lambda + \Lambda^T\right)\left(\mathbf{f}^T B_{k_i}\mathbf{f} + \Lambda\right)^{-1}\Big]\end{aligned} \tag{39}$$

Appendix B: Gradient of L_2 cost function

The average expected cost across all the stimuli is

$$\bar{C} = \frac{1}{N}\sum_{k,l}\bar{C}_{kl} \tag{40}$$

Given the squared error loss function, the expected cost per stimuli can be written as

$$\bar{C}_{kl} = E_{\mathbf{R}(k,l)}\left[\left(\hat{X}_{kl}^{opt} - X_k\right)^2\right] \tag{41}$$

where $\hat{X}_{kl}^{\text{opt}} = \sum_{u=1}^{N_{lvl}} X_u\mathcal{P}(X_u|\mathbf{R}(k,l))$ since the optimal estimate for a squared error function is the mean of the posterior, i.e., $E[X_u|\mathbf{R}(k,l)]$. Using the approximation that the expected cost of each stimulus is equal to the cost given the expected response (Geisler et al., 2009) yields

$$\bar{C}_{kl} \cong \left(\sum_{u=1}^{N_{lvl}} X_u\mathcal{P}(X_u|\mathbf{r}(k,l)) - X_k\right)^2 \tag{42}$$

Therefore, to evaluate the gradient of the total cost we just need to evaluate the expression for the gradient of the expected cost of each stimulus. Hence,

$$\nabla_{\mathbf{f}_q}\bar{C}_{kl} = \nabla_{\mathbf{f}_q}\left(\hat{X}_{kl}^{opt} - X_k\right)^2 = 2\left(\hat{X}_{kl}^{opt} - X_k\right)\nabla_{\mathbf{f}_q}\hat{X}_{kl}^{opt} \tag{43}$$

The gradient of the optimal estimate given the mean response is

$$\nabla_{\mathbf{f}_q}\hat{X}_{kl}^{opt} = \sum_{u=1}^{N_{lvl}} X_u\left[\nabla_{\mathbf{f}_q}\mathcal{P}(X_u|\mathbf{r}(k,l))\right] \tag{44}$$

Hence, the problem reduces to finding $[\nabla_{\mathbf{f}_q}\mathcal{P}(X_u|\mathbf{r}(k,l))]$

$$\mathcal{P}(X_u|\mathbf{r}(k,l)) = \frac{\mathcal{N}(\mathbf{r}(k,l); \mu_u, \Sigma_u)}{\sum_{i=1}^{N_{lvl}}\mathcal{N}(\mathbf{r}(k,l); \mu_i, \Sigma_i)} \tag{45}$$

Making substitutions in Equation 45 gives

$$\mathcal{P}(X_u|\mathbf{r}(k,l)) \;= \frac{|\Sigma_u|^{-0.5}\exp\left[-0.5(\mathbf{r}(k,l)-\mu_u)^T\Sigma_u^{-1}(\mathbf{r}(k,l)-\mu_u)\right]}{\sum_{i=1}^{N_{lvl}}|\,\Sigma_i|^{-0.5}exp\left[-0.5(\mathbf{r}(k,l)-\mu_i)^T\Sigma_i^{-1}(\mathbf{r}(k,l)-\mu_i)\right]} \quad (46)$$

$$= \frac{|\mathbf{f}^TB_u\mathbf{f}+\Lambda|^{-0.5}\exp\left[-0.5\mathbf{A}^T_{kl,u}\mathbf{f}\left(\mathbf{f}^TB_u\mathbf{f}+\Lambda\right)^{-1}\mathbf{f}^T\mathbf{A}_{kl,u}\right]}{\sum_{i=1}^{N_{lvl}}|\,\mathbf{f}^TB_i\mathbf{f}+\Lambda|^{-0.5}\exp\left[-0.5\mathbf{A}^T_{kl,i}\mathbf{f}\left(\mathbf{f}^TB_i\mathbf{f}+\Lambda\right)^{-1}\mathbf{f}^T\mathbf{A}_{kl,i}\right]} \quad (47)$$

where $\mathbf{A}_{kl,u}=\mathbf{s}_{kl}-\mathbf{s}_u$. The gradient of the posterior probability can then be evaluated using the following relation with the gradient of the logarithm of the posterior probability

$$\nabla_{\mathbf{f}_q}\mathcal{P}(X_u|\mathbf{r}(k,l)) = \mathcal{P}(X_u|\mathbf{r}(k,l))\left[\nabla_{\mathbf{f}_q}\log\mathcal{P}(X_u|\mathbf{r}(k,l))\right] \quad (48)$$

