SENSORY REPRESENTATIONS OPTIMIZED FOR THE NATURAL ENVIRONMENT

Lingqi Zhang

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2023

Co-Supervisor of Dissertation	Co-Supervisor of Dissertation
David H. Brainard	Alan A. Stocker
Professor of Psychology	Associate Professor of Psychology

Graduate Group Chairperson

Russell Epstein, Professor of Psychology

Dissertation Committee

Johannes Burge, Associate Professor of Psychology
Konrad Kording, Professor of Bioengineering and Neuroscience

SENSORY REPRESENTATIONS OPTIMIZED FOR THE NATURAL ENVIRONMENT

To my wife, Tingting Wang.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my advisors, David Brainard and Alan Stocker. I would like to thank them for both their dedicated guidance and support, and the intellectual freedom I was granted to pursue questions I found truly interesting. I have learned so much from both of them over the years, but most importantly, how to always ask questions that push the boundary of our understanding, and never be satisfied with "just-so" answers. The work presented in this thesis would not have been possible without them.

I would like to thank my thesis committee, Konrad Kording and Johannes Burge, for their numerous feedback and suggestions. I am especially grateful to Johannes, who is always happy to answer all my random questions, no matter how trivial they may seem.

There are many other people who, in important ways, influenced and contributed to this thesis. I thank Jean-Paul Noel for the fun collaboration project we worked on together, which inspired Chapter 3. I thank Ozzy Taskin and Geoffrey Aguirre for teaching me everything I have learned about neuroimaging. I greatly enjoyed working with Cheng Qiu and Ari Benjamin, contemplating efficient codes in neural networks. I thank Nicolas Cottaris for all the help with ISETBio. I also thank Zahra Kadkhodaie and Eero Simoncelli for the many engaging discussions on the implicit image prior, and the incredible summer I spent at Flatiron, which led to the work in Chapter 5.

I am grateful for the wonderful people with whom I was able to share most of my graduate school journey. I thank Noam Roth for always being there when I needed someone to talk to; Takahiro Doi for the so many fun conversations we had ranging from academia to politics; Michael Barnett for keeping me company during late working hours; Ben Chin for showing me all the exciting psychophysics experiments he was working on; and Lisheng He for sharing with me his knowledge about human decision making.

I am incredibly lucky to have been a part of Goddard 420 and the larger CNI community, and to have spent time with a group of the most wonderful friends one could ever hope for. There are so many people so I shall not list them all. But for the particularly fond memories of the many The Great British Bake Off episodes and StarCraft gameplays, I would like to thank Kyra Shapiro, Catrina Hacker, Kara McGaughey, Jiang Mao, Dave White, Barnes Januzzi, Simon Bohn, and Liangyu Tao. I will truly miss this place.

Lastly, I would like to thank the love of my life, Tingting Wang, for her unconditioned love and support for me all these years, especially through the most difficult times. I also would like to thank my parents, Xuehong Cai and Junquan Zhang, for everything they have done that has shaped who I am today, and for always encouraging me to pursue my dreams.

ABSTRACT

SENSORY REPRESENTATIONS OPTIMIZED FOR THE NATURAL ENVIRONMENT

Lingqi Zhang

David H. Brainard

Alan A. Stocker

The limited resources available to the visual system must be allocated efficiently to support its function. To achieve this, our brain needs to take advantage of the statistical regularities of our visual environment. In this thesis, I systematically explore how different aspects of natural stimulus statistics can impact and determine perceptual behavior and sensory representation in both biological and artificial systems. In Chapter 1, I provide a brief review of the theory of efficient coding, models of natural image statistics, and the interplay between these two fields. In Chapter 2, based on a Bayesian ideal observer model that is constrained by efficient coding, I show how simple stimulus priors can provide a quantitative link between psychophysics and neurophysiology in the domain of speed perception. In Chapter 3, I extend these ideas to the domain of sensory adaptation. In particular, I develop a method to quantify changes in sensory encoding in a tilt illusion experiment, and find that these changes are consistent with an efficient coding account for which the encoding is optimized toward the conditional statistics of orientation based on the surrounding context. In Chapter 4, I generalize the efficient coding principle to fully naturalistic stimuli by building models of natural image statistics and image-computable ideal observers to quantify the information encoded by the early stages of visual encoding. I show how features of the retinal encoding can be explained by an optimal design principle. In Chapter 5, I present a novel algorithm for directly solving the linear optimal coding problem by finding the set of linear measurements that minimize error in a Bayesian image reconstruction problem. This approach improves upon established methods such as principal component analysis and compressed sensing, and provides a unifying perspective. Lastly, in Chapter 6, I discuss open questions and future directions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

BACKGROUND AND INTRODUCTION

1.1. Foreword

I will start this thesis with an example in image processing. The left image below (Fig. 1.1A) is of low quality due to the lack of contrast. A simple procedure, termed histogram equalization (Castleman, 1996), applies a pixel-wise remapping based on the cumulative distribution function (CDF) of the histogram (Fig. 1.1B), and can effectively restore the contrast of the image (Fig. 1.1C).

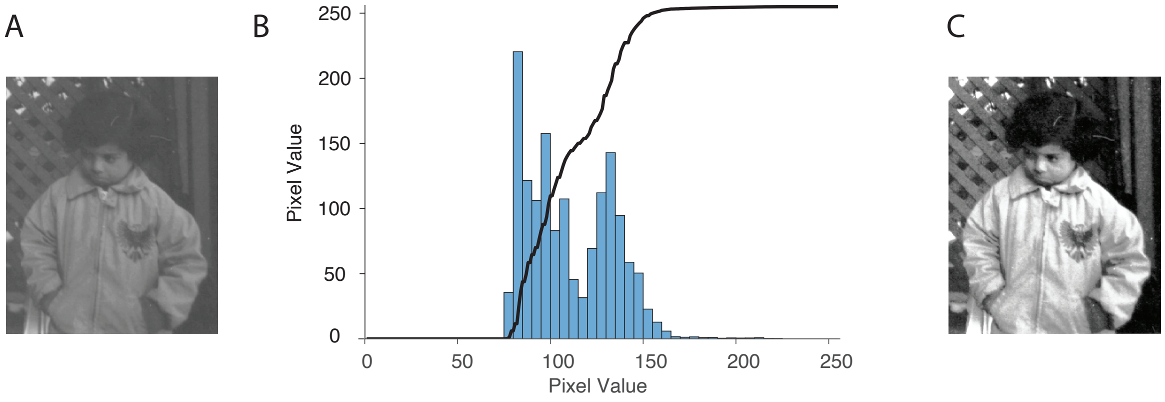

Figure 1.1: Histogram equalization. **A)** Original image. **B)** Histogram (blue) of the pixel values of the image on the left, and a remapping (black) from the original to a new image based on the CDF of the histogram. **C)** New image with restored image contrast.

This simple example demonstrates a few core general principles: 1) Displays are *constrained*, in that they can only show a limited range of luminance (typically at integer levels 0-255 in a digital device). 2) The goal of the image enhancement mechanism (Fig. 1.1B) is to make *efficient* use of the full range of the display. 3) The encoding is dependent on the ensemble of pixel values of the original image, that is, the *statistical* property (i.e., histogram)) of the signal.

It turns out, the same set of principles applies to the study of our own visual system, as the limited resources available to the visual system must be allocated efficiently to support its goal of veridically representing the external world. Theories developed under these principles are generally associated with the term “efficient coding” (Barlow et al., 1961). In the next few sections, I will provide a brief overview of the literature on efficient coding in studying vision, and perhaps equally important, the literature on statistical models of natural images. We will see how advancements made in these models provide new insight into the visual system through the lens of efficient coding. When appropriate, I will also give an overview of how the work presented in this thesis fits into and advances the broader literature.

1.2. Theory of Efficient Coding

1.2.1. Efficient Neural Codes

One early result about efficient neural codes is, in fact, a direct parallel to our image processing example above. Laughlin (1981) found that the nonlinear contrast response function of the interneurons in the blow-fly’s compound eye can be precisely derived based on the CDF of contrast in its natural environment. In general, theories of efficient coding aim to explain aspects of neural responses as optimally representing sensory stimuli based on their statistical regularities. While earlier results focus on single neurons and simple definitions of optimality such as redundancy reduction (van Hateren, 1992; Dan et al., 1996), more recent work has extended the concept to population codes (Tkačik et al., 2010; Pitkow and Meister, 2012), and incorporated more principled definitions of optimality such as those based on information (Strong et al., 1998) and estimation accuracy (Wang et al., 2016b; Park and Pillow, 2017). These results have provided normative explanations for neural tuning curves (Ganguli and Simoncelli, 2014), gain control (Schwartz and Simoncelli, 2001), and even shown that spike trains transmit information at the statistical limit in some conditions (Palmer et al., 2015). Recent computational models developed within this framework are able to explain more intricate features of the visual system beyond spikes and tuning curves, such as the organization of the early visual pathway (Karklin and Simoncelli, 2011; Zhou et al., 2020; Roy et al., 2021; Jun et al., 2021), and properties of receptive fields in early visual cortex (Olshausen and Field, 1996; Caywood et al., 2004).

1.2.2. Perceptual Behavior

The principle of efficient coding allows us to derive perceptual properties of the organism as a whole. Based on the principle of redundancy reduction (i.e., decorrelation), Atick and Redlich (1992) derived the idealized spatial contrast sensitivity functions across different luminance levels, which matched precisely the corresponding psychophysical measurements. Recent work is able to establish a general relationship between stimulus statistics and perceptual threshold sensitivity (Ganguli and Simoncelli, 2016; Wei and Stocker, 2017) based on a solution to the efficient coding problem in terms of Fisher information (Ganguli and Simoncelli, 2014; Wei and Stocker, 2016). One prominent empirical example in this domain is the oblique effect: People exhibit greater sensitivity to orientations around the cardinal directions than the obliques (Appelle, 1972). This can be seen as a consequence of efficient coding of orientation, which has an uneven distribution that bias towards vertical and horizontal in the natural environment (Girshick et al., 2011). In addition, it has been shown that this anisotropic representation of orientation is the exact source of the repulsive bias observed when people are asked to estimate the orientation of a stimulus (Wei and Stocker, 2015).

Although efficient coding provides a normative explanation of how stimulus statistics shape both neural representation and perceptual behavior, previous work tends to address these two aspects independently. In Chapter 2 of this thesis (Zhang and Stocker, 2022), I will present a theoretical framework that makes joint predictions for both neural coding and behavior, based on the assumption that neural representations of sensory information are efficient but also optimally used in generating a percept. I will validate the prediction of the model in the domain of speed perception, against electrophysiology data from neurons in the middle temporal (MT) cortex of macaque monkeys, and psychophysics data from human participants in a two-alternative forced choice (2AFC) speed discrimination task.

1.2.3. Adaptation

The statistical regularities of our environment exist on multiple spatial and temporal scales, and can change drastically depending on the context (Coppola et al., 1998; Schwartz et al., 2007). Efficient coding predicts that the brain should also adapt its encoding dynamically based on contextual cues. Thus, efficient coding provides a normative framework for understanding sensory adaptation. In fact, the empirical literature has indeed shown adaptation in neural codes ranging from timescales of milliseconds to minutes in order to maximize information transmission (Fairhall et al., 2001; Wark et al., 2009). Recent theory has also proposed neural codes that can balance the objective of both detecting context change and maximizing information within a specific context (Młynarski and Hermundstad, 2018, 2021). On the behavioral side, sensory adaptation has distinct perceptual signatures (Clifford et al., 2007), including changes in the discrimination threshold (Regan and Beverley, 1985; Clifford et al., 2001) and perceptual bias (Mitchell and Muir, 1976). Although it has been proposed that the changes in sensory encoding can potentially explain the observed behavior (Stocker and Simoncelli, 2005; Webster, 2011), there is no current consensus regarding how to attribute adaptation to changes at different stages of computation (e.g., encoding vs. decoding, Seriès et al. 2009). To bridge the gap between our understanding of sensory encoding and perceptual behavior in sensory adaptation, in Chapter 3 of this thesis, I develop a method for directly characterizing sensory encoding from psychophysical behavior. I will show in a tilt illusion experiment (Magnussen and Johnsen, 1986; Gibson and Radner, 1937) that the changes in orientation encoding across different spatial contexts are indeed consistent with the conditional statistics of orientation (Schwartz et al., 2007), as predicted by the efficient coding framework.

1.3. Natural Image Statistics

One of the crucial elements of efficient coding theory is the statistical properties of the input signal itself. While it is relatively straightforward to identify in certain cases such as for orientation (Coppola et al., 1998), the general problem is a challenging one, as it requires building probabilistic models of the natural scenes. Below, I will provide a brief overview of the progress that has been made in modeling natural image statistics. I will also discuss how this progress has helped in gaining insights into the visual system.

1.3.1. Spectral Model

We define images as high-dimensional vector $x \in \mathcal{R}^n$, where n is the number of pixels in the image across all color channels. The high dimensionality makes it extremely difficult to fully characterize $p(x)$. To simplify the problem, one can assume that $p(x)$ is a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Thus, to characterize natural images, we need to specify the covariance matrix Σ. Furthermore, assuming natural images are translation invariant, Σ will have a unique structure known as circulant. Specifically, each row of the matrix is a one-element rotated copy of the row preceding it. The eigenvectors of Σ in this case are known to be the discrete Fourier transform (DFT). Therefore, Σ can be diagonalized in the frequency domain, and the coefficients can be examined independently across frequencies as they become decorrelated (Simoncelli, 2005).

It turns out empirically that the covariance structure of natural images is indeed close to being translational invariant. In addition, the variance of the frequency components in the Fourier domain follows roughly a power law function with an exponent of around two (Deriugin, 1956; Tolhurst et al., 1992). This spectral model of natural images, while simplistic, has provided an important foundation for our understanding of images. Theories of efficient coding also often hypothesize that the goal of the visual system is to remove the correlations presented in natural signals in order to achieve redundancy reduction (Giridhar et al., 2011; Benucci et al., 2013; Duong et al., 2023).

1.3.2. Sparse Coding

If natural images are indeed Gaussian distributed, then any linear transformation applied to them should result in Gaussian random variables. However, when natural images are transformed using a wavelet basis, such as the steerable pyramid (Simoncelli and Freeman, 1995), the resulting coefficients, although approximately decorrelated, are highly non-Gaussian: The distribution has a higher concentration around zero and a more extended tail. This observation led to the hypothesis that natural images are “sparse”, meaning each individual image is composed of only a small subset of basis images.

Building upon the idea of sparsity, Olshausen and Field (1996) constructed a set of overcomplete basis functions that can faithfully reconstruct natural images while also being maximally sparse. The resulting basis images turn out to be a collection of localized, oriented edges at different scales, similar to the receptive field of neurons in the early visual cortex (Fig. 1.2). This has been taken as evidence that the visual system employs a sparse firing pattern across the population to represent images efficiently. It is worth noting, however, that the definition of efficiency in this context differs from the rest of this thesis, as it is not explicitly related to information content.

Regardless of the interpretation, the sparse wavelet model captures important structures of natural images, and has superior performance in applications such as image restoration compared to the spectral model (Elad and Aharon, 2006). In Chapter 4 of this thesis, I use this method to build prior over natural color images. By integrating it with an accurate model of the cone mosaic, I develop an image reconstruction-based method to quantify the information available during the initial stage of visual encoding (Zhang et al., 2022a). I will further demonstrate how the characteristics of the cone mosaic can be understood in an optimal coding framework.

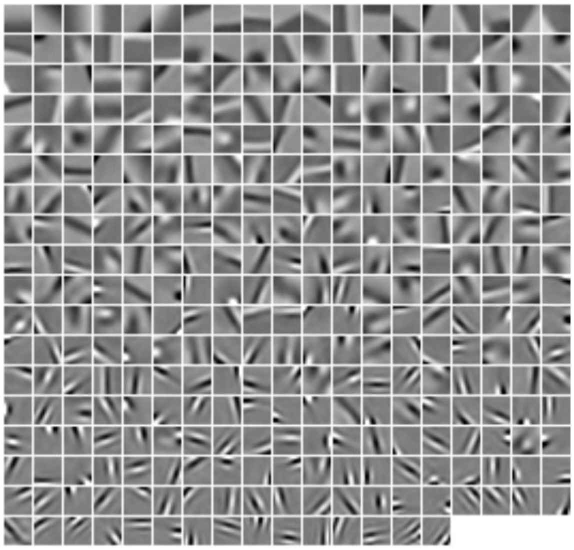

Figure 1.2: Sparse coding basis. A set of basis images learned through optimizing the sparse coding objective on natural images. Figure adapted from Foldiak and Endres (2008).

1.3.3. Gaussian Scale Mixture

In both the spectral and wavelet model, the coefficients are approximately decorrelated in the transformed domain. However, there still exists an important higher-order dependency among the coefficients (Portilla et al., 2003). The figure below shows the conditional histograms of two wavelet coefficients, w_1 and w_2, in adjacent spatial locations and scales (Fig. 1.3). As we can see, while the mean value of w_2 is independent of w_1, indicating decorrelation, the variance of w_2 grows almost linearly as the absolute value of w_1 increases.

One way to model this variance dependency is to employ an infinite mixture of Gaussian distribution (Portilla et al., 2003). Concretely, define $\mathbf{y} = \sqrt{z} \cdot \mathbf{x}$, where $\mathbf{x}$ is multivariate Gaussian and z is scalar random variable. The common multiplier z will capture this variance relationship. In addition, the marginal distribution of $\mathbf{y}$ in this model also accounts for the non-Gaussian behavior we have mentioned above. This Gaussian scale mixture (GSM) model further improves upon the wavelet sparse model, and it excels particularly at modeling texture statistics (Portilla and Simoncelli, 2000).

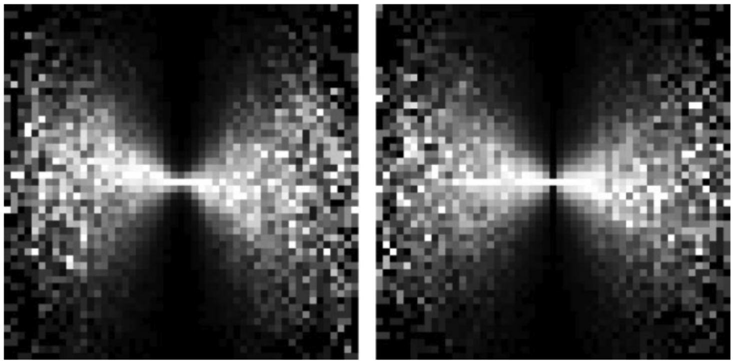

Figure 1.3: Conditional histogram in the wavelet domain for two different pairs of wavelets that are adjacent in spatial locations and scales. The x-axis represents the value of one wavelet coefficient w_1, while each column (y-axis) is a histogram of another wavelet w_2 conditioned on the value of w_1. Figure adapted from Portilla et al. (2003).

The GSM model also provides insight into divisive normalization, a commonly observed biological mechanism in cortical neurons (Carandini and Heeger, 2012). The model suggests that if we normalize $\mathbf{y}$ with an appropriately chosen scalar, namely $\sqrt{z}$, we can recover the normally distributed $\mathbf{x}$. In fact, it has been shown that the mechanism for computing the local normalization factor in neural circuits can be viewed as an estimator for the multiplier z (Schwartz and Simoncelli, 2001). Recent research has proposed a more complex mechanism for determining the normalization factor based on the local context of image content (Coen-Cagli et al., 2015). The early visual neurons have receptive fields that are similar to those used in building the GSM model. Thus, from a coding perspective, divisive normalization has two advantages: First, it removes the variance dependency between neurons, achieving further redundancy reduction (Schwartz and Simoncelli, 2001). Secondly, the normally distributed responses also maximize the information capacity of a single neuron (Iyer and Burge, 2019).

1.3.4. Diffusion Model

Despite the progress that has been made, the basic parametric models discussed above are still inadequate to fully characterize natural images. In fact, the samples drawn from these models are not realistic, and fail to capture the intricacies of real-world images. However, with the recent advancements in machine learning, specifically the development of diffusion probability models (Song and Ermon, 2019; Rombach et al., 2022), this limitation has been overcome. These models allow us to generate fully realistic images that are almost indistinguishable from real images by simply sampling from a distribution. In the paragraph below, I will provide a brief overview of diffusion models.

The goal here is to build a complex probability distribution $p(x)$. We can parameterize $p(x) : \mathcal{R}^n \rightarrow R_0^+$ by defining:

$$p_\theta(x) = \frac{1}{Z_\theta} \exp(-f_\theta(x)), \tag{1.1}$$

where $-f_\theta(x)$ is an energy function that can be arbitrarily complex, such as a neural network. However, calculating the normalizing constant Z_θ requires evaluating the integral $\int \exp(-f_\theta(x))dx$, which is infeasible when x is high-dimensional, as in the case of natural images. Alternatively, we can define a score function $s(x) : \mathcal{R}^n \rightarrow \mathcal{R}^n$, as

$$s(x) = \nabla_x \log p(x) = -\nabla_x f_\theta(x). \tag{1.2}$$

By taking the derivative of the log probability density, we avoid calculating the normalizing constant Z_θ. We can estimate the score function $s(x)$ directly from data, based on a technique called score matching (Hyvärinen and Dayan, 2005). Once learned, the score function can be used to draw samples from the corresponding prior $p(x)$ through Langevin dynamics (Bussi and Parrinello, 2007):

$$x_{t+1} = x_t + \epsilon \nabla_{x_t} \log p(x) + \sqrt{2\epsilon}\; z_t,\;\; z_t \sim \mathcal{N}(0, 1). \tag{1.3}$$

The stationary distribution of x_t converges to $p(x)$ as $\epsilon \to 0$ and $t \to \infty$. Furthermore, there is a precise mathematical relationship between the score function and image denoising (Miyasawa, 1961). As such, it allows the score function to be learned by simply training models to perform additive Gaussian noise removal, and the sampling procedure can be viewed as an iterative denoising process (Vincent, 2011; Kadkhodaie and Simoncelli, 2021).

The intricate prior represented by these models opens up new possibilities for studying how natural image statistics shape the visual system. In our recent work, we have incorporated a more complex prior model using the image prior implicit in a convolutional neural network denoiser (Kadkhodaie and Simoncelli, 2021; Zhang et al., 2022b). In Chapter 5, I present a novel method for identifying the optimal set of linear measurements that, when combined with an image prior, minimizes the Bayesian reconstruction error on a set of natural images. This can be viewed as a form of linear efficient coding, and my analysis examines the effect of the image prior on the optimal solution. The approach also improves upon and provides a unifying perspective on other linear methods such as principle component analysis (PCA) and compressed sensing (CS).

CHAPTER 2

PRIOR EXPECTATIONS IN VISUAL SPEED PERCEPTION PREDICT ENCODING CHARACTERISTICS OF NEURONS IN AREA MT

This chapter was previously published as: Ling-Qi Zhang and Alan A Stocker. Prior expectations in visual speed perception predict encoding characteristics of neurons in area MT. *Journal of Neuroscience*, 42(14):2951–2962, 2022. I contributed to the conceptualization, formal analysis, methodology, validation, software, visualization, and writing of this work.

Abstract

Bayesian inference provides an elegant theoretical framework for understanding the characteristic biases and discrimination thresholds in visual speed perception. However, the framework is difficult to validate due to its flexibility and the fact that suitable constraints on the structure of the sensory uncertainty have been missing. Here, we demonstrate that a Bayesian observer model constrained by efficient coding not only well explains human visual speed perception but also provides an accurate quantitative account of the tuning characteristics of neurons known for representing visual speed. Specifically, we found that the population coding accuracy for visual speed in area MT ("neural prior") is precisely predicted by the power-law, slow-speed prior extracted from fitting the Bayesian model to psychophysical data ("behavioral prior") to the point that the two priors are indistinguishable in a cross-validation model comparison. Our results demonstrate a quantitative validation of the Bayesian observer model constrained by efficient coding at both the behavioral and neural levels.

2.1. Introduction

Human perception of visual speed is typically biased and depends on stimulus attributes other than the actual motion of the stimulus. Contrast, for example, strongly affects perceived stimulus speed such that a low-contrast drifting grating typically appears to move slower than a high-contrast grating (Thompson, 1982; Stone and Thompson, 1992; Blakemore and Snowden, 1999; Stocker and Simoncelli, 2006). These biases and perceptual distortions are qualitatively consistent with a Bayesian observer that combines noisy sensory measurements with a prior preference for lower speeds (Simoncelli, 1993; Weiss et al., 2002; Stocker, 2006). Previous work has also shown that by embedding the Bayesian observer within a two-alternative forced choice (2AFC) decision process one can "reverse-engineer" the noise characteristics (i.e., likelihood) and prior expectations of individual human subjects from their behavior in a speed-discrimination task (Stocker and Simoncelli, 2004, 2006). This provided both a quantitative validation of the Bayesian observer model and a normative interpretation of human behavior in visual speed perception tasks, which has been confirmed in various later studies (e.g. Welchman et al., 2008; Hedges et al., 2011; Sotiropoulos et al., 2014; Jogan and Stocker, 2015). However, recovering the parameters of a Bayesian observer model from behavioral data is typically difficult due to the intrinsic non-specificity of its probabilistic formulation, which has been grounds for a critical view of the Bayesian modeling approach altogether (Jones and Love, 2011; Bowers and Davis, 2012). The reverse-engineered speed priors in previous studies indeed all showed large variations across subjects, indicating a potential case of over-fitting due to insufficient model constraints (Stocker and Simoncelli, 2006; Hedges et al., 2011; Sotiropoulos et al., 2014; Jogan and Stocker, 2015).

In this article, we show how we addressed this potential problem by developing and validating a tightly constrained Bayesian observer model. We followed a recent proposal to use efficient coding as a constraint that links the likelihood function to the prior expectations of a Bayesian observer (Wei and Stocker, 2012, 2015). The efficient coding hypothesis posits that biological neural systems allocate their limited coding capacity such that overall information transmission is optimized given the stimulus distribution in the natural environment (Barlow et al., 1961; Laughlin, 1981). It thus establishes a direct relationship between the stimulus distribution and the accuracy of neural representations in sensory systems (Linsker, 1988; McDonnell and Stocks, 2008; Wang et al., 2012; Ganguli and Simoncelli, 2014; Yerxa et al., 2020; Roy et al., 2021). Wei and Stocker (2015) showed how to formulate efficient coding as an information constraint that can be embedded within the probabilistic language of the Bayesian framework. The resulting Bayesian observer model has proven to account for a wide range of phenomena in perception including repulsive biases in perceived visual orientation (Wei and Stocker, 2015; Taylor and Bays, 2018) and the lawful relationship between perceptual bias and discrimination threshold (Wei and Stocker, 2017), but also in more cognitive domains such as subjective preferences judgments (Polania et al., 2019) or the representation of numbers (Cheyette and Piantadosi, 2020; Prat-Carrabin and Woodford, 2021).

The overall goal of our current work was two-fold. First, we aimed for quantitative validation of this new Bayesian observer model in the domain of visual speed perception. We fit the model to speed discrimination data collected by Stocker and Simoncelli (2006). We found that compared to the model in this original study, the new model allowed us to reverse-engineer much more reliable and consistent estimates of subjects' prior beliefs while still accurately accounting for subjects' psychophysical behavior. Second, based on the efficient coding hypothesis we wanted to test whether the reverse-engineered prior expectations are mirrored in the population encoding characteristics of neurons in the motion-sensitive area in the primate brain. The middle temporal (MT) area is widely recognized as the cortical area in the primate brain that selectively represents direction and speed of moving visual stimuli (Zeki, 1974; Newsome and Pare, 1988; Britten et al., 1993; Movshon and Newsome, 1996; Priebe et al., 2003). By analyzing single-cell recordings of a large population of MT neurons (Nover et al., 2005), we found that the sensitivity with which visual speed is encoded in this population ("neural prior") is precisely predicted by the prior beliefs extracted from the psychophysical data ("behavioral prior"). Our results provide important quantitative validation of the Bayesian observer model constrained by efficient coding at both the behavioral and neural levels.

2.2. Methods

2.2.1. Behavioral prior: Bayesian observer model constrained by efficient coding

Data

We reanalyzed the two-alternative-forced-choice (2AFC) speed discrimination data collected in Stocker and Simoncelli (2006). In each trial of the experiment a subject was shown a pair of horizontally drifting gratings (reference and test), and was asked to choose which one of them was moving faster. The reference grating had one of two contrast levels [0.075, 0.5] and one of six different drifting speeds [0.5, 1, 2, 4, 8, 12] deg/s. The test grating had one of seven different contrast levels [0.05, 0.075, 0.1, 0.2, 0.4, 0.5, 0.8] and its speed was determined by an adaptive staircase procedure (one-up/one-down). There were 72 different individual conditions (i.e., psychometric curves) and each condition contained 80 trials, resulting in a total of 5,760 trials. We excluded one subject (labeled as Subject 3 in Stocker and Simoncelli (2006)) that was only tested at two contrasts and two test speed levels. While we were able to recover a prior from this subject that was highly consistent with the rest of the subjects, it was not possible to perform a meaningful model comparison and cross-validation due to the low number of trials.

Model formulation

We use the Bayesian observer model by Wei and Stocker (2015) and embed it within a decision process to predict the binary judgments in the 2AFC experiment (Stocker and Simoncelli, 2006).

Specifically, we assume that "encoding" of the stimulus is governed by an efficient coding constraint such that encoding accuracy, measured as the square-root of Fisher Information (FI), is proportional to the stimulus prior (Brunel and Nadal, 1998; McDonnell and Stocks, 2008; Wei and Stocker, 2016), that is

$$\sqrt{I_F}(v) \propto p(v) \ . \tag{2.1}$$

Encoding is described as the conditional probability distribution $p(m|v)$. It determines how stimulus speed v is transformed probabilistically into a noisy sensory measurement m. We can satisfy the efficient coding constraint (Eq. (2.1)) by assuming the following encoding distribution

$$p(m|v) = \mathcal{N}(m; \mu = F(v), \sigma^2 = h^2(c)) \ , \tag{2.2}$$

where $F(v) = \int_{-\infty}^{v} p(v)dv$ is the cumulative density function (CDF) of v. We parameterized the speed prior distribution as the modified power-law function

$$p(v) \propto (|v| + c_1)^{c_0} + c_2 \ , \tag{2.3}$$

where $c_{0,1,2}$ are free and unconstrained parameters. We also tested alternative parameterizations (Fig. 2.4). The scalar $h(c)$ determines the amount of total encoding resources (i.e., the overall magnitude of internal noise) at different contrast levels. It can be shown that

$$\sqrt{I_F(v)} = F'(v)/\sigma = p(v)/h(c) \ . \tag{2.4}$$

The total amount of encoding resource is measured by $\int \sqrt{I_F(v)}dv$ which evaluates to $1/h(c)$. In order to numerically handle the unbounded nature of a magnitude variable such as speed (compared to a circular variable such as orientation), we added a small constant $(2.5 * 10^{-3})$ to $p(v)$ such that its CDF did not saturate (i.e., $F(v)$ is not upper bounded by 1).

To decode (i.e., estimate) the stimulus v given a particular sensory representation m, we first determine the likelihood function

$$l(v) = p(m|v) \tag{2.5}$$

by considering the encoding distribution (Eq. (2.2)) as a function of v. Applying Bayes' rule and multiplying the likelihood function with the prior $p(v)$ (Eq. (2.3)), we then can compute the posterior as

$$p(v|m) \propto l(v)p(v) \ . \tag{2.6}$$

Assuming an L_0 loss function (Stocker and Simoncelli, 2006), the estimate $\hat{v}$ of the stimulus v is given as

$$\hat{v} = \underset{\hat{v}}{\operatorname{argmax}} \ \ l(\hat{v})p(\hat{v}) \ . \tag{2.7}$$

The estimate represents the optimally decoded stimulus $\hat{v}$ given m. It is a deterministic function of m (implicit in the likelihood function $l(v)$), which we can explicitly express as $\hat{v}(m)$. However, m is not directly observable in a psychophysical experiment. Thus, we marginalize over m to obtain the estimate distribution for a given stimulus v,

$$p(\hat{v}|v) = \int p(\hat{v}|m)p(m|v)dm = \int \delta(\hat{v} - \hat{v}(m))p(m|v)dm \ , \tag{2.8}$$

where $\delta(\cdot)$ is the Dirac delta function.

In a 2AFC speed discrimination experiment, subjects report a binary decision and not a continuous estimate $\hat{v}$. We assume subjects make their choice (i.e., which one is faster) by comparing their estimate $\hat{v}_r$ of the reference stimulus with their estimate $\hat{v}_t$ of the test stimulus. For a pair of v_t and v_r, across many repeated trials, these choices follow a binomial distribution with the probability of the test stimulus being perceived faster given as

$$p_{v_t,v_r}(\hat{v}_t > \hat{v}_r) = \int_{-\infty}^{+\infty} p(\hat{v_t}|v_t) \int_{-\infty}^{\hat{v}_t} p(\hat{v_r}|v_r) d\hat{v}_r d\hat{v}_t \ . \tag{2.9}$$

Model fitting

If we represent the data in our experiment as N triplets (v_{ir}, v_{it}, k_i), where $k_i \in \{0, 1\}$ represents the binary choice, then the overall log-likelihood of the model given the data is

$$\mathcal{L} = \sum_{i=1}^{N} \{k_i \log[p_{v_{it},v_{ir}}(\hat{v}_{it} > \hat{v}_{ir})] + (1 - k_i) \log[1 - p_{v_{it},v_{ir}}(\hat{v}_{it} > \hat{v}_{ir})]\} \ . \tag{2.10}$$

We find the model parameters c_0, c_1, c_2 and $h(c)$ by maximizing $\mathcal{L}$ using MATLAB's *fminsearchbnd* algorithm. Note that the model is highly constrained: For each subject, we jointly fit a single three-parameter prior distribution plus one scalar noise parameter $h(c)$ for each of the 7 contrast levels to the data from all 72 conditions.

Alternative prior parameterization

In order to assess the consistency and stability of our reverse-engineered prior distributions, we also tested two alternative parameterizations (Fig. 2.4):

- a Gamma distribution: $p(v; \alpha, \beta) \propto |v|^{\alpha-1} e^{-\beta|v|}$
- a piece-wise log-linear function with 18 sample points $v^*_{1:18}$ equally distributed in logarithmic space in the range $v = [0...50]$ deg/s. Each corresponding $p(v^*_{1:18})$ value is a free prior parameter; prior density values are linearly interpolated between those values.

For comparison we also fit a Gaussian prior with $p(v; \sigma^2) = \mathcal{N}(v; \mu = 0, \sigma^2)$.

Weber's law and power-law prior

With our model, it is possible to analytically predict discrimination threshold Δ_v and Weber fraction $\frac{\Delta_v}{v}$ for any given prior distribution. It has been shown that discrimination threshold is inversely proportional to the square-root of FI (Seriès et al., 2009; Wei and Stocker, 2017), thus

$$\Delta_v \propto \frac{1}{\sqrt{I_F}} \ . \tag{2.11}$$

According to the efficient coding constraint (Eq. (2.1)), we can substitute $\sqrt{I_F}$ with $p(v)$ and find

$$\Delta_v \propto \frac{1}{p(v)} \ . \tag{2.12}$$

This equation allows us to predict discrimination threshold for any prior density (up to a scale factor). For the modified power-law prior with exponent $c_0 = -1$ and $c_2 = 0$ (Eq. (2.3)) we can find

$$\Delta_v \propto (v + c_1) \ . \tag{2.13}$$

By further setting $c_1 = 0$, we obtain $v \propto \Delta_v$, which is the definition of Weber's law (i.e., a constant Weber fraction). For non-zero c_1 the Weber fraction changes to

$$\frac{\Delta_v}{v} \propto (1 + \frac{c_1}{v}) \ . \tag{2.14}$$

At high speeds, $\frac{c_1}{v} \approx 0$ and thus the Weber fraction is constant. At low speeds, $\frac{c_1}{v} \to \infty$ causing $\frac{\Delta_v}{v}$ to increase.

Our efficient coding constraint implies that the stimulus is transformed according to the CDF of v (Eq. (2.2)). For a power-law prior with exponent $c_0 = -1$ and $c_2 = 0$, the CDF is

$$\int z_1 (v + c_1)^{-1} dv = z_1 \log(v + c_1) + z_2 \ , \tag{2.15}$$

which is precisely the logarithmic transformation that has been previously used for describing the speed tuning of MT neurons (Nover et al., 2005).

2.2.2. Neural prior: MT encoding analysis

Data

We reanalyzed the electrophysiological recording data from Nover et al. (2005). Neurons in area MT of several macaque monkeys were individually identified. Each identified neuron was then tested with a random-dot motion stimulus moving with one of eight speeds [0, 0.5, 1, 2, 4, 8, 16, 32] deg/s. Stimulus location, direction, size and disparity were individually optimized for each neuron. Every stimulus speed was presented three to seven times. We considered the mean firing rate over the entire stimulus duration (1.5 s) as a neuron's single-trial response. We analyzed a total of 480 neurons.

Population Fisher information

Following Nover et al. (2005), we fit each neuron's mean firing rate as a function of stimulus speed with a Gaussian tuning curve in log-speed

$$R(v) = R_0 + A * \exp(-\frac{\log[q(v)]^2}{2\sigma^2}) \ , \tag{2.16}$$

where $q(v) = \frac{v+v_0}{v_p+v_0}$. Parameters R_0, A, σ^2, v_0, and v_p are determined by minimizing the sum of squared difference of the observed and predicted firing rates. A maximum-likelihood fit assuming Poisson distributed firing rate variability produced very similar results.

We computed the population FI for different assumptions about neurons' response variabilities and their pair-wise noise correlations within the population. First, we assumed that response noise is independent between neurons in the population and response variability is well-described by a Poisson process. In this case, the population FI is calculated as

$$I_F(v) = \sum_{i=1}^{N} \frac{[R_i'(v)]^2}{R_i(v)} \ . \tag{2.17}$$

The "neural prior" (the prior that corresponds to the measured MT encoding precision assuming efficient encoding) is then equivalent to the normalized square-root of FI, thus

$$p(v) = \frac{\sqrt{I_F(v)}}{\int \sqrt{I_F(v)} dv} \ . \tag{2.18}$$

As in Nover et al. (2005), we also repeated the above analysis using an alternative tuning-curve model (Gamma distribution function) and obtained very similar results.

Next, we estimated the population FI by adjusting the Poisson model with an explicit estimate of the Fano factor F_i for each neuron

$$I_F(v) = \sum_{i=1}^{N} \frac{[R_i'(v)]^2}{F_i * R_i(v)} \ , \tag{2.19}$$

where F_i was obtained by linearly regressing the neuron's firing rate variance against its firing rate mean. Finally, we computed the linear Fisher information (Kanitscheider et al., 2015; Kohn et al., 2016)

$$I_F(v) = \vec{R}'(v)^T \Sigma^{-1} \vec{R}'(v) \ , \tag{2.20}$$

where Σ is the noise correlation matrix and is the identity matrix for the independent noise case. To understand the effect of speed tuning preference-dependent noise correlations observed in area MT (Huang and Lisberger, 2009), we adopted the limited-range correlation model

$$\begin{aligned} \Sigma_{ij} &= \sigma^2 \rho_{ij} \\ \rho_{ij} &= \exp(-|\Delta_{ij}|/L) \ , \end{aligned} \tag{2.21}$$

where Δ_{ij} is difference in log-speed preference between neuron i and neuron j, σ^2 is the noise variance, and L is the overall correlation strength (Abbott and Dayan, 1999). For the simulations shown in Fig. 2.7E, we set $\sigma^2 = 1$ as the prior is only determined by the shape of the population FI and $L = [0.5, 1.0, 2.5]$ for low, medium and high correlation strengths.