Taking the natural logarithm of the posterior yields

$$\log\mathcal{P}(X_u|\mathbf{r}(k,l)) = -\log\sum_{i=1}^{N_{lvl}}\frac{|\mathbf{f}^TB_u\mathbf{f}+\Lambda|^{0.5}}{|\mathbf{f}^TB_i\mathbf{f}+\Lambda|^{0.5}}\exp\left[\frac{1}{2}\left(\mathbf{A}^T_{kl,u}\mathbf{f}\left(\mathbf{f}^TB_u\mathbf{f}+\Lambda\right)^{-1}\mathbf{f}^T\mathbf{A}_{kl,u}-\mathbf{A}^T_{kl,i}\mathbf{f}\left(\mathbf{f}^TB_i\mathbf{f}+\Lambda\right)^{-1}\mathbf{f}^T\mathbf{A}_{kl,i}\right)\right] \quad (49)$$

Next, we define new variables to simplify this expression for the log posterior probability and the subsequent derivation of its gradient. Let each term in the summation in Equation 49 be

$$Z_i(u,k,l,\mathbf{f}) = T_i(u,k,l,\mathbf{f})\exp[1/2U_i(u,k,l,\mathbf{f})] \quad (50)$$

where $T_i(u,k,l,\mathbf{f}) = \frac{|\mathbf{f}^TB_u\mathbf{f}+\Lambda|^{0.5}}{|\mathbf{f}^TB_i\mathbf{f}+\Lambda|^{0.5}}$ is the scale factor in each term in the summation in Equation 50 and where $U_i(u,k,l,\mathbf{f}) = \mathbf{A}^T_{kl,u}\mathbf{f}(\mathbf{f}^TB_u\mathbf{f}+\Lambda)^{-1}\mathbf{f}^T\mathbf{A}_{kl,u}-\mathbf{A}^T_{kl,i}\mathbf{f}(\mathbf{f}^TB_i\mathbf{f}+\Lambda)^{-1}\mathbf{f}^T\mathbf{A}_{kl,i}$ is the exponentiated term in each term of the sum in Equation 50. Hence, by substituting Equation 50 into Equation 49 the simplified expression for the log posterior is

$$\log\mathcal{P}(X_u|\mathbf{r}(k,l)) = -\log\sum_{i=1}^{N_{lvl}}Z_i(u,k,l,\mathbf{f}) \quad (51)$$

The gradient of the log posterior probability can therefore be expressed as

$$\nabla_{\mathbf{f}}\log\mathcal{P}(X_u|\mathbf{r}(k,l)) = \nabla_{\mathbf{f}}\left(-\log\sum_{i=1}^{N_{lvl}}Z_i(u,k,l,\mathbf{f})\right) \quad (52)$$

The gradient of the log is

$$\nabla_{\mathbf{f}}\log\mathcal{P}(X_u|\mathbf{r}(k,l)) = \frac{\sum_{i=1}^{N_{lvl}}\nabla_{\mathbf{f}}Z_i(u,k,l,\mathbf{f})}{\sum_{i=1}^{N_{lvl}}Z_i(u,k,l,\mathbf{f})} \quad (53)$$

Expanding the numerator by substituting Equation 50 using the chain rule for differentiation

$$\nabla_{\mathbf{f}}\log\mathcal{P}(X_u|\mathbf{r}(k,l)) = -\frac{1}{\sum_{v=1}^{N_{lvl}}Z_v(u,k,l,\mathbf{f})}\sum_{i=1}^{N_{lvl}}\Bigg(\exp[(1/2U_i(u,k,l,\mathbf{f})]\nabla_{\mathbf{f}}T_i(u,k,l,\mathbf{f}) + \frac{1}{2}T_i(u,k,l,\mathbf{f})\exp[1/2U_i(u,k,l,\mathbf{f})]\nabla_{\mathbf{f}}U_i(u,k,l,\mathbf{f})\Bigg) \quad (54)$$

The remaining terms to be evaluated are $\nabla_{\mathbf{f}}T_i(u,k,l,\mathbf{f})$ and$\nabla_{\mathbf{f}}U_i(u,k,l,\mathbf{f})$.