2.2.3. Cross-validation

We performed a five-fold cross-validation procedure. The trial data for each condition were first randomly and equally divided into 5 groups. For each group, the model was fit to the data of the remaining four groups (training), and then evaluated on the group's data (validation). Model validation performance was measured as the log-likelihood of the fit model given the validation data. The entire procedure was repeated 20 times, resulting in 100 estimates of the model validation likelihood. For the behavioral prior condition, we considered the full observer model using the fit prior for that run. For the neural prior condition, we assumed the prior to be fixed and equal to the prior extracted from the population FI analysis with only the contrast-dependent noise parameters being fit on each run. The same procedure was used to compute the validation likelihoods of the original, less constrained Bayesian observer model (Stocker and Simoncelli, 2006), and of individual Weibull fits to every condition. Log-likelihood values in Fig. 2.8B were normalized to the range set by a lower bound given by the log-likelihoods of a coin-flip model for the decision (i.e. a model with a fixed decision probability of 0.5), and an upper bound determined by the values of the Weibull fits.

2.2.4. Data and Code Accessibility

Data and analysis code, including the instruction to create a full display of the psychometric curves and model fits for individual subjects, are available through GitHub:
https://github.com/lingqiz/Speed_Prior_2021

2.3. Results

We model speed perception as an efficient encoding, Bayesian decoding process (Fig. 2.1A). On any given trial the speed v of a visual stimulus is represented by a noisy and bandwidth-limited sensory measurement m. Following Wei and Stocker (2012, 2015), we assume that "encoding" of the stimulus is governed by an efficient coding constraint (Eq. (2.1)) such that encoding accuracy, measured as the square-root of Fisher Information (FI), is proportional to the stimulus prior $p(v)$ (Brunel and Nadal, 1998; McDonnell and Stocks, 2008; Wei and Stocker, 2016). This constraint promotes a more accurate encoding of speeds for which the prior density is high. It determines the observer's uncertainty about the actual stimulus speed given a particular sensory measurement (i.e., the likelihood function $p(m|v)$). For "decoding", this likelihood function $p(m|v)$ is combined with the stimulus prior $p(v)$, resulting in the posterior $p(v|m)$. Lastly, a percept $\hat{v}$ (i.e., an estimate) is computed based on the posterior and a loss function (see *Methods* for details).

A unique feature of the new model is that the stimulus distribution jointly determines encoding and decoding of the Bayesian observer model (Wei and Stocker, 2012). Thus, both the encoding characteristics of neurons representing visual speed (Fig. 2.1B), and the psychophysical behavior of subjects in speed perception (Fig. 2.1C) should be consistent with the prior belief of the observer about the statistical regularities of visual speed.

2.3.1. Extracting the "behavioral prior"

We fit our model to the psychophysical speed discrimination data collected previously in another study (Stocker and Simoncelli, 2006). On each trial of their experiment, subjects were shown a pair of horizontally drifting gratings (reference and test stimulus), and were asked to choose which one was moving faster (Fig. 2.1C). For each combination of stimulus contrast and reference speed, a full psychometric curve was measured by repeating the trials at different test speeds chosen by an adaptive staircase procedure. A total combination of 72 conditions representing reference and test stimuli at different speeds and contrast levels were tested, resulting in 72 different psychometric functions (see *Methods* and Stocker and Simoncelli (2006) for details).

In contrast to the original model (Stocker and Simoncelli, 2006), the new observer model directly links the likelihood function and the prior distribution (Wei and Stocker, 2012). Thus perceived speed is fully determined by subjects' prior expectations and a contrast-dependent internal noise parameter that reflects the total amount of represented sensory information (Wei and Stocker, 2015, 2016). Our goal was to find the prior distribution $p(v)$ and the noise parameters $h(c)$ that best accounted for subjects' individual perceptual behavior. In order to fit the observer model, we embedded it within a binary decision process (Fig. 2.1C). On each trial, speed estimates for both the reference and the test stimuli are performed, and then subjects are assumed to respond according to which estimate is faster. Entire psychometric functions are predicted by marginalizing over the

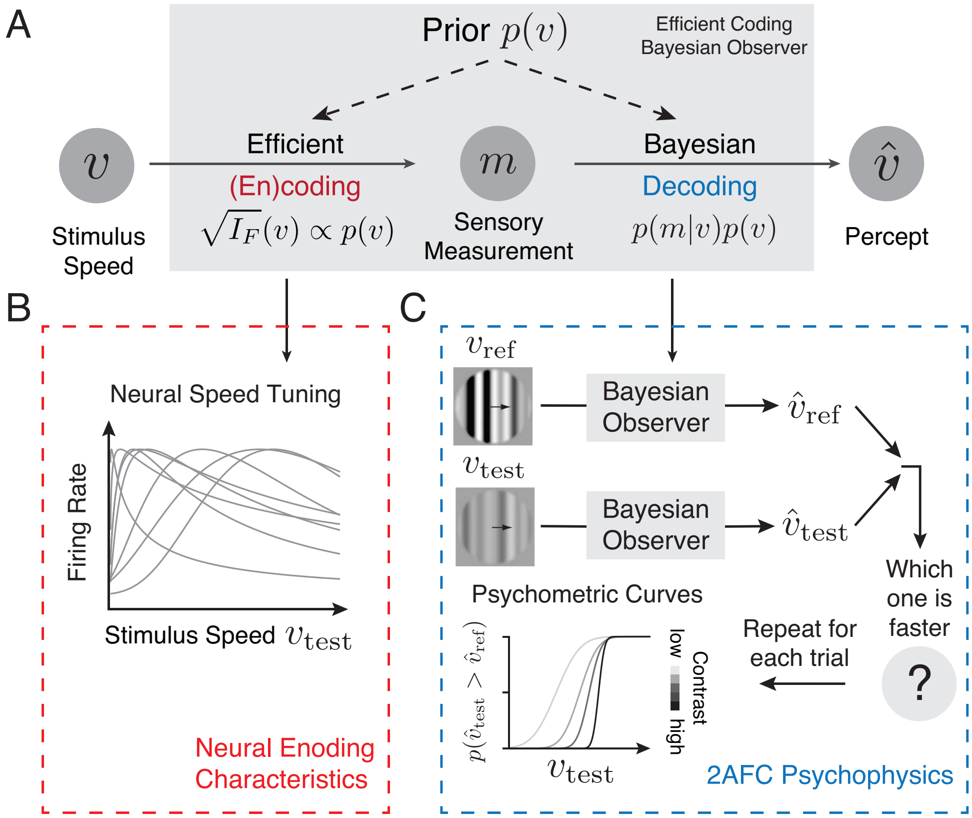

Figure 2.1: Bayesian observer model constrained by efficient coding. **A)** We model speed perception as an efficient encoding, Bayesian decoding process (Wei and Stocker, 2012, 2015). Stimulus speed v is encoded in a noisy and resource-limited sensory measurement m with an encoding accuracy that is determined by the stimulus prior $p(v)$ via the efficient coding constraint (Eq. (2.1)). Ultimately, a percept is formed through a Bayesian decoding process that combines the likelihood $p(m|v)$ and prior $p(v)$ to compute the posterior $p(v|m)$, and then selects the optimal estimate $\hat{v}$ according to a loss function. Encoding and decoding are linked and jointly determined by the prior distribution over speed. **B)** Efficient coding determines the accuracy of the neural representation of visual speed (i.e., the tuning characteristics of neurons in area MT). **C)** Embedding the Bayesian observer within a decision process provides a model to predict psychophysical behavior in a two-alternative forced choice (2AFC) speed discrimination task.

Subject	c_0	c_1	c_2	$h(0.05)$	$h(0.075)$	$h(0.10)$	$h(0.20)$	$h(0.40)$	$h(0.50)$	$h(0.80)$
1	-0.790	0.003	$6 * 10^{-5}$	0.035	0.027	0.028	0.019	0.016	0.013	0.012
2	-0.867	0.002	10^{-8}	0.045	0.041	0.038	0.030	0.023	0.020	0.010
3	-1.045	0.220	$1 * 10^{-5}$	0.061	0.045	0.041	0.029	0.017	0.019	0.026
4	-1.097	0.248	$7 * 10^{-6}$	0.043	0.040	0.036	0.028	0.018	0.017	0.016

Table 2.1: Fit parameter values of the prior density $p(v) \propto (|v| + c_1)^{c_0} + c_2$ and the contrast-dependent noise $\sigma = h(c)$ for every subject.

unobserved sensory measurement (*Methods*).

We jointly fit our model for every subject to all 72 conditions using a maximum-likelihood procedure. The free parameters of the model consisted of a parametric description of the prior and one noise parameter for each stimulus contrast. Following previous studies (Stocker and Simoncelli, 2006; Hedges et al., 2011; Jogan and Stocker, 2015), we parameterized the prior distribution as a modified power-law function (Eq. (2.3)). Figure 2.2A shows the data and model fit for a few example conditions for exemplary Subject 1. Fit parameter values for all subjects are listed in Table 2.1. Overall, the model predicts psychometric curves that are similar to those obtained from fitting a Weibull function. The log-likelihood of the new model is close to that of separate Weibull fits to every individual condition (Fig. 2.2B). Figure 2.2C further illustrates that the new model performs as well as the original, less constrained Bayesian observer model (Stocker and Simoncelli, 2006). A more detailed model comparison utilizing cross-validation is provided in a later section.

Importantly, the reverse-engineered prior expectations are much more consistent across subjects than those obtained from using the original model (Stocker and Simoncelli, 2006; Hedges et al., 2011; Sotiropoulos et al., 2014; Jogan and Stocker, 2015). The exponent c_0, for example, is now close to a value of -1 for every subject rather than varying over an order of magnitude (Table 2.1). Furthermore, values of the contrast-dependent noise parameter monotonically decrease as a function of contrast as expected and are consistent with the functional description of the contrast response curve of cortical neurons (Fig. 2.3).

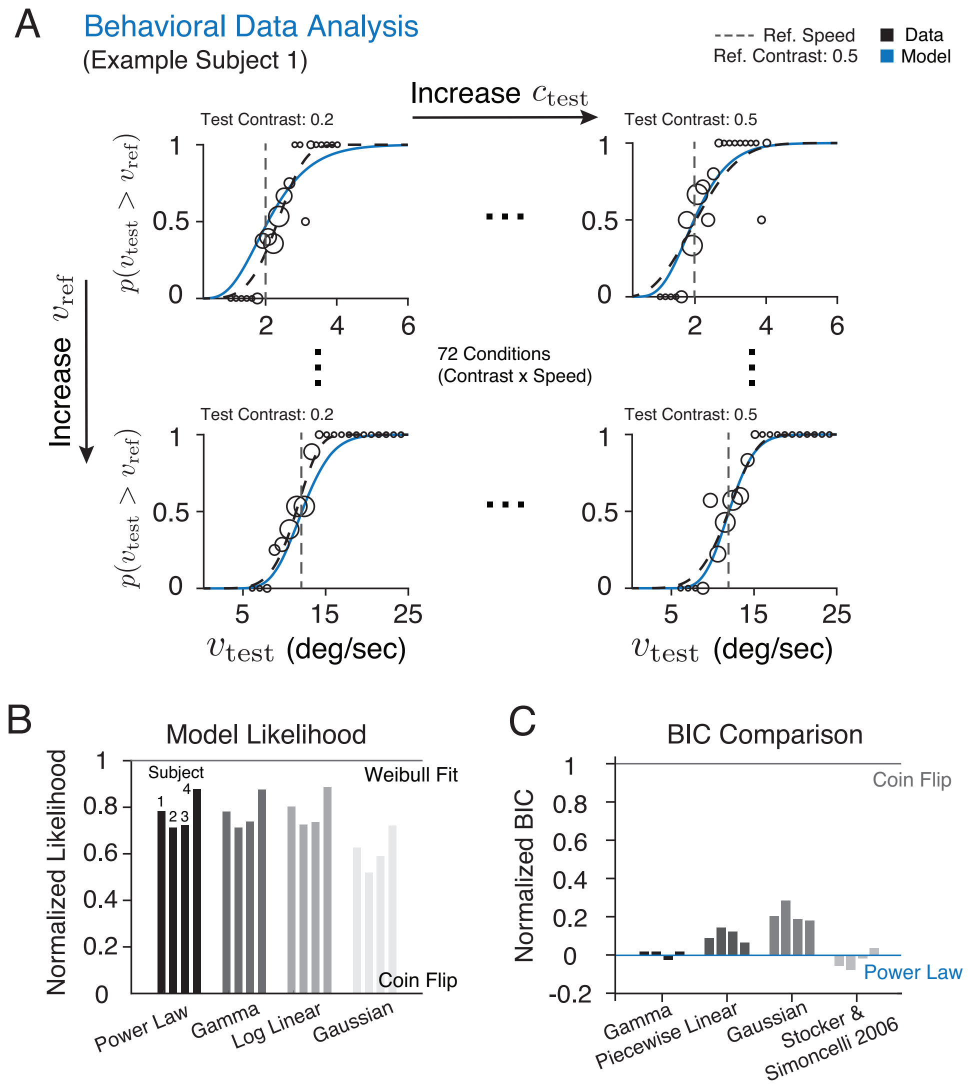

Figure 2.2: Extracting the "behavioral prior". **A)** We jointly fit the Bayesian observer model to psychophysical data across all contrast and reference speed conditions (72 conditions total). Shown are a few conditions for exemplary Subject 1. Circle sizes are proportional to the number of trials at that test speed. The dashed curves are Weibull fits to each conditions, and the solid blue curves represent the model prediction. See *Data and Code Accessibility* for instructions to create a full display of psychometric curves and model fits of all 72 conditions for individual subjects. **B)** Log-likelihood values of the best-fitting model for each subject using four different prior parameterizations including a power-law function, Gamma distribution, piece-wise log-linear function, and a Gaussian distribution, respectively. Values are normalized to the range set by a coin flip model (lower bound) and Weibull fits to individual psychometric curves (upper bound). **C)** The relative Bayesian Information Criterion (BIC) values for the different parameterizations as well as the original, less constrained Bayesian observer model by Stocker and Simoncelli (2006). Values are normalized to the range set by the efficient coding Bayesian model with power-law parameterization and the coin flip model (lower is better). See *Methods* for details.

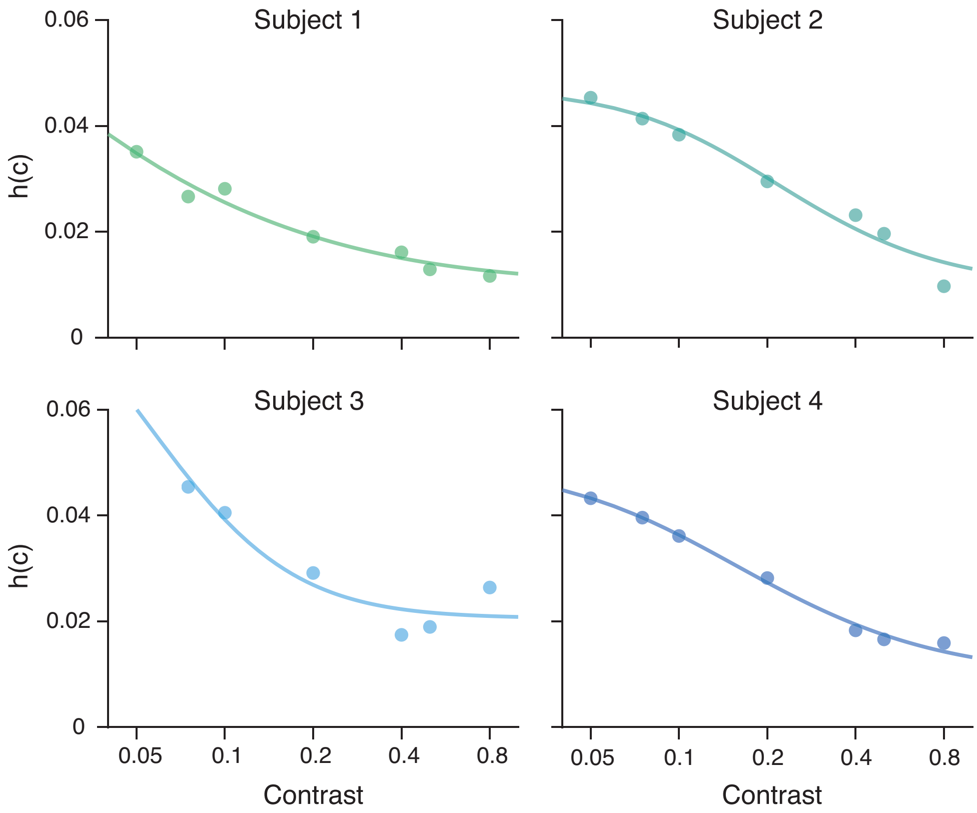

Figure 2.3: Contrast-dependent noise. Fit parameters values as a function of stimulus contrast plotted for every subject. Bold lines represent fits with a parametric description of the contrast response function of cortical neurons $h(c) = [r_{\max} c^q/(c^q + c_{50}^q) + r_{\text{base}}]^{-1/2}$ (Sclar et al., 1990; Albrecht and Hamilton, 1982; Heuer and Britten, 2002).

In order to test the impact of choosing a power-law parameterization for the prior distribution (Eq. 2.3), we performed model fits using two other parameterizations with increasing degrees of freedom (i.e., a Gamma distribution and a piece-wise log-linear function), and also a Gaussian prior for comparison (*Methods*). The model fits well for all but the Gaussian prior, resulting in similar log-likelihood values (Fig. 2.2B) although the BIC value is higher for the log-linear parameterization due to its large number of parameters (Fig. 2.2C). Crucially, however, the shapes of the fit prior distributions are very similar across the different parameterizations, all exhibiting a power-law like, slow-speed preferred distribution (Fig. 2.4). The obvious exception is the Gaussian parameterization because it is unsuited to approximate a power-law function.