The expression for $\nabla_{\mathbf{f}} T_i(u,k,l,\mathbf{f})$ is

$$\nabla_{\mathbf{f}} T_i(u,k,l,\mathbf{f}) = \nabla_{\mathbf{f}} \frac{|\mathbf{f}^T B_u \mathbf{f} + \Lambda|^{0.5}}{|\mathbf{f}^T B_i \mathbf{f} + \Lambda|^{0.5}} = \frac{|\mathbf{f}^T B_u \mathbf{f} + \Lambda|^{0.5} |\mathbf{f}^T B_i \mathbf{f} + \Lambda|^{0.5} \left(\left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} B_u \mathbf{f} - \left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} B_i \mathbf{f} \right)}{|\mathbf{f}^T B_i \mathbf{f} + \Lambda|}$$
$$= \frac{|\mathbf{f}^T B_u \mathbf{f} + \Lambda|^{0.5}}{|\mathbf{f}^T B_i \mathbf{f} + \Lambda|^{0.5}} \left(B_u \mathbf{f} \left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} - B_i \mathbf{f} \left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} \right) \quad (55)$$

The expression for $\nabla_{\mathbf{f}} U_i(u,k,l,\mathbf{f})$ is

$$U_i(u,k,l,\mathbf{f}) = \mathrm{Tr}\left(\mathbf{A}_{kl,u}^T \mathbf{f} \left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T \mathbf{A}_{kl,u} - \mathbf{A}_{kl,i}^T \mathbf{f} \left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T \mathbf{A}_{kl,i} \right)$$
$$= \mathrm{Tr}\left(\mathbf{A}_{kl,u}^T \mathbf{f} \left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T \mathbf{A}_{kl,u} \right) - \mathrm{Tr}\left(\mathbf{A}_{kl,i}^T \mathbf{f} \left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T \mathbf{A}_{kl,i} \right)$$
$$= \mathrm{Tr}\left(\left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T \mathbf{A}_{kl,u} \mathbf{A}_{kl,u}^T \mathbf{f} \right) - \mathrm{Tr}\left(\left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T \mathbf{A}_{kl,i} \mathbf{A}_{kl,i}^T \mathbf{f} \right)$$
$$= \mathrm{Tr}\left(\left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T D_{kl,u} \mathbf{f} \right) - \mathrm{Tr}\left(\left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T D_{kl,i} \mathbf{f} \right) \nabla_{\mathbf{f}} U_i(u,k,l,\mathbf{f})$$
$$= \nabla_{\mathbf{f}} \mathrm{Tr}\left(\left(\mathbf{f}^T B_u \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T D_{kl,u} \mathbf{f} \right) - \nabla_{\mathbf{f}} \mathrm{Tr}\left(\left(\mathbf{f}^T B_i \mathbf{f} + \Lambda\right)^{-1} \mathbf{f}^T D_{kl,i} \mathbf{f} \right) \quad (56)$$

where $D_{kl,u} = \mathbf{A}_{kl,u} \mathbf{A}_{kl,u}^T$. The expression for the gradient of the trace in Equation 56 is obtained by using Equation 32. Thus,

$$\nabla_{\mathbf{f}} Tr((\mathbf{f}^T B_u \mathbf{f} + \Lambda)^{-1} \mathbf{f}^T D_{kl,u} \mathbf{f}) = -2 B_u \mathbf{f} (\mathbf{f}^T B_u \mathbf{f} + \Lambda)^{-1} \mathbf{f}^T D_{kl,u} \mathbf{f} (\mathbf{f}^T B_u \mathbf{f} + \Lambda)^{-1} + 2 D_{kl,u} \mathbf{f} (\mathbf{f}^T B_u \mathbf{f} + \Lambda)^{-1} \quad (57)$$

The gradient $\nabla_{\mathbf{f}} U_i(u,k,l,\mathbf{f})$ is obtained by substituting Equation 57 into Equation 56. The gradient of $\log \mathcal{P}(X_u|\mathbf{r}(k,l))$ is obtained by substituting Equation 55 and Equation 56 into Equation 54. The gradient of the posterior probability is obtained by plugging Equation 54 into Equation 48. The gradient of the cost for each stimulus is obtained by plugging Equation 48 into Equation 44, and then plugging that result into Equation 43.

Appendix C: AMA–Gauss gradient with L_0 / KL-divergence cost function

The total cost for a set of filters is given by the average expected cost across all stimuli

$$\bar{C} = \frac{1}{N} \sum_{k,l}^{N} E_{\mathbf{R}(k,l)}[C_{kl}] \quad (58)$$

Given the 0,1 cost function, the cost associated with the filter response to an arbitrary stimulus is given by$C_{kl} = 1 - \mathcal{P}(X_k|\mathbf{R}(k,l))$. This cost is monotonic with KL-divergence and we refer to this cost as the KL-cost.