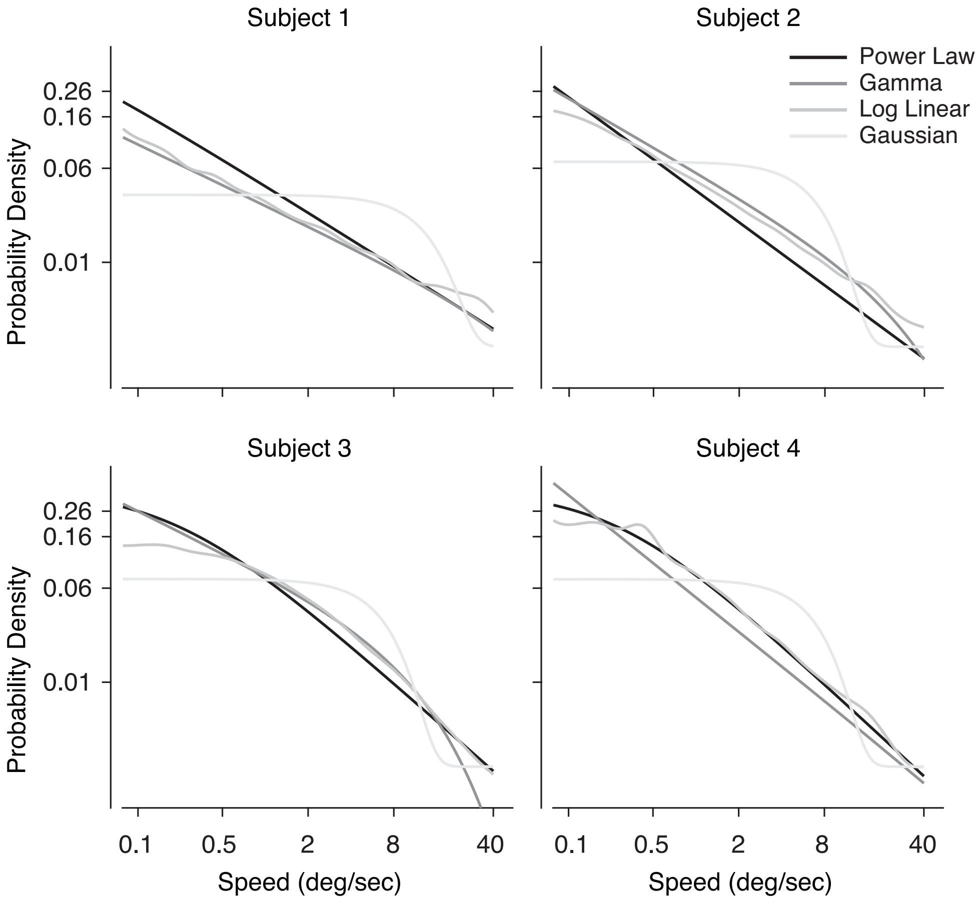

Figure 2.4: Effect of prior parameterization. The best fit prior density functions for each subject using four different parameterizations including a power-law function, a Gamma distribution, a piece-wise log-linear function, and a Gaussian distribution, respectively. Note that the Gaussian provides a relative poor fit of the data (see also Fig. 2.2B).

We further validated our model by comparing its predictions for contrast-induced biases and discrimination thresholds to subjects' data. To quantify bias, we computed the ratio of test speed to reference speed at the point of subjective equality (PSE, defined as the 50% point of the psychometric curve). If a lower-contrast test stimulus is indeed perceived to be slower, then its physical speed will need to be higher in order to match the perceived speed of the higher-contrast reference. Thus, a contrast-induced slow-speed bias is manifested by a PSE ratio greater than 1 when the test contrast is lower than the reference, and vice versa. As shown in Fig. 2.5A, subjects clearly underestimated the speeds of low-contrast stimuli, an effect that occurred at any contrast level and speed.

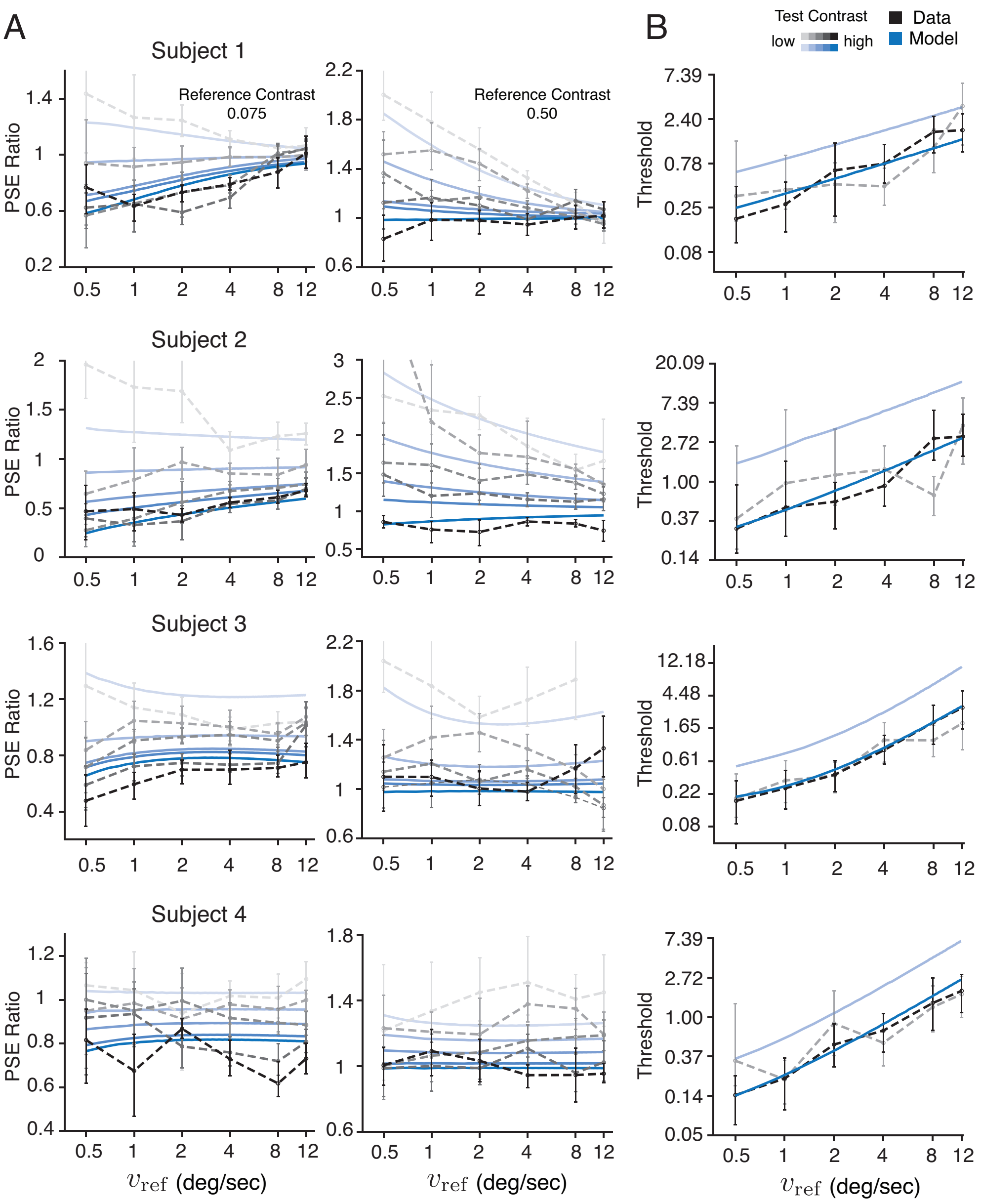

Figure 2.5: Predicted contrast-dependent bias and discrimination threshold. **A)** Ratios of the test relative to reference speed at the point of subjective equality (PSE) extracted from individual Weibull fits to the data (black) and our model (blue). Shading levels correspond to different contrast levels (0.05, 0.1, 0.2, 0.4, 0.8) of the test stimulus (darker means higher contrast). The reference stimulus has a contrast of 0.075 in left column, and 0.5 in the right column. **B)** Speed discrimination thresholds, defined as the difference in stimulus speed at the 50% and 75% points of the psychometric curve, at two different contrast levels (0.075, 0.5), extracted from individual Weibull fits to the data (black) and our model (blue). Error bars indicate the 95% confidence interval across 500 bootstrap runs.

Furthermore, subjects' thresholds increase monotonically with speed (Fig. 2.5B). While they follow Weber's law at higher speeds, they deviate from a constant Weber-fraction at slow speeds, which is well documented (McKee et al., 1986; De Bruyn and Orban, 1988; Stocker and Simoncelli, 2006). Our new model is able to capture both the contrast-induced slow-speed bias and the discrimination threshold behavior with an accuracy comparable to the original model (Stocker and Simoncelli, 2006) (also see Fig. 2.2C).

It is worth noting that the predicted higher threshold values for the low-contrast stimulus condition are not particularly evident in the data (Fig. 2.5B). Previous studies, however, have convincingly demonstrated that lower stimulus contrasts lead to higher speed discrimination thresholds under various stimulus configurations (Panish, 1988; Turano and Pantle, 1989; Horswill and Plooy, 2008; Champion and Warren, 2017). Thus we believe that the experimental design in the original study (Stocker and Simoncelli, 2006), in particular the deliberate compromise in choosing a low number of trials per condition (only 80 trials per psychometric curve) in order to test subjects over a large range of different contrast/speed combinations, may be responsible for the noisy, overlapping threshold estimates. The ability to obtain reliable threshold estimates was further limited by the use of a stair-case procedure optimized for inferring the PSE rather than the slope of the psychometric curves (see Stocker and Simoncelli (2006) for details). Future investigations will be required to fully resolve this discrepancy.

Despite its constrained nature, the model can well account for individual differences across subjects. Differences in the values of the contrast-dependent noise parameter determine individual variations in bias and threshold magnitude. In addition, when the prior exponent is close to -1, the PSE ratios are mostly constant across different speeds (e.g., Fig. 2.5A, Subject 3 and 4) (Wei and Stocker, 2017), whereas an exponent larger than -1 predicts relative biases that decrease for higher stimulus speeds (e.g., Fig. 2.5A, Subject 1).

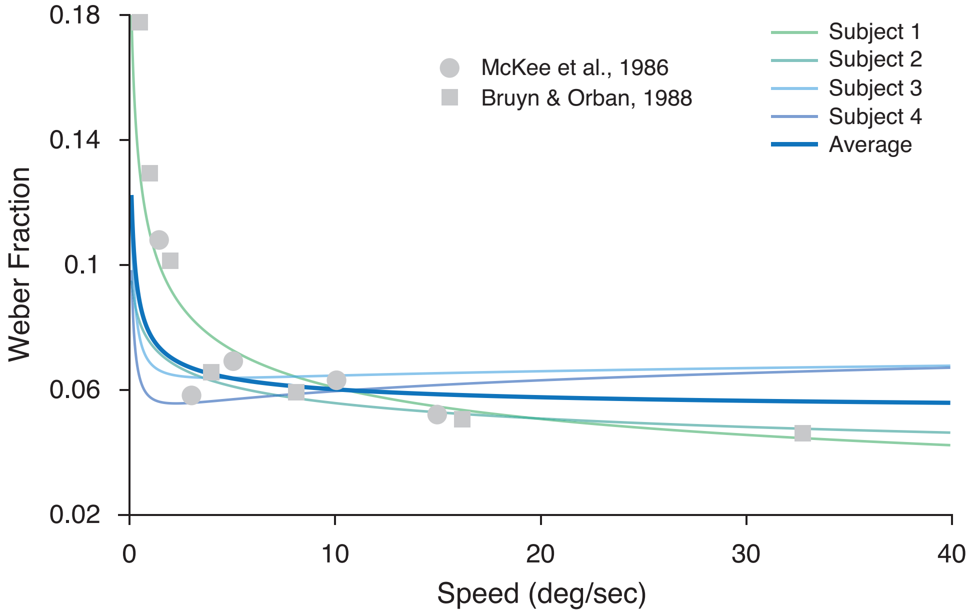

Figure 2.6: Predicted Weber fraction. Predicted Weber fractions $\Delta v/v$ based on the reverse-engineered behavioral priors of the four individuals and the average subject are shown in comparison to previously reported psychophysical measurements (McKee et al., 1986; De Bruyn and Orban, 1988). We can analytically show that the modified power-law prior with an exponent $c_0 = -1$ will predict both the constant Weber fraction at higher speed and its deviation at slow speeds (*Methods*). Note that since the Weber fraction is predicted up to a factor, the predictions are scaled to the level of the data. De Bruyn and Orban (1988) also found deviations from Weber's law at extremely high speeds (256 cycles/deg) which is not depicted here.

Lastly, previous models have mainly focused on the contrast-induced speed bias and how it can be attributed to a slow-speed prior that shifts the percept toward slower speeds for increasing levels of sensory uncertainty (Weiss et al., 2002; Hürlimann et al., 2002; Stocker, 2006; Stocker and Simoncelli, 2006; Hedges et al., 2011; Rokers et al., 2018; Lakshminarasimhan et al., 2018). While this is still the case for our new model, the monotonic increase in threshold is now also a direct consequence of the slow-speed prior: Since higher speeds are less likely, efficient coding dictates that less neural resources are allocated for their representation, resulting in a larger threshold. In fact, the predicted Weber fractions based on subjects' reverse-engineered priors closely resemble previous psychophysical measurements (Fig. 2.6; also see *Methods*).

2.3.2. Extracting the "neural prior"

The efficient coding constraint of the model predicts that the neural encoding of visual speed should reflect the stimulus prior distribution (Wei and Stocker, 2012, 2016; Ganguli and Simoncelli, 2014). Thus, if our new model is correct, then the reverse-engineered "behavioral prior" should be a good predictor of the neural encoding characteristics of visual speed. Specifically, we expect the neural encoding accuracy, measured as the square-root of the neural population FI, to match the extracted speed prior (Eq. (2.1)).

To test this prediction, we analyzed the encoding characteristics of a large population of neurons in area MT. Our analysis was based on single-cell recorded data from the macaque brain (Nover et al., 2005). The data contained repeated spike counts from 480 MT neurons responding to random dot motion stimuli moving at eight different speeds. Following the original study (Nover et al., 2005), we fit a log-normal speed tuning curve model for each neuron in the data set (Fig. 2.7A). This tuning curve model accurately described the mean firing rates of the majority of the neurons (Fig. 2.7B). Under the assumption that neural response variability was well-captured by a Poisson distribution, it is then straight-forward to compute the FI of individual neurons (Fig. 2.7A; *Methods*).

Given the inherent limitations of single-unit recorded data, assumptions about the noise and its correlation structure in the population are necessary in order to compute the population FI (Abbott and Dayan, 1999; Averbeck et al., 2006; Moreno-Bote et al., 2014; Kohn et al., 2016). We first considered the noise to be independent among the neurons in the recorded population. In this simplified scenario, the population FI is the sum of the FI of individual neurons (Fig. 2.7C). The shape of the resulting population FI is very close to a power-law function; that is, when plotted on a log-log scale it closely resembles a straight line. Two slightly different methods of calculating the FI of individual neurons, either by estimating the Fano factor explicitly or by computing the linear FI (Kanitscheider et al., 2015; Kohn et al., 2016), produced nearly identical estimates of the shape of the population FI (Fig. 2.7D; *Methods*).

Further, assuming a correlation pattern that is speed-independent simply reduces the magnitude of the population FI but does not change its overall shape compared to the independent noise assumption. However, speed tuning-dependent noise correlations between pairs of neurons have been reported for area MT (Huang and Lisberger, 2009). Thus, to assess the potential impact of such correlations (Zohary et al., 1994), we computed the linear population FI with a limited-range correlation model based on the relative speed preferences of individual neurons (Abbott and Dayan, 1999) (*Methods*). We found that although the magnitude of FI decreases with increasing correlation strength, the shape of the population FI is largely invariant within a large range of simulated correlation strengths (Fig. 2.7E). The reason why these correlations have little effect on the shape of the population FI is that the tuning characteristics of MT neurons are relatively "homogeneous" (i.e., the parameters of the tuning curve, such as the tuning width, are mostly independent of speed preference), and close to uniformly tile the logarithmic speed space (Nover et al., 2005). Thus, we argue that given the available evidence, estimating the *shape* of the population FI assuming independent noise is a reliable approximation.

The efficient coding constraint makes the additional prediction that the overall magnitude of the population FI corresponds to the total represented sensory information, and thus should be directly related to the contrast-dependent noise parameter of our observer model (Wei and Stocker, 2015, 2016; Noel et al., 2021). Although the contrast-dependent noise parameter values are consistent with the typical contrast response function of cortical neurons (Fig. 2.3), a rigorous test of the prediction requires characterization of MT speed encoding at different levels of stimulus contrasts, which is something the current data do not provide. Preliminary neural data (Stocker et al., 2009) suggest, however, that stimulus contrast indeed simply scales the population FI without changing its shape. This is intriguing given the well-documented diversity and heterogeneity by which stimulus contrast affects the shape and position of speed tuning curves in area MT (Pack et al., 2005; Krekelberg et al., 2006; Stocker et al., 2009).

2.3.3. Comparing the behavior and neural prior

Finally, we compared the extracted behavioral and neural priors. If our observer model is correct then the prior expectation with which a subject perceives the speed of a moving stimulus should

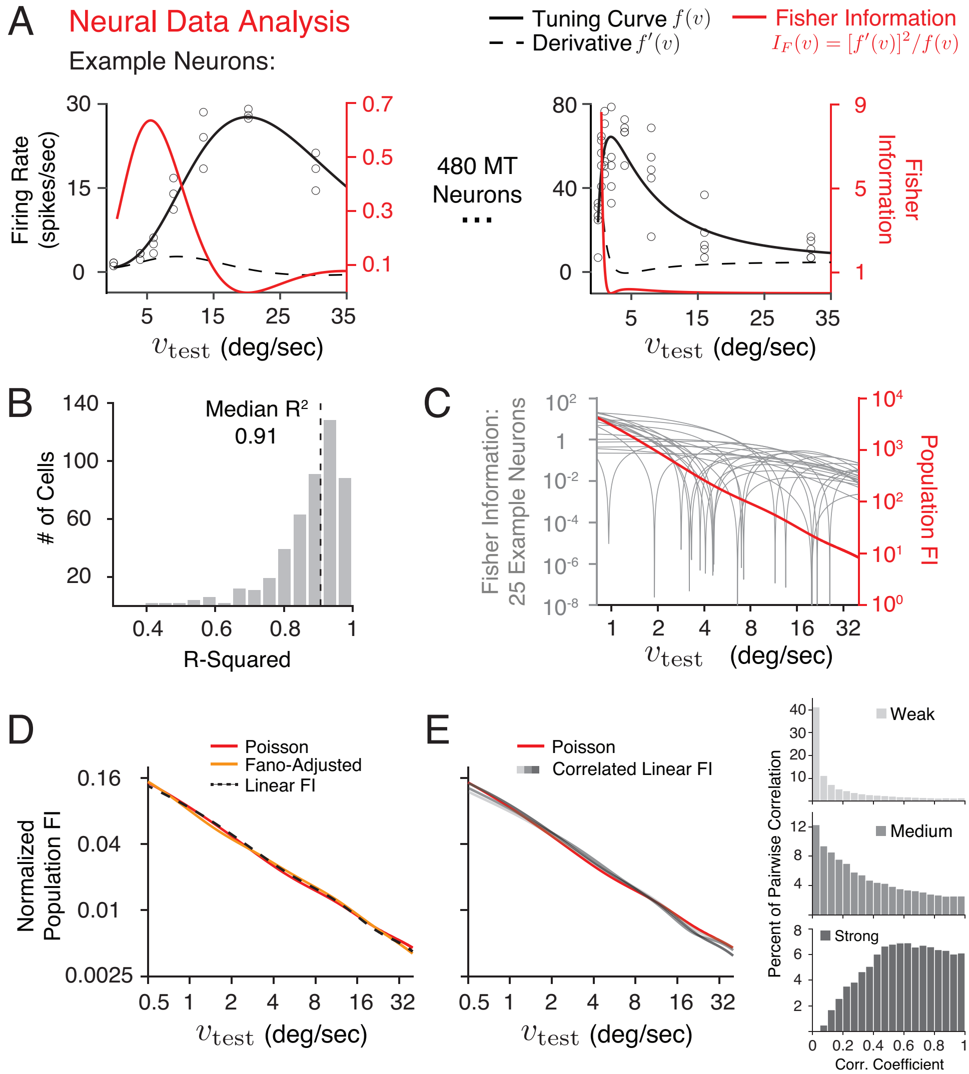

Figure 2.7: Extracting the "neural prior". According to the efficient coding constraint (Eq. (2.1)) the population Fisher information (FI) should directly reflect the prior distribution of visual speed. **A)** Single-trial mean firing rates as a function of stimulus speed v_{test} (dots) shown for two example MT neurons from the data set (Nover et al., 2005), together with their fit log-normal tuning curves (black, left y-axis) and corresponding FI (red, right y-axis) assuming a Poisson noise model. **B)** Histogram of the goodness of the log-normal tuning curve fit across all neurons measured by R^2. **C)** Individual FI of 25 example neurons (gray, left y-axis), and the population FI (red, right y-axis), calculated as the sum of the FI over all 480 neurons in the data set. **D)** The normalized square-root of population FI assuming independent Poisson noise (red), adjusted for the variance by estimating the Fano factor explicitly (orange), and the linear Fisher information (dashed black). **E)** The normalized square-root of population FI assuming independent Poisson noise (red), and the linear population FI based on a speed tuning-dependent, limited-range noise correlation model (Abbott and Dayan, 1999) for three levels of correlation strength as illustrated by the histogram of pairwise correlation coefficients on the right. See *Methods* for details.

be quantitatively identical to, and thus predictive of, the stimulus distribution neural encoding is optimized for. Figure 2.8A shows the extracted behavioral prior of every subject and the neural prior. The prior distributions are indeed very similar and are consistent with a power-law function with an exponent of approximately -1.