$$C_{kl} = -\log \mathcal{P}(X_k|\mathbf{R}(k,l)) \quad (59)$$

We approximate the expected cost associated with each stimulus with the expected cost given the mean response (Geisler et al., 2009). Thus, we have

$$E_{\mathbf{R}(k,l)}[C_{kl}] = -\int_{-\infty}^{\infty} \log \mathcal{P}(X_k|\mathbf{R}(k,l)) \mathcal{P}(\mathbf{R}(k,l)|\mathbf{s}_{kl}) d\mathbf{R}(k,l) \quad (60)$$

$$\cong -\log \mathcal{P}(X_k|\mathbf{r}(k,l)) \quad (61)$$

Therefore, the total cost for a set of filters is given by

$$\bar{C} = -\frac{1}{N} \sum_{k,l}^{N} \log \mathcal{P}(X_k|\mathbf{r}(k,l)) \quad (62)$$

Hence, the gradient of the total expected cost $\bar{C}$ can then be written as

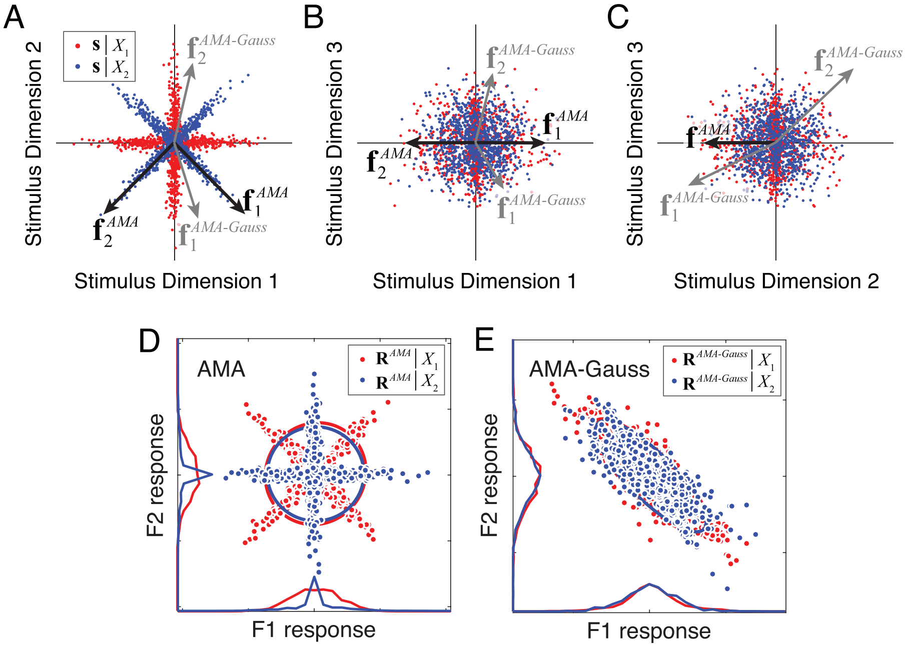

Figure A1. AMA vs. AMA–Gauss with simulated non-Gaussian class-conditional stimulus distributions. (A–C) Simulated stimuli from category 1 (red) and category 2 (blue). Each subplot plots two of the three stimulus dimensions against each other. All of the information for discriminating the categories is in the first two stimulus dimensions. AMA filters (black) and AMA–Gauss (gray) filters (which have the same dimensionality as the stimuli) are represented by arrows. AMA filters place all weight in the first two informative stimulus dimensions; that is, the two-dimensional subspace defined by the two AMA filters coincides with the most informative stimulus dimensions. AMA–Gauss filter weights are placed randomly across the three stimulus dimensions; that is, the two-dimensional subspace defined by the two AMA-Gauss filters will be random with respect to the most informative subspace. (D) Class-conditional AMA filter responses $\mathbf{R}^{AMA}$ allow the categories to be discriminated. (E) Class-conditional AMA–Gauss filter responses $\mathbf{R}^{AMA\text{–}Gauss}$ do not allow the categories to be discriminated.

$$\nabla_{\mathbf{f}}\bar{C} = \frac{1}{N}\sum_{k,l}^{N}\nabla_{\mathbf{f}_q}[\log\mathcal{P}(X_k|\mathbf{r}(k,l))] \quad (63)$$

The full expression for the expected cost $\bar{C}$ is obtained by substituting the expression for $\nabla_{\mathbf{f}_q}[\log\mathcal{P}(X_k|\mathbf{r}(k,l))]$ given by Equations 54, 55, and 56 in Appendix B.