In order to quantitatively assess the effective similarity between the behavior and neural prior, we constructed a “neural observer” model for which the prior was fixed to be the neural prior extracted from the MT data; only the contrast-dependent noise parameters $h(c)$ were free parameters. We used a cross-validation procedure to compare this neural observer with the unconstrained observer model and the original model (Stocker and Simoncelli, 2006). As illustrated in Fig. 2.8B, the validation performances are highly similar across all three models and closely match the performance of individual Weibull fits. This comparison demonstrates several aspects. First, it confirms that the behavioral and neural prior are behaviorally indistinguishable and thus effectively equivalent; if the neural data did not exist, we would have been able to accurately predict the encoding accuracy of MT neurons at the population level. Second, it highlights the excellent quality of the Bayesian observer model as its account of human behavior is close to that of the best possible parametric description of the data (i.e. individual Weibull fits). And finally, it shows that the complexity of our new observer model is appropriate and does not lead to over-fitting.

2.3.4. Weber-Fechner law

The power-law shape of the behavioral and neural prior distribution also sheds a new normative light on the interpretation of Weber’s law. Famously, Fechner proposed that Weber’s law emerges from a logarithmic neural encoding of a stimulus variable (Fechner, 1860). Neural encoding of visual speed in area MT is indeed considered logarithmic (Nover et al., 2005; Burge and Geisler, 2015): When analyzed in the logarithmic speed domain, the tuning curves of MT neurons are approximately bell-shaped, scale-invariant, and tile the stimulus space with near-uniform density (Nover et al., 2005; Pack et al., 2005).

With our new model, we have already demonstrated that a modified power-law prior (Eq. (2.2)) with an exponent of approximately -1 can well account for Weber’s law behavior and the devia-

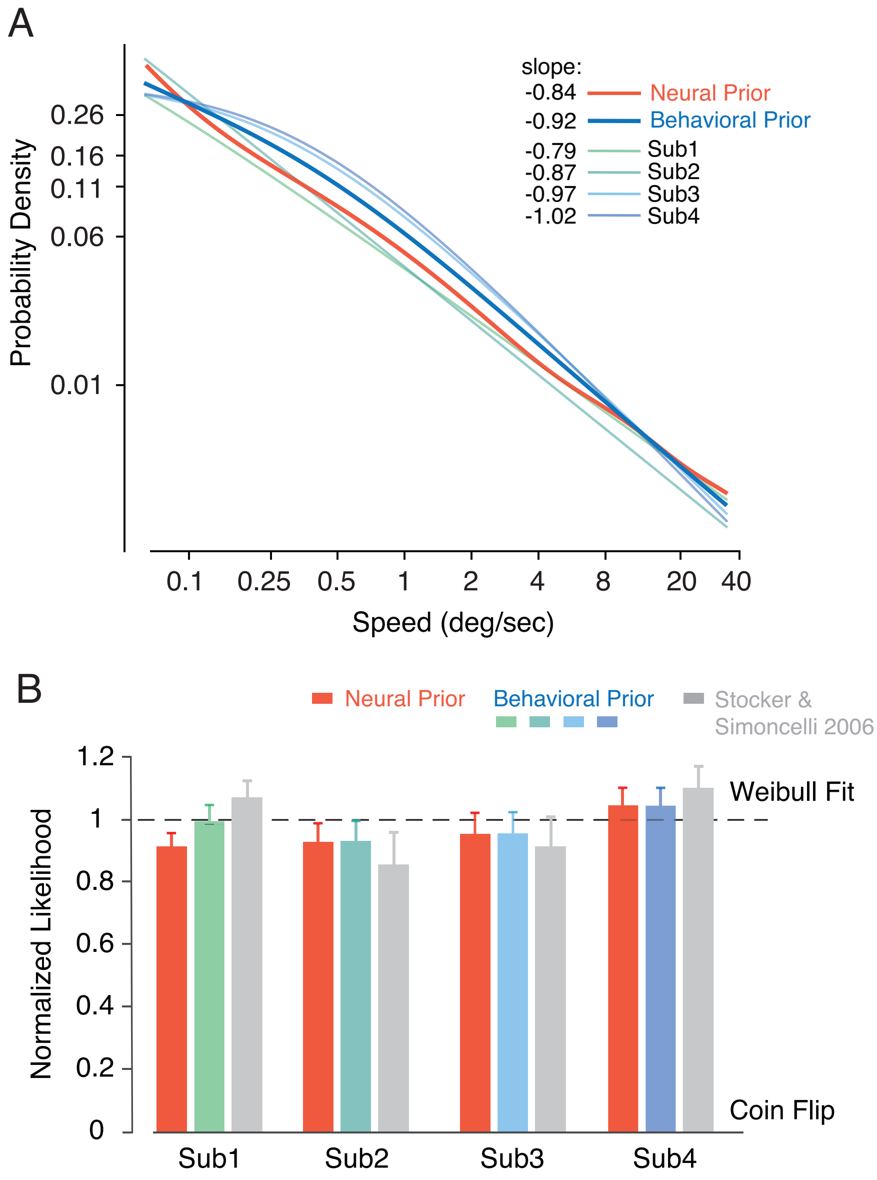

Figure 2.8: Comparing neural and behavioral prior. **A)** Reverse-engineered behavioral priors of every subject and their average (dark blue) superimposed by the neural prior (red). Slope values are computed from a linear fit of the curves in log-log coordinates. **B)** The cross-validated log-likelihood of the model using the subject's best-fitting behavioral prior (blue) or the fixed neural prior (red), and the log-likelihood of the original model (Stocker and Simoncelli, 2006) (gray). The log-likelihood value is normalized to the range defined by a "coin-flip" model (lower bound) and the Weibull fit to each psychometric curve (upper bound). Error bars represent ± SD across 100 validation runs according to a five-fold cross-validation procedure. See *Methods* for details.

tion from it at slow speeds (Fig. 2.6). We can now show that this power-law prior also predicts logarithmic neural encoding. Specifically, one way to implement the efficient coding constraint (Eq. (2.1)) is to assume a homogeneous neural encoding (i.e., identical tuning curves that uniformly tile the sensory space) of the variable of interest *transformed* by its cumulative distribution function (CDF) (Ganguli and Simoncelli, 2010; Wei and Stocker, 2012; Wang et al., 2016a), which is sometimes also referred to as histogram equalization (Acharya and Ray, 2005). With an exponent $c_0 = -1$, the CDF of the speed prior is exactly the logarithmic function that well-described MT tuning characteristics (Nover et al., 2005) (*Methods*). Thus, the Bayesian observer model constrained by efficient coding provides a normative explanation for both Weber's law and the logarithmic encoding of visual speed in area MT.

2.4. Discussion

We presented a Bayesian observer model constrained by efficient coding for human visual speed perception. We fit this model to existing human 2AFC speed discrimination data recorded over a wide range of stimulus contrasts and speeds, which allowed us to reverse-engineer the "behavior prior" that best accounts for the psychophysical behavior of individual subjects. In addition, we analyzed the population encoding accuracy of visual speed based on an existing set of single-cell recordings in area MT, thereby extracting the "neural prior" according to the efficient coding constraint of our observer model. We found that the behavioral prior estimated from the psychophysical data accurately predicts the neural prior reflected in the encoding characteristics of the MT neural population.

Our results provide a successful, quantitative validation of the Bayesian observer model constrained by efficient coding in the domain of visual speed perception. We demonstrate that this model can accurately account for the behavioral characteristics of bias and threshold in visual speed perception if subjects' prior belief about the statistical distribution of visual speed resembles a power-law function with an exponent of approximately -1. Cross-validation revealed no significant difference between the best possible parametric description of the behavioral data (i.e., individual Weibull fits) and our model fits. Compared to the original, more flexible Bayesian model formulation (Stocker and Simoncelli, 2006), the added efficient coding constraint results in estimates of behavioral priors that are not only much more consistent across subjects but also remarkably predictive of the population encoding characteristics of neurons in the motion-sensitive area MT in the primate brain. Our work substantially strengthens the evidence for the slow-speed prior interpretation of motion illusions (Weiss et al., 2002; Stocker and Simoncelli, 2004, 2006; Welchman et al., 2008; Sotiropoulos et al., 2014; Jogan and Stocker, 2015; Senna et al., 2015) (but see (Rideaux and Welchman, 2020)) by the demonstrated quantitative support from electrophysiological neural data.

We also offer an explanation for why certain perceptual variables have a logarithmic neural representation and thus follow Weber's law (Fechner, 1860). According to our model, logarithmic encoding and Weber's law both follow from the efficient representation of a perceptual variable with a power-law prior distribution. We thus predict that perceptual variables that conform to Weber's law have power-law distributions with an exponent of approximately -1 *and* are logarithmically encoded in the brain (although alternative encoding solutions that satisfy the efficient coding constraint are possible, see (e.g. Wei and Stocker, 2015)). Indeed, perceptual variables that are known to approximately follow Weber's law such as weight (Fechner et al., 1966), light intensity (Treisman, 1964), and numerosity (Nieder and Miller, 2003; Cheyette and Piantadosi, 2020; Prat-Carrabin and Woodford, 2021), exhibit heavy tails in their statistical distributions under natural environmental conditions (Dehaene and Mehler, 1992; Dror et al., 2004; Peters et al., 2015; Piantadosi and Cantlon, 2017), a defining feature of a power-law function. Conversely, any deviation from Weber's law and the logarithmic encoding should be reflected in deviations of the statistical stimulus distributions from a power-law function. Future studies of natural stimulus statistics, modeling of psychophysical data, and neural recordings will be needed to further and more quantitatively validate the generality of this prediction.

Other recent work has used efficient coding assumptions to link perceptual discriminability to the statistical prior distribution of perceptual variables (Gu et al., 2010; Ganguli and Simoncelli, 2016; Sims, 2018). Our approach is a substantial step forward in that it embeds this link within a full behavioral observer model. Thus rather than relying on a single summary metric of behavior (i.e. discrimination threshold), the predictions of our model are constrained by the full richness of the psychophysical data, i.e., every single datum in the set. This not only provides a much more stringent test of the observer model but also permits more robust and precise predictions of neural coding accuracy and priors.

The presented comparison between behavioral and neural prior is limited to the extent that there were substantial experimental differences between the behavioral and neural data. For example, we compared human with non-human primate data and estimated the neural tuning characterization based on single-cell responses to random-dot motion rather than the broadband, drifting grating stimuli used in the psychophysical experiment. Yet, the surprisingly accurate match of the extracted neural and behavioral priors suggests that they may reflect the "true" stimulus prior, in which case these differences in stimuli and model systems should indeed matter little because the stimulus prior is largely a property of the environment and not the observer nor the particular stimulus pattern. Recent studies have demonstrated that it is possible to quantitatively characterize the accuracy with which a perceptual variable is represented in the human brain using voxel-level encoding models of functional magnetic resonance imaging signals (Van Bergen et al., 2015). Future work may exploit this technique to validate and potentially refine our estimates of the "neural prior" in human subjects. Such work would also permit a more thorough investigation of individual differences at both behavioral and neural levels through matched task and stimulus designs.

The specific shape of the extracted neural and behavioral prior depends on the assumed efficient coding objective. The chosen efficient coding constraint (Eq. (2.1)) results from the objective to maximize the mutual information between neural representation and stimulus (Wei and Stocker, 2016). It is possible, although unlikely given the exceptional good quantitative match, that with a different combination of efficient coding constraint and loss function, a power-law prior with a different exponent could also be consistent with both the behavioral and neural data (see Wang et al., 2012; Morais and Pillow, 2018; Rast and Drugowitsch, 2020). This is difficult to validate conclusively without access to an accurate characterization of the stimulus prior of visual speed, as the search space over all possible combinations is extensive. Previous work has shown, however, that the encoding characteristics in early visual cortex for visual stimulus variables for which good estimates of the stimulus prior exist (e.g., luminance contrast and local orientation) are closely accounted for by the mutual information maximization objective (Wang et al., 2012).

An important assumption of our observer model is that the neural and behavioral priors not only match but are also consistent with the statistical distribution of visual speeds in the natural environment ("stimulus prior"). As such, our results predict that the stimulus prior approximates a power-law distribution that lies within the range given by the neural and behavioral priors shown in Fig. 2.8A. However, empirical validation of this prediction by directly measuring the stimulus prior distribution is rather challenging. Object motion, but also the ego-motion of the observer in terms of its body, head, and eye movements all contribute to the visual motion signal. Thus, the precise characterization of the visual speed distribution would require accurate measurements and calibrations of these different types of motions, as well as of the algorithm used to extract the motion information from the visual signal. Previous studies have approximated these relative movements to various degrees and used different algorithms to extract local visual speed from spatio-temporal images, resulting in different characterization of the prior distribution (Dong and Atick, 1995; Roth and Black, 2007; Baker et al., 2011; Sinha et al., 2021). However, common to all these measured stimulus priors is that they have higher probabilities at slow speeds and form long-tailed distributions. Future work using more comprehensive data (DuTell et al., 2020) may provide a better characterization of visual speed priors under ecologically valid, natural conditions.

We expect our model and analytic approach to be applicable to any other perceptual variable and task that exhibit characteristic patterns of perceptual biases and discrimination thresholds. However, of particular interest and posing a strong test of our model are changes in perceptual bias and threshold that are induced by spatio-temporal context such as adaptation aftereffects or the tilt-illusion (Schwartz et al., 2007, 2009; Clifford et al., 2007). It is traditionally assumed that these biases are caused by a mismatch in expectation between encoding and decoding (i.e. the "coding catastrophe" (Schwartz et al., 2007)), which is in sharp contrast to one of the main features of our model. Preliminary results are promising (Wei et al., 2015; Wei and Stocker, 2017). However, more quantitative analyses are necessary to test how well the framework can account for the data and what neural and behavioral priors it will predict.

In summary, within the context of visual speed perception, we have demonstrated that the Bayesian observer model constrained by efficient coding has the potential to provide a unifying framework that can quantitatively link natural scene statistics with psychophysical behavior and neural representation. Our results represent a rare example in cognitive science where behavioral and neural data *quantitatively* match within the predictions of a normative theory.

CHAPTER 3

CONTEXTUAL MODULATION OF ORIENTATION ENCODING IN TILT ILLUSION

The work presented in this chapter is performed in collaboration with Huseyin O. Taskin, Geoffrey K. Aguirre, and Alan A. Stocker. I contributed to the conceptualization, funding acquisition, data collection, formal analysis, methodology, validation, software, visualization, and writing of this work.

Abstract

The perceived orientation of a grating is dependent on its surrounding orientation. This "tilt illusion" is one of the examples of how perception is influenced by both the spatial and temporal context. Explanations of these effects tend to focus on the mechanism by which the response properties of sensory neurons are modified in response to stimulus context. However, linking changes in neural encoding to behavior is difficult because it requires specific assumptions regarding both how stimulus information is represented and also how it is interpreted. Here, based on the lawful relationship between perceptual bias, variance, and Fisher information, we developed a method that allows us to directly characterize the contextual modulations of sensory encoding based on psychophysical behavior in a tilt-illusion experiment. We found that for a non-oriented surround, sensory encoding accurately reflected the statistics of visual orientations in natural scenes. However, with oriented stimulus surrounds, encoding accuracy was significantly boosted at the corresponding surround orientation. Our results are consistent with the notion that contextual modulations of sensory encoding represent a form of efficient coding where encoding accuracy is optimized toward the conditional stimulus statistics within the specific context.

3.1. Introduction

Human perception of sensory stimuli is influenced by the context within which they are presented. For instance, the perceived orientation of a stimulus can be altered by preceding it with a series of stimuli with similar orientations. Similarly, the perceived orientation may differ in the presence of an adjacent spatially oriented surround. These phenomena, known as the tilt aftereffect (Magnussen and Johnsen, 1986) and tilt illusion (Gibson and Radner, 1937), illustrate how both temporal and spatial context can affect our perception. In both cases, the changes in perception are associated with a characteristic pattern of repulsive perceptual bias (Mitchell and Muir, 1976) and reduction in discrimination threshold (Regan and Beverley, 1985) near the adapting orientation. Lastly, it has been shown that even in the absence of a systematic relationship between stimuli, our perception can still be influenced by the immediate past history of stimuli and response sequences (Fischer and Whitney, 2014; Fritsche et al., 2020).

From a normative perspective, it can be beneficial for our sensory representation to adapt to the local temporal and spatial statistics in order to maximize information capacity (Wark et al., 2007). Such sensory adaptation in response to changes in stimulus statistics has been demonstrated at the level of single neurons (Fairhall et al., 2001; Wark et al., 2009) and across neural population (Benucci et al., 2013). In the case of the tilt aftereffect, electrophysiology studies have shown that adaptation causes neurons in the early visual cortex to both suppress their responses at and shift their tuning preferences away from the adapting orientation (Dragoi et al., 2000, 2001). Theoretical work suggested that these observed effects can be predicted based on a model of the relationship between divisive normalization in sensory neurons and natural scene statistics (Schwartz et al., 2009). Similar changes in neural activities have been observed in other studies using fMRI (Fang et al., 2005) and EEG (Rideaux et al., 2023) measurements.

To link these changes in sensory representation to perceptual behavior, however, requires specific assumptions about how sensory signals are transformed into perceptual estimates. Formally, perception is commonly described using an encoding-decoding framework, where stimulus θ is first encoded by noisy neural measurement m. An estimate $\hat{\theta}$ is then generated based on a decoder $\hat{\theta}(m)$ (Fig. 3.1A). Since both the encoding and decoding processes can be altered during sensory adaptation, it is difficult to delineate their individual contributions to behavior. In fact, previous modeling attempts have assumed decoders ranging from completely fixed and "unaware" of changes (which leads to a "coding catastrophe", see Schwartz et al. (2007); Seriès et al. (2009)), to decoders that match the changing encoders (Stocker and Simoncelli, 2005; Schwartz et al., 2009), and to decoders that dynamically update the stimulus prior simultaneously as the encoders (Fritsche et al., 2020).

In this article, we directly measure the changes in sensory encoding in terms of Fisher Information (FI) across different adaptation contexts, by leveraging a lawful relationship between estimation bias, variance, and FI (Fig. 3.1A-C). Our method requires only minimal assumptions regarding the decoder, and thus is agnostic to potential changes in the decoding process (Fig. 3.1D-F). We apply our methods to estimation data collected in a tilt illusion experiment, and found that the changes in encoding are consistent with an efficient coding account for which the sensory representation is optimized towards the conditional statistics of orientation for the specific context.

3.2. Methods

3.2.1. Experiment design

Five participants were asked to estimate the orientation of a briefly presented Gabor grating. The orientation of the grating was randomly sampled between 0 and 180 deg. The stimulus was presented for 1500 ms, had a spatial frequency of 1 cyc/deg with random phase, and had a peak contrast of 0.2 with 1 Hz contrast modulation. After an average 4.5 sec delay, participants rotated a line probe with a two-button keypad to report their response. The average response window is 4.0 sec, as indicated by the line probe gradually fading out. There were a total of 20 trials within each block, and the trials were separated by an interval of 2 sec. Participants were allowed to take short breaks in between the blocks.

In each block, gratings were presented within a stimulus surround consisting of either non-oriented filtered noise, or gratings with one of two possible fixed orientations (35 or 145 deg) with the same spatial frequency. The surround had a diameter of 18 visual degrees, with its center 10 degrees being the target grating. The three types of blocks were interleaved randomly in groups of three, and each participant finished a total of 60 blocks (a total of 1200 trials, 400 trials for each condition).

3.2.2. Data analysis

Bias and variance

We use a procedure similar to that of Noel et al. (2021). We model each participant as a generic estimator of the true orientation θ. Participants' responses $\hat{\theta}$ are stochastic. For a fixed θ, they will produce a distribution $p(\hat{\theta}|\theta)$. We denote the bias $b(\theta)$ and variance $\sigma^2(\theta)$ of the participants as (Fig. 3.1A):

$$b(\theta) = E_{\hat{\theta} \sim p(\hat{\theta}|\theta)}[\hat{\theta}] - \theta, \tag{3.1}$$

$$\sigma^2(\theta) = E_{\hat{\theta} \sim p(\hat{\theta}|\theta)}[\hat{\theta}^2] - E_{\hat{\theta} \sim p(\hat{\theta}|\theta)}[\hat{\theta}]^2. \tag{3.2}$$

Note that both are defined as a function of θ. To compute these quantities from response data, we apply a sliding window analysis with a size of 12 deg and a step size of 0.5 deg. The mean and variance are computed within each window, with the true θ chosen as the center of that window.

Cramer-Rao lower bound

Define the encoding model as $p(m|\theta)$. Fisher Information (FI) is the variance of the score function:

$$I_F(\theta) = E_{m \sim p(m|\theta)}[(\frac{\partial}{\partial \theta} \log p(m|\theta))^2], \tag{3.3}$$

which quantifies the accuracy of the encoding. In particular, Cramer-Rao lower bound states that the variance of any unbiased estimator $\hat{\theta}$ is bounded below by FI as $\mathrm{Var}(\hat{\theta}|\theta) \geq 1/I_F(\theta)$. For any estimator $\hat{\theta}$ with a bias is given by $b(\theta)$, we have (Casella and Berger, 2021):

$$\mathrm{Var}(\hat{\theta}|\theta) \geq \frac{[1 + b'(\theta)]^2}{I_F(\theta)}. \tag{3.4}$$

Here $b'(\theta)$ denotes the derivative of the bias $b(\theta)$. That is, for any generic estimator that does attain the Cramer-Rao lower bound, there is a lawful relationship between encoding accuracy and the bias and variance of the estimates (Wei and Stocker, 2017). Thus, we obtain a lower bound on the FI based on $b(\theta)$ and $\sigma^2(\theta)$ extracted from individual subjects (see Fig. 3.1C and Fig. A.1):

$$I_F(\theta) \geq \frac{[1 + b'(\theta)]^2}{\sigma^2(\theta)}. \tag{3.5}$$

In our analysis, we further assumed the encoding FI is at the lower bound. Decoders commonly used in the literature, such as maximum likelihood and Bayesian estimation, are both in this category.

To confirm the validity of our analysis, we simulated an observer with anisotropic encoding with the peak FI at 90 degrees (Fig. 3.1D), and three arbitrary decoders. Although each decoder produces a distinct pattern of estimation bias and variance, they all attain the Cramer-Rao lower bound (Fig. 3.1E). Thus, we can correctly recover the encoding FI in all three cases based on Eq. 3.5 (Fig. 3.1F).

Link to stimulus prior

Efficient coding hypothesizes that neural coding aims to maximize information transmission, given the relevant biological constraints. One possible solution to the efficient coding problem is expressed in terms of FI (Wei and Stocker, 2016):

$$p(\theta) \propto \sqrt{I_F(\theta)}. \tag{3.6}$$

Thus, we can derive the stimulus prior for which the neural encoding is optimized, assuming the representation is indeed efficient, based on the normalized FI:

$$p(\theta) = \frac{\sqrt{I_F(\theta)}}{\int \sqrt{I_F(\theta)}\, d\theta}. \tag{3.7}$$

The denominator, $\int \sqrt{I_F(\theta)}\, d\theta$, also quantifies the total amount of information in the encoding. In our results below, we will use the normalized (root) FI as the primary measure of sensory encoding.

3.3. Results

We first examine the behavior of the combined subject by examining the bias $b(\theta)$, that is, the systematic error in participants' estimate of the stimulus orientation. In all three surround conditions, the subject showed an oblique bias (Jastrow, 1892), that is, the perceived orientation is biased away from the cardinal directions (i.e., vertical and horizontal) and toward the obliques. In addition, having an oriented surround induced a robust repulsive bias (i.e., away from) at the surround orientation on top of the overall bias pattern (Fig. 3.2A). See Fig. A.1 for similar plots for

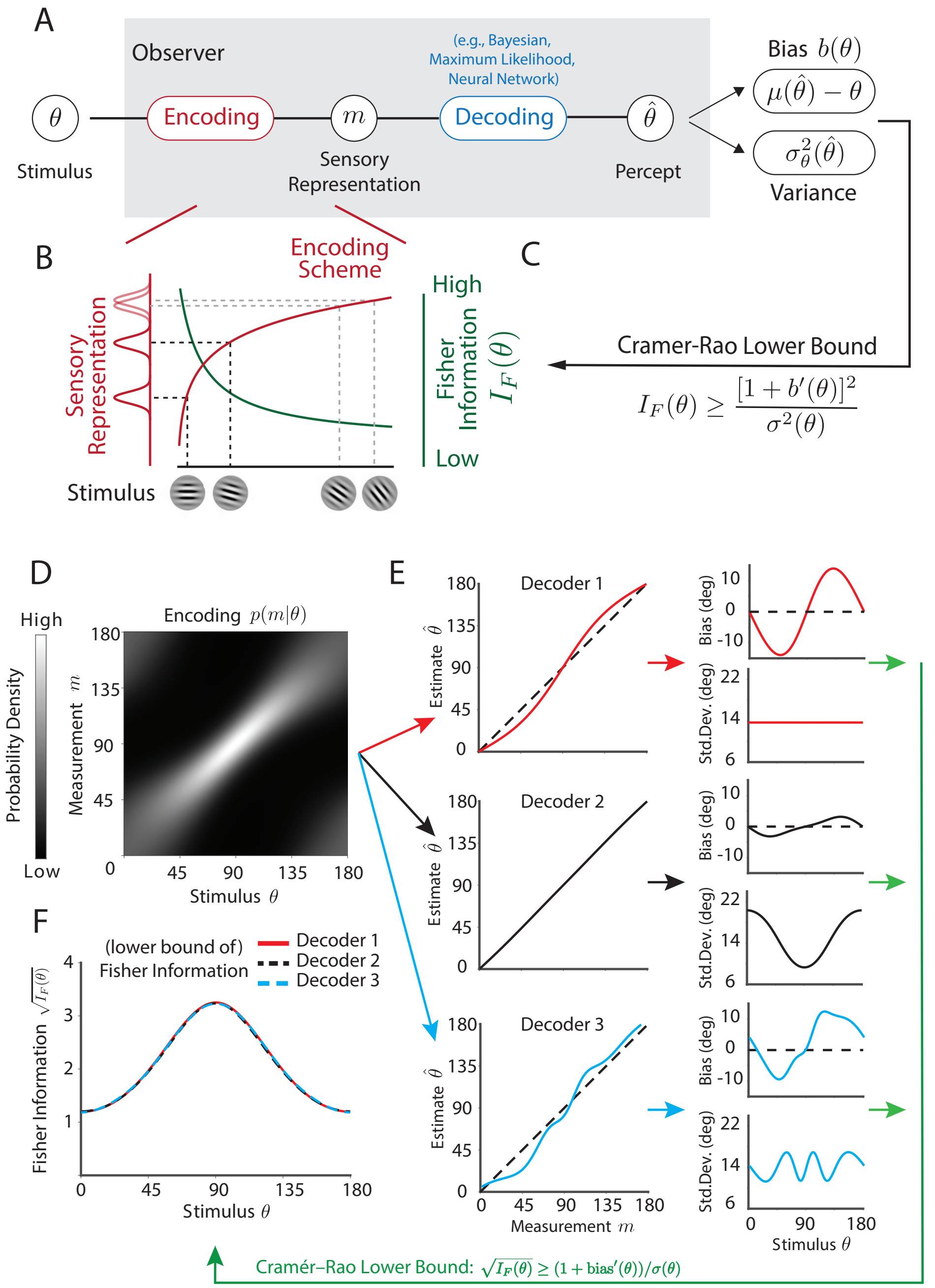

Figure 3.1: Theoretical framework and data analysis. Figure adapted from Noel et al. (2021). **A)** The encoding–decoding framework for modeling perception. **B)** Example encoding with anisotropic coding accuracy as a function of the stimulus, quantified by the Fisher Information (FI). **C)** Cramer–Rao lower bound specifies the lawful relationship between estimation bias, variance, and FI. **D)** An simulated observer with anisotropic encoding, for which the peak FI is at 90 deg. **E)** Three arbitrary decoders produce distinct patterns of estimation biases and variances. **F)** Estimated encoding FI based on the lower bound.

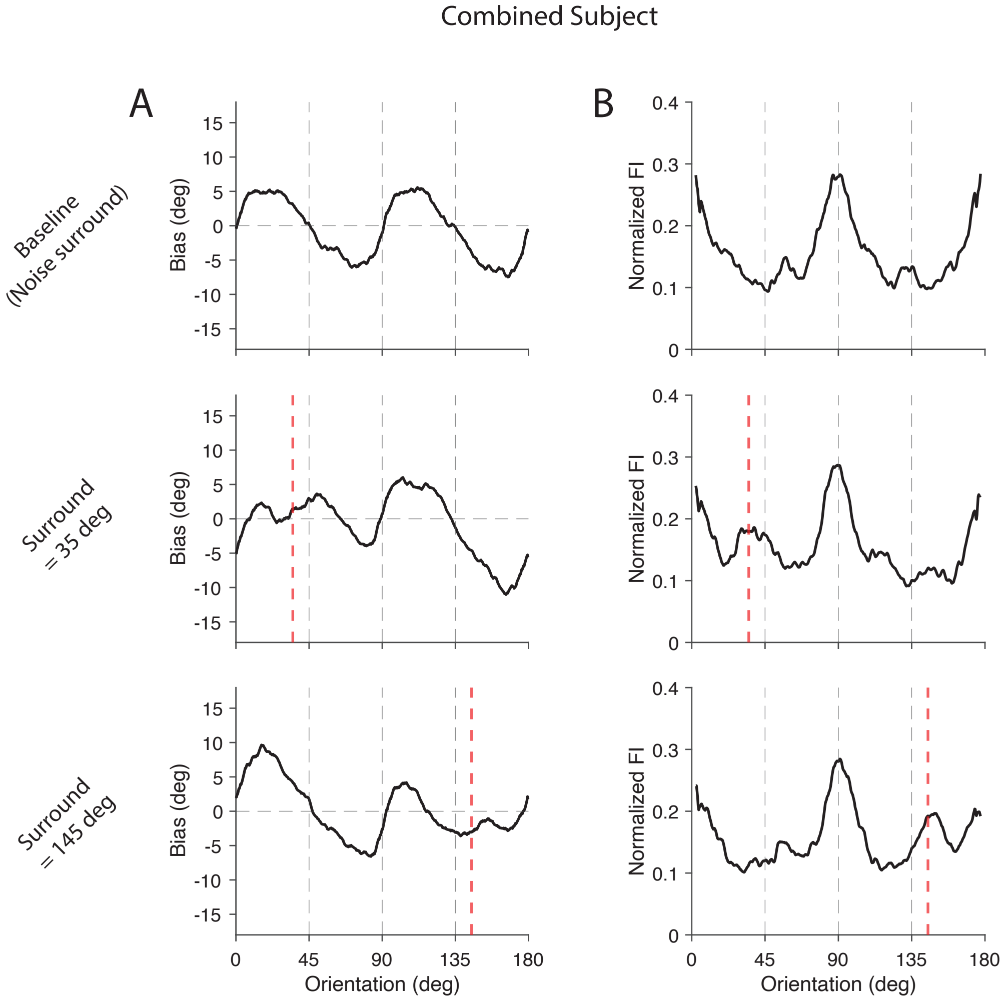

Figure 3.2: Average bias and normalized Fisher Information (FI) for the combined subject. **A)** Average response bias as a function of the target orientation for the combined subject. **B)** Normalized FI as a function of orientation for the combined subject. The three rows in both panels correspond to the baseline (noise surround) and two oriented surround conditions, respectively. Red dotted lines indicate the surround orientations. See Fig. A.1 for the same plot for individual subjects, including the raw response data.

individual subjects, including the raw response data.

To characterize the sensory encoding across the three conditions, we applied our method described above to extract the normalized FI (Eq. 3.5, 3.7). We found that in the baseline condition, FI peaked at cardinal directions. The shape of FI qualitatively matched previously measured distribution of orientations in natural images (Coppola et al., 1998; Girshick et al., 2011), consistent with the idea that the encoding is optimized towards the natural orientation statistics (Fig. 3.2B).

When the stimulus is presented within an oriented surround, the FI near the surround orientation is significantly boosted (Fig. 3.2B). That is, center orientations similar to the surround are represented with increased accuracy. It has been shown that there exists a strong dependency between adjacent spatial locations in natural scenes: Felsen et al. (2005) calculated the distribution of the orientation differences between the center and surround regions from a set of natural images. They found that the distribution is centered at 0 deg, with a variance of around 10 deg. Similar findings have been observed for consecutive frames in natural movies (Felsen et al., 2005; Schwartz et al., 2007; Van Bergen and Jehee, 2019). Thus, the surround orientation provides a reliable expectation for the possible center orientations. Therefore, this local increase in FI can be explained as optimizing encoding towards the conditional orientation distribution, informed by the surround context. Lastly, we have conducted the same analysis for individual subjects, and have found a consistent pattern across all five participants (Fig. A.1).

To highlight the changes in encoding due to spatial context, we define an orientation adaptation kernel, as the difference in the normalized FI between the surround and baseline condition (Fig. 3.3). In both cases, the adaptation kernel is clearly unimodal, with a prominent peak at the surround orientation. There also seems to be an overall reduction in FI further away from the peak, but it did not reach statistical significance in our current data.

3.4. Future Work

We developed a method for extracting the FI of sensory encoding based on psychophysical estimation data. We applied our method to participants' responses in a tilt-illusion experiment, and found that

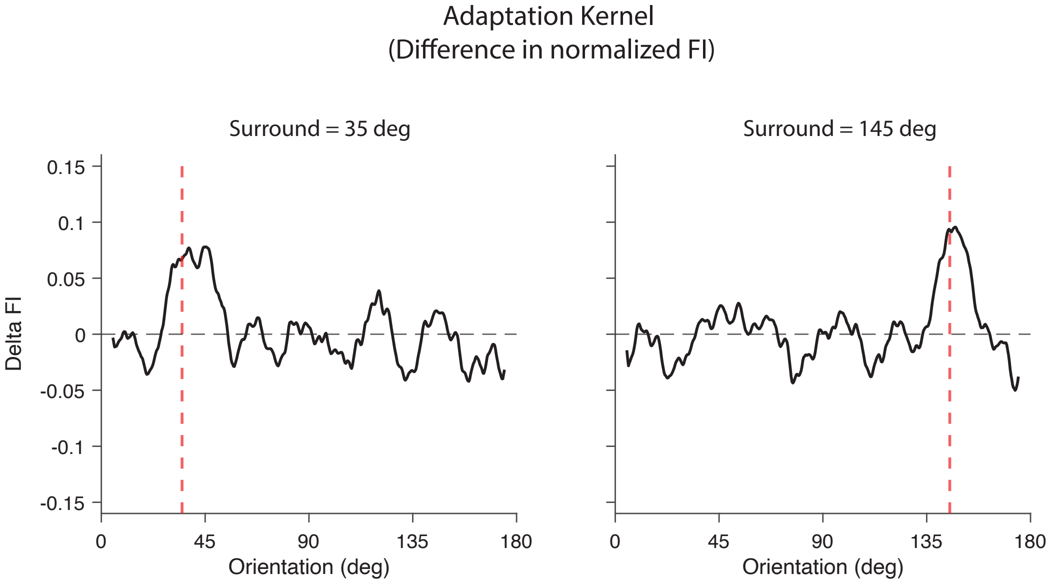

Figure 3.3: Orientation adaptation kernel. Adaptation kernel, defined as the difference in the normalized FI between the two surround and baseline conditions, computed from the combined subject. Red dotted lines indicate the surround orientations.

the changes in encoding are consistent with an efficient coding account, for which the representation of orientation is optimized toward the spatial conditional statistics informed by the surround context.

Similar to Chapter 2, we would like to further validate our method against direct physiological measurements of sensory encoding. In particular, we are interested in quantifying the neural coding accuracy based on blood-oxygen-level-dependent (BOLD) signals recorded when participants are viewing orientation stimuli. Below, we will provide a description of the analysis methods, together with some preliminary results for the no-surround condition based on previous data collected in Van Bergen et al. (2015).

The stimulus design in Van Bergen et al. (2015) is similar to what we have described above, except that the gratings are presented without any surround context. To quantify the neural responses to each stimulus presentation, ROIs for each subject are defined with a separate retinotopy session. The voxels from the V1, V2, and V3 regions are selected for subsequent analysis. The voxel response for each trial is computed by averaging the BOLD time course within a 4-second window after stimulus onset, after adding 4 seconds to account for the delay in the hemodynamic response function (HRF).