Appendix D: AMA vs. AMA–Gauss: A simulated example yielding discrepant results

Here, we show AMA and AMA–Gauss are not guaranteed to give equivalent results, and therefore that the similarity of the results presented in the main text is not a foregone conclusion. We show that AMA learns the correct filters and that AMA–Gauss does not when the class-conditional response distributions are non-Gaussian. We simulate stimuli $\mathbf{s}$ with three stimulus dimensions, from each of two categories X_1 and X_2. (For comparison, in the main text the speed and disparity stimuli had 256 and 64 stimulus dimensions, respectively). The simulated stimuli are shown in Figure A1A–C. The first two dimensions of the simulated stimuli contain the information for discriminating the categories; the third stimulus dimension is useless. Specifically, the stimulus distributions are given by

$$\mathcal{P}(\mathbf{s}|X_1) = \frac{1}{2}\mathcal{N}(\mathbf{s};\mathbf{0},\Sigma_A) + \frac{1}{2}\mathcal{N}(\mathbf{s};\mathbf{0},\Sigma_B) \quad (64)$$

$$\mathcal{P}(\mathbf{s}|X_2) = \frac{1}{2}\mathcal{N}\left(\mathbf{s};\mathbf{0},R\Sigma_A R^T\right) + \frac{1}{2}\mathcal{N}\left(\mathbf{s};\mathbf{0},R\Sigma_B R^T\right) \quad (65)$$

$$\text{where} \qquad \Sigma_A = \begin{bmatrix} V_{large} & 0 & 0 \\ 0 & V_{small} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Sigma_B = \begin{bmatrix} V_{small} & 0 & 0 \\ 0 & V_{large} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and R is a 45° rotation matrix that operates on the first two dimensions. Stimuli in both categories are therefore distributed as mixtures of Gaussians with identical first- and second-order stimulus statistics (i.e., same mean and covariance). Thus, all information for discriminating the categories exists in higher-order statistical moments of the first two stimulus dimensions. Hence, because AMA–Gauss is sensitive only to class-conditional mean and covariance differences, it will be blind to the stimulus differences that define the categories.

AMA and AMA–Gauss were each tasked with learning two filters that discriminate the stimulus categories. AMA learns filters that enable the categories to be discriminated; its filters place all weight on the first two informative stimulus dimensions (Figure A1A, black arrows), and zero weight on the third uninformative stimulus dimension (Figure A1B, C). AMA–Gauss is blind to the stimulus statistics that enable the stimulus categories to be discriminated; its filters place their weights randomly, often putting substantial weight on the uninformative third stimulus dimension (Figure A1A–C, gray arrows). AMA–Gauss fails to learn filters that allow the categories to be nicely discriminated.

Figure A1D, E show AMA and AMA–Gauss conditional response distributions. AMA filter responses capture the mixture distribution that defines each category, and AMA–Gauss does not. Thus, the results in the main text are not simply due to the fact that Gaussians often provide good generic approximations of distributions with a lot of probability mass in one place.

Appendix E: Connection between AMA–Gauss and GQM

The log-likelihood of latent variable X_u using Equation 28 can be written as

$$l(X_u) = -\frac{1}{2}\left(\mathbf{x}^T\mathbf{f}\Sigma_u^{-1}\mathbf{f}^T\mathbf{x} - 2\left(\mu_u^T\Sigma_u^{-1} - \eta_u^T\Sigma_u^{-1}\right)\mathbf{f}^T\mathbf{x} + \mu_u^T\Sigma_u^{-1}\mu_u - \eta_u^T\Sigma_u^{-1}\eta_u + 2\eta_k^T\Sigma_u^{-1}\mu_u\right) + \zeta_u \quad (66)$$

where $\zeta_u = -\frac{1}{2}\log|2\pi\Sigma_u|$ is a constant. The expected log-likelihood can then be written as

$$E_\eta[l(X_u)] = -\frac{1}{2}\left(\mathbf{x}^T\mathbf{f}\Sigma_u^{-1}\mathbf{f}^T\mathbf{x} - 2\mu_u^T\Sigma_u^{-1}\mathbf{f}^T\mathbf{x} + \mu_u^T\Sigma_u^{-1}\mu_u - Tr\left(\Sigma_u^{-1}\Lambda\right)\right) + \zeta_u \quad (67)$$

It is evident from Equation 67 that $E_\eta[l(X_u)]$ is of the form $\mathbf{x}^TC\mathbf{x} + \mathbf{b}^T\mathbf{x} + a$ where

$$C = -\frac{1}{2}\mathbf{f}\Sigma_u^{-1}\mathbf{f}^T \quad (68)$$

$$\mathbf{b}^T = \mu_u^T\Sigma_u^{-1}\mathbf{f}^T \quad (69)$$

$$\text{and} \quad a = -\frac{1}{2}\mu_u^T\Sigma_u^{-1}\mu_u + \frac{1}{2}Tr\left(\Sigma_u^{-1}\Lambda\right) + \zeta_u \quad (70)$$