To characterize orientation encoding, Van Bergen et al. (2015) developed a probabilistic encoding model for the voxel responses:

$$m_i = \sum_{k=1}^{K} W_{ik}(f_k(\theta) + \eta_k) + v_i, \tag{3.8}$$

where m_i's are the individual voxels, W is the weight matrix, and $f_k(\theta)$'s are basis orientation tuning functions. η_k's are the encoding noise specific for each basis function, and the v_i's are residual noise for each voxel. Each tuning function with a preference of ϕ_k is of the form:

$$f_k(\theta) = \max(0, \cos(\theta - \phi_k))^5. \tag{3.9}$$

The encoding model defines the likelihood function $p(m|\theta)$ over all the voxels in the ROI. The complete form of the encoding model is a multivariate Gaussian distribution. Denote the variance of η_k's as σ^2, and the residual variance v_i's as a vector τ^2. The mean and covariance are given by:

$$\mu = W\vec{f}(\theta), \quad \Sigma = \rho\boldsymbol{\tau}\boldsymbol{\tau}^T + (1-\rho)I \circ \boldsymbol{\tau}\boldsymbol{\tau}^T + \sigma^2 W W^T, \tag{3.10}$$

respectively (see Van Bergen et al. (2015); Van Bergen and Jehee (2021)). Note $W, \rho, \boldsymbol{\tau}$ and σ^2 are the free parameters of the model. These parameters are estimated using a two-step procedure (Van Bergen et al., 2015): First, W is estimated using linear regression, and then τ, ρ, σ are estimated using maximum likelihood with the fixed $\hat{W}$. We re-implemented the model in PyTroch, allowing us to take advantage of automatic differentiation and GPU computation:
https://github.com/lingqiz/Orientation-Encode/blob/main/analysis/encode.py

This encoding model can be used to decode orientation based on the voxel activity. For example, the posterior mean estimator is as follows:

$$\hat{\theta}(m) = \int \theta \; p(\theta|m) d\theta = \int \theta \; \frac{p(m|\theta)p(\theta)}{\int p(m|\theta)p(\theta)d\theta} d\theta \tag{3.11}$$

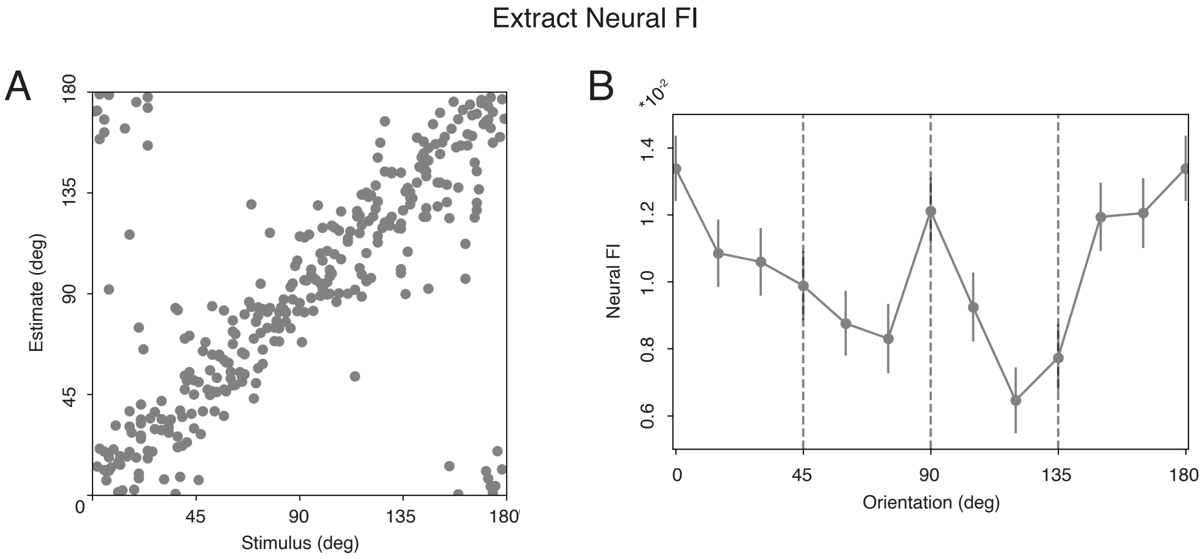

Figure 3.4: Extract neural FI. **A)** Decoding results for one example subject from Van Bergen et al. (2015) based on the probabilistic decoder. **B)** The neural Fisher Information as a function of orientation for the combined subject based on 18 individuals in Van Bergen et al. (2015). FI is defined as the mean of the negative second derivative of the orientation likelihood, computed from trials in the validation data. Error bars represent ±2 SEM.

In Fig. 3.4A, we show the result of orientation decoding from one example subject using a cross-validation procedure: For each iteration, we withhold 10 trials from the data. We fit the encoding model (Eq. 3.8) on the remaining data (~ 300 trials) using the two-step procedure described above. The estimates are then obtained for the 10 trials according to Eq. 3.11. The process is then repeated until every trial has become a validation trial exactly once.

Our goal here is to see if it is possible to quantify the coding accuracy based on the voxel encoding model. As a proof of concept, below we demonstrate that we can extract neural FI from the BOLD activities, based on the data from Van Bergen et al. (2015). Concretely, for a stimulus θ and the corresponding voxel response m on a given trial, the observed FI is defined at θ as the second derivative of the likelihood function:

$$I_F(\theta)^* = \frac{\partial^2 p(m|\theta)}{\partial \theta^2}. \tag{3.12}$$

Here the likelihood $p(m|\theta)$ is obtained by fitting the model (Eq. 3.10) using the same cross-validation procedure as decoding. We compute the $I_F(\theta)^*$ for each trial, and then combined the data from all 18 individual subjects in the dataset, and applied a sliding window average to compute the average neural FI for the combined subject (Fig. 3.4) as a function of stimulus orientation. The results corroborate the FI we extracted based on behavior in the baseline condition (Fig. 3.2B): The FI peaks at the cardinal directions, and is the lowest near the obliques.

Based on our preliminary results, it appears that we can accurately characterize neural coding of orientation using BOLD responses. Moving forward, our research will aim to measure BOLD signals while subjects view tilt-illusion stimuli, in order to quantify changes in orientation encoding in different contexts based on voxel data. Our hope is that these results will align with the conclusion based on the psychophysical estimation data.

The results we presented here are complementary to those in Chapter 2. We developed an alternative method to extract behavioral prior from psychophysical estimation data, and obtained neural prior using different measurement techniques at a coarser spatial resolution but across a larger neural population. In both cases, we have found precise correspondences between the behavioral and neural prior. Thus, our results demonstrate the generality of our framework and constitute a strong quantitative validation of the efficient coding theory.

CHAPTER 4

IMAGE RECONSTRUCTION FRAMEWORK FOR CHARACTERIZING INITIAL VISUAL ENCODING

This chapter was previously published as: Ling-Qi Zhang, Nicolas P Cottaris, and David H Brainard. An image reconstruction framework for characterizing initial visual encoding. *eLife*, 11:e71132, 2022a. I contributed to the conceptualization, formal analysis, methodology, validation, software, visualization, and writing of this work.

Abstract

We developed an image-computable observer model of the initial visual encoding that operates on natural image input, based on the framework of Bayesian image reconstruction from the excitations of the retinal cone mosaic. Our model extends previous work on ideal observer analysis and evaluation of performance beyond psychophysical discrimination, takes into account the statistical regularities of the visual environment, and provides a unifying framework for answering a wide range of questions regarding the visual front end. Using the error in the reconstructions as a metric, we analyzed variations of the number of different photoreceptor types on human retina as an optimal design problem. In addition, the reconstructions allow both visualization and quantification of information loss due to physiological optics and cone mosaic sampling, and how these vary with eccentricity. Furthermore, in simulations of color deficiencies and interferometric experiments, we found that the reconstructed images provide a reasonable proxy for modeling subjects' percepts. Lastly, we used the reconstruction-based observer for the analysis of psychophysical threshold, and found notable interactions between spatial frequency and chromatic direction in the resulting spatial contrast sensitivity function. Our method is widely applicable to experiments and applications in which the initial visual encoding plays an important role.

4.1. Introduction

Visual perception begins at the retina, which takes sensory measurements of the light incident at the eyes. This initial representation is then transformed by computations that support perceptual inferences about the external world. Even these earliest sensory measurements, however, do not preserve all of the information available in the light signal. Factors such as optical aberrations, spatial and spectral sampling by the cone mosaic, and noise in the cone excitations all limit the information available downstream.

One approach to understanding the implications of such information loss is ideal observer analysis, which evaluates the optimal performance on psychophysical discrimination tasks. This allows for quantification of the limits imposed by features of the initial visual encoding, as well as predictions of the effect of variation in these features (Geisler, 1989, 2011). Ideal observer analysis separates effects due to the visual representation from inefficiencies in the processes that mediate the discrimination decisions themselves. Such analyses have often been applied to analyze performance for simple artificial stimuli, assuming that the stimuli to be discriminated are known exactly (Banks et al., 1987; Davila and Geisler, 1991) or known statistically with some uncertainty (Pelli, 1985; Geisler, 2018). The ideal observer approach has been extended to consider decision processes that learn aspects of the stimuli being discriminated, rather than being provided with these a priori, and extended to handle discrimination and estimation tasks with naturalistic stimuli (Burge and Geisler, 2011, 2014; Singh et al., 2018; Chin and Burge, 2020; Kim and Burge, 2020). For a recent review see Burge (2020); also see Tjan and Legge (1998); Cottaris et al. (2019, 2020).

It is generally accepted that the visual system has internalized the statistical regularities of natural scenes, so as to take advantage of these regularities for making perceptual inferences (Attneave, 1954; Field, 1987; Shepard, 1987; Kersten and Yuille, 1996). This motivates interest in extending ideal observer analysis to apply to fully naturalistic input, while incorporating the statistical regularities of natural scenes (Burge, 2020). Here we pursue an approach to this goal that, in addition, extends the evaluation of performance to a diverse set of objectives.

We developed a method that, under certain assumptions, optimally reconstructs images from noisy cone excitations, with the excitations generated from an accurate image-computable model of the front end of the visual system (Cottaris et al., 2019, 2020). We use the term “image-computable" here in contrast with observer models that operate on abstract and/or hypothetical internal representations. The image reconstruction approach provides us with a unified framework for characterizing the information loss due to various factors in the initial encoding. In the next sections, we show analyses that: 1) use image reconstruction error as an information metric to understand the retinal mosaic "design" problem, with one example examining the implications of different allocations of retinal cone types; 2) allow both visualization and quantification of information loss due to physiological optics and cone mosaic sampling and how this varies with eccentricity, as well as with different types of color deficiency; 3) combine the image reconstruction approach with analysis of psychophysical discrimination, providing a way to incorporate into such analyses the assumption that our visual system takes into account the statistical regularities of natural images.

4.2. Methods

The problem of reconstructing images from neural signals can be considered in the general framework of estimating a signal x, given an (often lower-dimensional and noisy) measurement m. We take a Bayesian approach. Specifically, we model the generative process of measurement as the conditional probability $p(m|x)$ and the prior distribution of the signal as the probability density $p(x)$. We then take the estimate of the signal, $\hat{x}$, as the maximum a posteriori estimate $argmax\ p(m|\hat{x})p(\hat{x})$. We next explain in detail how each part of the Bayesian estimate is constructed.

4.2.1. Likelihood function

In our particular problem, x is a column vector containing the (vectorized) RGB pixel values of an input image of dimension $N * N * 3$, where N is the linear pixel size of the display. Below we will generalize from RGB images to hyperspectral images. The column vector m contains the excitations of the M cone photoreceptors. The relationship between x and m is modeled by the ISETBio software (Cottaris et al. 2019, 2020; Fig. 4.1). ISETBio simulates in detail the process of displaying an image on the monitor, the wavelength-dependent optical blur of the human eye and spectral transmission through the lens and the macular pigment, as well as the interleaved sampling of the retinal image by the L, M and S cone mosaic. For the majority of simulations presented in our paper, we simulate a 1-deg foveal retina mosaic, which contains approximately 11,000 cone photoreceptors. A stochastic procedure was used to generate approximately hexagonal mosaics with eccentricity-varying cone density matched to that of the human retina (Curcio et al., 1990)). See Cottaris et al. (2019) for a detailed description of the algorithm. We use a wavelength-dependent point spread function empirically measured in human subjects (Marimont and Wandell, 1994; Cottaris et al., 2019), with a pupil size of 3-mm. We took the cone integration time to be 50 ms. The input images of size $128 * 128 * 3$ were displayed on a simulated typical CRT monitor (simulated with a 12 bit-depth in each of the RGB channels to avoid quantization artifacts).

Once the RGB pixel values in the original image are linearized, all the processes involved in the relation between x and m, including image formation by the optics of the eye and the relation between retinal irradiance and cone excitations, are well described as linear operations. Furthermore, the instance-to-instance variability in cone excitations is described by a Poisson process acting independently in each cone. Thus $p(m|x)$ is the product of Poisson probability mass functions, one for each cone, with the Poisson mean parameter λ_i for each cone determined by a linear transformation of the input image x. We describe the linear transformation between x and the vector of Poisson mean parameters λ by a matrix R, and thus obtain:

$$p(m|x) = \prod_{i=1}^{M} Poisson(m_i|\lambda_i = [Rx]_i). \tag{4.1}$$

We refer to the matrix R as the *render matrix*. This matrix together with the Poisson variability encapsulates the properties of the initial visual encoding through to the level of the cone excitations. In cases where we parameterize properties of the initial visual encoding (parameters denoted by θ in the text above), the render matrix is a function of these parameters.

Although ISETBio can compute the relation between the linearized RGB image values at each pixel and the mean excitation of each cone, it does so in a general way that does not exploit the linearity of the relation. To speed the computations, we use ISETBio to precompute R. Each column of R is a vector of mean cone excitations r_j to a basis image x_j with one entry set to one and the remaining entry set to zero. To determine R, we use ISETBio to compute explicitly each of its columns r_j. We verified that calculating mean cone excitations from an image via Rx yields the same result as applying the ISETBio pipeline directly to the image.

See *Code and Data Availability* for parameters used in the simulation including display specifications (i.e., RGB channel spectra, display gamma function) and cone mosaic setup (i.e., cone spectral sensitivities, lens pigment and macular pigment density and absorption spectra), as well as some of the pre-computed render matrices.

4.2.2. Null space of render matrix

To understand the information lost between an original RGB image and the mean cone excitations, we can take advantage of the linearity property of the render matrix. Variations in the image space that are within the null space of the (low-rank) render matrix R will have no effect on the likelihood. That is, the cone excitation pattern provides no information to disambiguate between image variants that differ only by vectors within the null space of R. To obtain the null space of R, we used MATLAB function *null*, which computes the singular value decomposition of R. The set of right singular vectors whose associated singular values are 0 form a basis for the null space.

As an illustration, we generated random samples of images from the null space by taking linear combinations of its orthonormal basis vectors, where the weights are sampled independently from a Gaussian distribution with a mean of 0 and a standard deviation of 0.3. As shown in Fig. 4.3D, altering an image by adding to it samples from the null space has no effect on the likelihood.

4.2.3. Prior distribution

We also need to specify a prior distribution $p(x)$. The problem of developing statistical models of natural images has been studied extensively using numerous approaches, and remains challenging (Simoncelli, 2005). The high-dimensionality and complex structure of natural images makes it difficult to determine a high-dimensional joint distribution that properly captures the various forms of correlation and higher-order dependencies of natural images. Here, we have implemented two relatively simple forms of $p(x)$.

We first introduce a simple Gaussian prior $p\left(x\right)$ to set up the basic concepts and notations for image prior based on basis vectors. In particular, for the Gaussian prior, we assume $p\left(x\right) = \mathcal{N}\left(x|\mu, \Sigma\right)$. For convenience, we zero-centered our images when building priors, making $\mu = 0$. The actual mean value of each pixel is added back to each image when computing the likelihood and at the end of the reconstruction procedure. The covariance matrix Σ can be estimated empirically, from a large dataset of natural images. Note that we can write the covariance matrix as its eigen-decomposition: $\Sigma = Q\Lambda Q^{-1}$. Defining $\beta = \Lambda^{-1/2}Q^{-1}x$, we have:

$$p(\beta) = \mathcal{N}(\beta|0, I) \tag{4.2}$$

This derivation provides a convenient way of expressing our image prior: We can project images onto an appropriate set of basis vectors, and impose a prior distribution on the projected coefficients. In the case above, if we choose the basis vectors as the column vectors of $\Lambda^{-1/2}Q^{-1}$, we obtain an image prior by assuming that the entries of β are each independently distributed as a univariant standard Gaussian (Simoncelli, 2005). Such a Gaussian prior can describe the first and second order statistics of natural images, but fails to capture important higher-order structure (Portilla et al., 2003).

Our second model of $p\left(x\right)$ emerges from the basis set formulation. Rather than choosing the basis vectors from the eigen-decomposition as above and using a Gaussian distribution over the weights β, we instead choose an over-complete set of basis vectors using independent components analysis, and model the distribution of the entries of weight vector β using the long-tailed distribution Laplace distribution. This leads to a sparse coding model of natural images (Olshausen and Field, 1996) (Simoncelli and Olshausen, 2001). More specifically, we learned a set of K $\left(K \geq 3N^2\right)$ basis vectors that lead to a sparse representation of our image dataset, through the reconstruction independent component analysis (RICA) algorithm (Le et al., 2011) applied to whitened images, and took these as the columns of the basis matrix E. Our image prior in this case can be written as $p\left(\beta\right)$, with $\beta = E^{+}x$. Here E^{+} represents the pseudoinverse of matrix E, and

$$p\left(\beta\right) = \prod_{k=1}^{K} \frac{1}{2b} \exp\left(-\left|\beta_k\right|/b\right). \tag{4.3}$$

Note that we further scaled each column of E to equalize the variance across β_k's.

Both methods outlined above can be applied directly to small image patches. They are computationally intractable for larger images, however, since the calculation of basis vectors will involve either an eigen-decomposition of a large covariance matrix or independent component analysis of a set of high-dimensional image vectors. To address this limitation, we iteratively apply the prior distributions we have constructed above to overlapping small patches of the same size within a large image (Guleryuz, 2006).

To illustrate the idea, consider the following example: Assume we have constructed a prior distribution $p(y)$, for small image patches y of size $N_{patch} * N_{patch}$. To model a larger image x of size $pN_{patch} * pN_{patch}$, we could consider viewing x as composed of $p * p$ independent patches of non-overlapping $y_j's$. Under this assumption, the prior on x could be expressed as the product:

$$p(x) \propto \prod_{j=1}^{p*p} p(y_j), \tag{4.4}$$

where $y_j's$ describe individual patches of size $N_{patch} * N_{patch}$ within x. The independence assumption is problematic, however, since $y_j's$ are far from independently sampled natural images: they need to be combined into a single coherent large image. Using this approach to approximate a prior would create block artifacts at the patch boundaries.

The basic idea above, however, can be extended heuristically to solve the block artifact problem by allowing y_j's to overlap with each other. The degree of overlap can be viewed as an additional parameter of the prior, which we refer to here as the stride. This effectively implements a convolutional form of the sparse coding prior Gu et al. (2015)). Again, for example, consider a large image x of size $pN_{patch} * pN_{patch}$. A stride of 1 will tile through all $(pN_{patch} - N_{patch} + 1) * (pN_{patch} - N_{patch} + 1)$ possible patches of size $N_{patch} * N_{patch}$ within x, yielding a prior distribution of the form:

$$p(x) \propto \prod_{j=1}^{(pN_{patch} - N_{patch} + 1)^2} p(y_j) \tag{4.5}$$

Although this form of prior is still an approximation, we have found it to work well in practice, and using it does not lead to visible block artifacts as long as the stride parameter is sufficiently smaller than N_{patch}.

4.2.4. Maximum a Posteriori estimation

To reconstruct the image $\hat{x}$ given a pattern of cone excitation m, we find the maximum a posteriori estimate:$\hat{x} = argmax\ p\left(m|\hat{x}\right)p\left(\hat{x}\right)$. In practice, this optimization is usually expressed in terms of its logarithmic counterpart: $\hat{x} = argmax\ \ [logp\left(m|\hat{x}\right) + logp\left(\hat{x}\right)]$.

For the Poisson likelihood and sparse coding prior, the equation above becomes:

$$\hat{x} = argmax\ \ \left[\sum_{i=1}^{M}\left(-\lambda_i + m_i * log\left(\lambda_i\right)\right) + \gamma * \sum_{j=1}^{J}\sum_{k=1}^{K}|\beta_{jk}| + c\right], \tag{4.6}$$

where $\lambda = R\hat{x}$, $\beta_j = E^{+}y_j$, $y_j's$ are individual patches of size $N_{patch} * N_{patch}$ within x. Each β_j is of length K and there are a total of J (overlapping) patches. Lastly, c is a constant that does not depend on $\hat{x}$.

In principle, the value of γ can be analytically derived based on the parametric form of the prior. However, due to the approximate nature of our prior, introduced especially by the aggregation over patches, we left γ as a free parameter. Treating γ as a free parameter also provides some level of robustness against misspecification of the prior more generally. For most of the reconstruction results presented in this paper, the value of γ was determined by maximizing reconstruction performance with a cross-validation procedure (see Fig. 4.2). We also found that the optimal γ values were similar across the two loss functions we considered. Note that the additional flexibility provided by this γ parameter also provides us with a parametric way to manipulate and isolate the relative contribution of the log-likelihood and log-prior terms to the reconstruction (e.g., Fig. 4.2; also compare Fig. 4.7 and Fig. B.4).

The optimization problem required to obtain $\hat{x}$ can be solved efficiently using the MATLAB function *fmincon* by providing the analytical gradient to the minimization function:

$$\frac{\partial \log p\,(m|x)}{\partial x} = \left(-1 + m \circ \frac{1}{\lambda}\right)^T * R, \quad \frac{\partial \log p\,(y)}{\partial y} = sign\,(\beta)^T * E^+, \tag{4.7}$$

where $\lambda = Rx, \beta = E^+y$, $\circ$ denotes element-wise product between two vectors,$\frac{1}{\lambda}$ is the element-wise inverse of vector λ, and

$$sign\,(\beta_i) = -1 \text{ for } \beta_i < 0 \text{ and } 1 \text{ for } \beta_i > 0 \tag{4.8}$$

4.2.5. RGB image dataset

We used the ImageNet ILSVRC (Russakovsky et al., 2015) as our dataset for natural RGB images. Fifty randomly sampled images were reserved as the evaluation set, and the rest of the images were used for learning the prior and for cross-validation. For the sparse prior, we constructed a basis set size of $K = 768$, on image patches of size $16 * 16$ sampled from the training set, and used a stride of 4 when tiling larger images. We randomly sampled 20 patches from each one of the $5,000$ images in the training set for learning the prior (ICA analysis), and 500 images for the cross-validation procedure to determine the γ parameter.

In our work, we simulate display of the RGB images on idealized monitor to generate spectral radiance as a linear combination of the monitor's RGB channel spectra. Thus, a prior over the linear RGB pixels values induces a full spatial-spectral prior. To make sure the constraints introduced by RGB images together with the monitor do not influence our results, we also conducted a control analysis using hyperspectral images directly, as described in the following section.

4.2.6. Hyperspectral images

As a control analysis, we developed priors and reconstructed images directly on small patches of hyperspectral images. The development is essentially the same as above, with the generalization being to increase the number of channels in the images from 3 to N. In addition, since our algorithm treats images as high-dimensional vectors, it can be directly applied to reconstruct hyperspectral images. Here, we used images from Nascimento et al. (2002) and Chakrabarti and Zickler (2011). The dataset of Nascimento et al. (2002) was pre-processed following the instructions provided by the authors, and the images of Chakrabarti and Zickler (2011) were converted to spectral radiance using the hyperspectral camera calibration data provided in that work. We further resampled the combined image dataset with a patch size of $18 * 18$ and 15 uniformly spaced wavelengths between 420 nm and 700 nm for a dataset of $\sim$ 5000 patches. We retained 300 of them as the evaluation set, and the rest for prior learning and cross-validation. The remaining of the analysis (i.e., prior and reconstruction algorithm) followed the same procedures as those used for the RGB images, using number of basis functions $K = 4860$ and applied directly to each small image without the patchwise procedure.

See *Code and Data Availability* for the curated RGB and hyperspectral image dataset, as well learned basis functions for each sparse prior.

4.2.7. Gaussian prior for synthetic images

We also reconstructed multivariate Gaussian distributed synthetic images with known chromatic and spatial correlations that we can explicitly manipulate (Fig. 4.5). To construct these signals $x \sim \mathcal{N}(\mu, \Sigma)$, where x is RGB image of size $N * N * 3$ ($N = 36$ in our current analysis), we set $\mu = 0.5$, and used a separable Σ along its two spatial dimensions and one chromatic dimension. That is:

$$\Sigma = \Sigma_c \otimes \Sigma_s \otimes \Sigma_s, \tag{4.9}$$

where Σ_c is the chromatic covariance matrix of size $(3 * 3)$

$$\Sigma_c(i, j) = \sigma_c^2 * \rho_c^{|i-j|}, \tag{4.10}$$

and Σ_s is the spatial covariance matrix of size$(N * N)$

$$\Sigma_s(i, j) = \sigma_s^2 * \rho_s^{|i-j|}. \tag{4.11}$$

In the covariance matrix constructions, i, j index into entries of Σ_c and Σ_s at i-th row and j-th column. Here $\otimes$ represents the Kronecker product, thus producing the signal covariance matrix Σ of size $\left(3N^2 * 3N^2\right)$ (Brainard et al., 2008; Manning and Brainard, 2009).

The parameters σ_c^2 and σ_s^2 determine the overall variance of the signal, which are fixed across all simulations, whereas by changing the value of ρ_c and ρ_s, we manipulate the degree of spatial and chromatic correlation presented in the synthetic images (Fig. 4.5).

We introduce an additional simplification for the case of reconstructions with respect to the synthetic Gaussian prior: We approximated the Poisson likelihood function with a Gaussian distribution with fixed variance. Thus, the reconstruction problem can be written as:

$$p(\beta) = \mathcal{N}(\beta | 0, I), \tag{4.12}$$

$$p\left(m|\beta\right) = \mathcal{N}\left(m; RQ\Lambda^{1/2}\beta, \sigma^2 I\right), \tag{4.13}$$

where R is the render matrix, and $\Sigma = Q\Lambda Q^{-1}$.

The reconstruction problem with Gaussian prior and Gaussian noise matches the ridge regression formulation, and can be solved analytically by the regularized normal equations, applied directly to each small image without the patchwise procedure. Denote the design matrix $D = RQ\Lambda^{1/2}$:

$$\hat{\beta} = \left(D^T D + \gamma I\right)^{-1} D^T m, \tag{4.14}$$

$$\hat{x} = Q\Lambda^{1/2}\hat{\beta}. \tag{4.15}$$

Note that the γ parameter here is also determined through a cross-validation routine. We adopted this simplification (using Gaussian noise) for the simulation results in Fig. 4.5, in order to make it computationally feasible to evaluate the average reconstruction error across a large number of synthetic image datasets.

4.2.8. Variations in retinal cone mosaic

To simulate a dichromatic observer, we constructed retinal mosaics with only two classes of cones but with similar spatial configuration. To simulate the deuteranomalous observer, we shifted the M cone spectral sensitivity function, setting its peak at 550 nm instead of the typical 530 nm. In both cases, the likelihood function (i.e., render matrix R) was computed using the procedure described above and the same Bayesian algorithm was applied to obtain the reconstructed images.

In Fig. 4.6 we also present the results of two comparison methods for visualizing dichromacy, those of Brettel et al. (1997) and Jiang et al. (2016), both are implemented as part of ISETBio routine. To determine the corresponding dichromatic images, we first computed the LMS trichromatic stimulus coordinates of the linear RGB value of each pixel of the input image, based on the parameters of the simulated CRT display. LMS coordinates were computed with respect to the Stockman-Sharpe 2-deg cone fundamentals (Stockman and Sharpe, 2000). The ISETBio function *lms2lmsDichromat* was then used to transform these LMS coordinates according to the two methods (see a brief description in the main text). Lastly, the transformed LMS coordinates were converted back to linear RGB values, and gamma corrected before rendering.

To simulate retinal mosaics at different eccentricities, we constructed retinal mosaics with the appropriate photoreceptor size, density (Curcio et al., 1990), and physiological optics (Polans et al., 2015), and computed their corresponding render matrices. The same Bayesian algorithm was applied to obtain the reconstructed images.

To simulate the interferometric experimental conditions of Williams (1985), we used diffraction-limited optics without longitudinal chromatic aberration (LCA) for the computation of the cone excitations, but used the likelihood function with normal optics for the reconstruction. This models subjects whose perceptual systems are matched to their normal optics and assumes there is no substantial adaptation within the short time span of the experiment.

4.2.9. Contrast sensitivity function

We compared the spatial Contrast Sensitivity Function (CSF) between a standard, Poisson 2AFC ideal observer, and an image reconstruction-based observer. We simulated stimulus modulations in two chromatic contrast directions, L+M and L-M. Contrast was measured as the vector length in the L and M cone contrast plane at 5 spatial frequencies, $[2, 4, 8, 16, 32]$ cycles per degree. For each chromatic direction and spatial frequency combination, the sensitivity is defined as the inverse of threshold contrast.

We used the QUEST+ procedure (Watson, 2017) as implemented in MATLAB by Brainard (https://github.com/BrainardLab/mQUESTPlus) for estimating the simulated threshold efficiently as follows: We initialized the procedure with the contrast near the middle of a pre-defined possible stimulus range. For each contrast, we first generated a null template T_{null}, which is the noise-free, average excitations of a 0.5 deg foveal mosaic with N_{cones} cones to a uniform background stimulus; and a target template T_{targ}, which is the noise-free, average cone excitations to a grating stimulus at that contrast level. We then simulated 128 two alternative forced choice (TAFC) trials at this contrast. For each trial, two Poisson-noise corrupted observed sets of cone excitations r_{null} and r_{targ}, are generated based on T_{null} and T_{targ}, respectively. We determine the accuracy of for TAFC trials with the target in the first interval. Based on the observer responses, the QUEST+ procedure chooses the next test contrast according to an information-maximization criterion (Watson, 2017)). The process is repeated 15 times, for a total of $15 \times 128 = 1920$ trials.

For the Poisson TAFC observer, we directly compute the likelihood ratio for the two possible orderings of the null and target stimulus:

$$\Lambda = \frac{Poisson\,(r_{\text{targ}}|T_{\text{targ}})\,Poisson\,(r_{\text{null}}|T_{\text{null}})}{Poisson\,(r_{\text{targ}}|T_{\text{null}})\,Poisson\,(r_{\text{null}}|T_{\text{targ}})}. \tag{4.16}$$

Taking the logarithm of the equation above, the decision rule simplifies to the following:

$$d = \sum_{i=1}^{N_{\text{cones}}} \{(r_{\text{targ}} \circ \log T_{\text{targ}} + r_{\text{null}} \circ \log T_{\text{null}}) - (r_{\text{null}} \circ \log T_{\text{targ}} + r_{\text{targ}} \circ \log T_{\text{null}})\}_i \tag{4.17}$$

where $\circ$ denotes element-wise product between two vectors. The simulated observer correctly chooses target in first interval when $d > 0$, and incorrectly test in second when $d < 0$. Because of symmetry, we only need to simulated one of the two TAFC orders.

For the image reconstruction-based observer, given the cone responses, it first applies the reconstruction algorithm to obtain the image template $\hat{T}_{\text{null}}$ and $\hat{T}_{\text{targ}}$ from T_{null} and T_{targ}, and also noisy image instances $\hat{r}_{\text{null}}$ and $\hat{r}_{\text{targ}}$ by applying the same algorithm to r_{null} and r_{targ}. We then perform a template-matching decision rule as follows:

$$d = \sqrt{||\hat{r}_{\text{targ}} - \hat{T}_{\text{targ}}|\,|_2^2 + ||\hat{r}_{\text{null}} - \hat{T}_{\text{null}}|\,|_2^2} - \sqrt{||\hat{r}_{\text{null}} - \hat{T}_{\text{targ}}|\,|_2^2 + ||\hat{r}_{\text{targ}} - \hat{T}_{\text{null}}|\,|_2^2}, \tag{4.18}$$

where $|| \cdot ||_2$ represents the L_2 norm of a vector. The template observer correctly chooses target in first interval when $d < 0$, and incorrectly target in second interval when $d > 0$. We choose the template matching procedure for computational convenience. Note that because the variability in the reconstructed images is not independent across pixels, this procedure is not ideal.

4.2.10. Code and data availability

The MATLAB code used for this paper is available at:
https://github.com/isetbio/ISETImagePipeline

In addition, the curated RGB and hyperspectral image datasets, parameters used in the simulation including display and cone mosaic setup, as well as the intermediate results such as the learned sparse priors, likelihood functions (i.e., render matrices), are available through: https://tinyurl.com/26r92c8y

4.3. Results

We developed a Bayesian method to reconstruct images from sensory measurements, which we describe briefly here (see *Methods*for details). We begin with a forward model that expresses the relation between an image and its visual representation at a well-defined stage in the visual pathway. Here that stage is the excitations of the photoreceptors of the retinal cone mosaic, so that our model accounts for blur in retinal image formation, spatial and spectral sampling by the cone mosaic, and the noise in the cone excitations. The approach is general, however, and may be applied to other sites in the visual pathways (see e.g., Naselaris et al. (2009); Parthasarathy et al. (2017)). Our forward model is implemented within the open-source software package ISETBio (isetbio.org; Fig. 4.1A-C) which encapsulates the probabilistic relationship between the stimulus (i.e., pixel values of a displayed RGB image) and the cone excitations (i.e., trial-by-trial photopigment isomerizations). ISETBio simulates the process of displaying an image on a monitor (Fig. 4.1A), the wavelength-dependent optical blur of the human eye and spectral transmission through the lens and the macular pigment (Fig. 4.1B), as well as the interleaved spatial and chromatic sampling of the retinal image by the L, M and S cones (Fig. 4.1C). Noise in the cone signals is characterized by a Poisson process. The forward model allows us to compute the *likelihood* function. The likelihood function represents the probability that an observed pattern of cone excitations was produced by any given image.

To obtain a *prior* over natural images, we applied independent components analysis (ICA, see *Methods*) to a large dataset of natural images (Russakovsky et al., 2015), and fit an exponential probability density function to the individual component weights (Fig. 4.1D). The prior serves as our description of the statistical structure of natural images.

Given the likelihood function, prior distribution, and an observed pattern of cone excitations, we can then obtain a reconstruction of the original image stimulus by applying Bayes rule to find the posterior probability of any image given that pattern. We take the reconstructed image as the one that maximizes the *a posteriori* probability (MAP estimate, see *Methods*) (Fig. 4.1D).

4.3.1. Basic properties of the reconstructions

To understand the consequences of initial visual encoding, we need to study the interaction between the likelihood function (i.e., our model of the initial encoding) and the statistics of natural images (i.e., the image prior). There are strong constraints on the statistical structure of natural images, such that natural images occupy only a small manifold within the space of all possible images. The properties of the initial encoding produce ambiguities with respect to what image is displayed when only the likelihood function is considered, but if these can be resolved by taking advantage of the statistical regularities of the visual environment, they should in principle, not prohibit effective visual perception. To illustrate this point, consider the simple example of discrete signal sampling: Based on the sampled signal, one cannot distinguish between the original signal from all its possible aliases (Bracewell and Bracewell, 1986)). However, with the prior knowledge that the original signal contains only frequencies below the Nyquist frequency of the sampling array, this ambiguity is resolved. In the context of our current study, the role of the natural image prior comes in several forms, as we will demonstrate in Results. First, since the reconstruction problem is underdetermined,

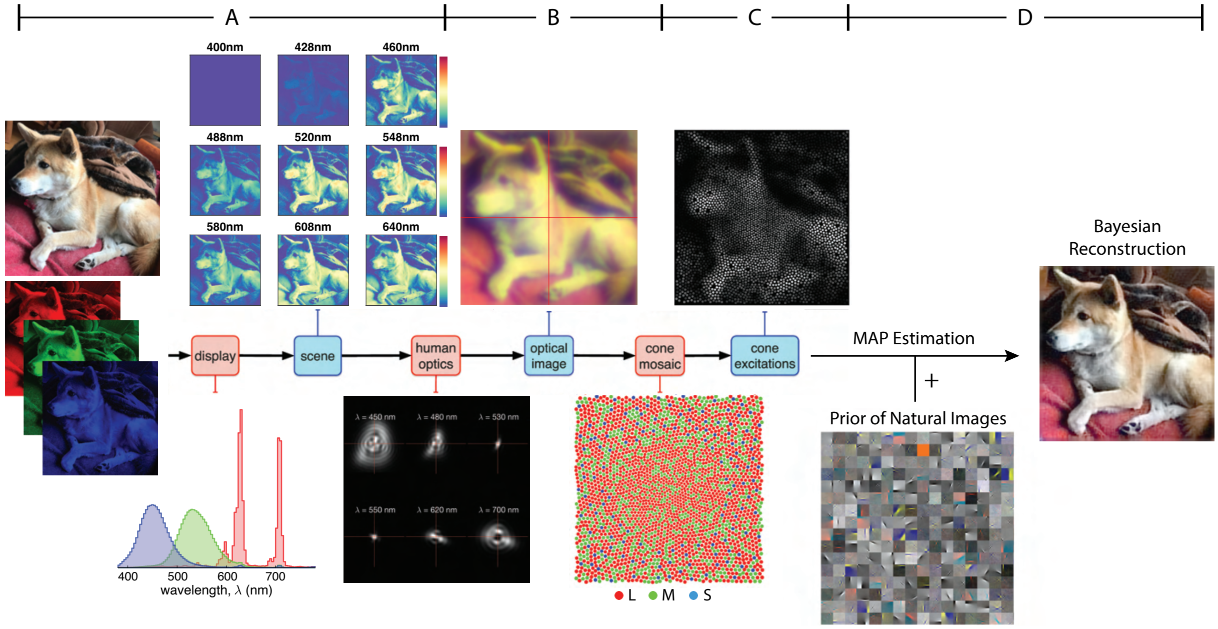

Figure 4.1: Model of the Initial Visual Encoding and Bayesian Reconstruction from Cone Mosaic Excitation. **A)** The visual stimulus, in our case a natural image in RGB format, is displayed on a simulated monitor, which generates a hyperspectral scene representation of that image. **B)** The hyperspectral image is blurred with a set of wavelength-dependent point-spread functions typical of human optics. We also account for spectral transmission through the lens and the macular pigment. This process produces the retinal image at the photoreceptor plane. **C)** The retinal image is then sampled by a realistic cone mosaic, which generates cone excitations (isomerizations) for each cone. The trial-by-trial variability in the cone excitations is modeled as a Poisson process. **D)** Our Bayesian reconstruction method takes the pattern of cone excitations as input and estimates the original stimulus (RGB image) based on the likelihood function and a statistical model (prior distribution) of natural images (See *Methods*).

the prior is a regularizer, providing a unique MAP estimate; Second, the prior acts as a denoiser, counteracting the Poisson noise in the cone excitation; Lastly, the prior guides the spatial and spectral demosaicing of the signals provided via the discrete sampling of the retinal image by the cone mosaic.

To highlight the importance of prior information while holding the likelihood function fixed, we can vary a parameter γ that adjusts the weight of the log-prior term in the reconstruction objective function (see *Methods*). Explicitly manipulating γ reveals the effect of the prior on the reconstruction (Fig. 4.2). When γ is small, the reconstruction is corrupted by the noise and the ambiguity of the initial visual encoding (Fig. 4.2A, B). When γ is large, the prior leads to desaturation and over-smoothing (Fig. 4.2E) in the reconstruction. For the rest of our simulations, the value of γ is determined on the training set by a cross-validation procedure that minimizes the reconstruction error, unless specified otherwise (Fig. 4.2C).

To further elucidate properties of the Bayesian reconstruction, especially the interaction between the likelihood and prior, we plotted a few representative images in a log-prior, log-likelihood coordinate system, given a particular instance of cone excitations (Fig. 4.3). The optimal reconstruction, taken as the MAP estimate, has both a high prior probability and likelihood value as expected (Fig. 4.3A). In fact, for our reconstruction algorithm, there should not exist any image above the $\gamma x + y = c$ line that goes through A (solid line, Fig. 4.3), otherwise the optimization routine has failed to find the global optimum. The original image stimulus (ground truth) has a slightly lower likelihood value, mainly due to noise present in the cone excitations, and also a slightly lower prior probability, possibly due to the fact that our prior is only an approximation to the true natural image distribution (Fig. 4.3B). The detrimental effect of noise becomes prominent in a maximum likelihood estimate (MLE, Fig. 4.3C): Noise in the cone excitations is interpreted as true variation in the original image stimulus, thus slightly increasing the likelihood value but also creating artifacts. Such artifacts are penalized by the prior in other reconstructions. Furthermore, even without the presence of noise, other features of the initial visual encoding (e.g., Fig. 4.1B, C) cause loss of information and ambiguity for the reconstruction. This is illustrated by a set of

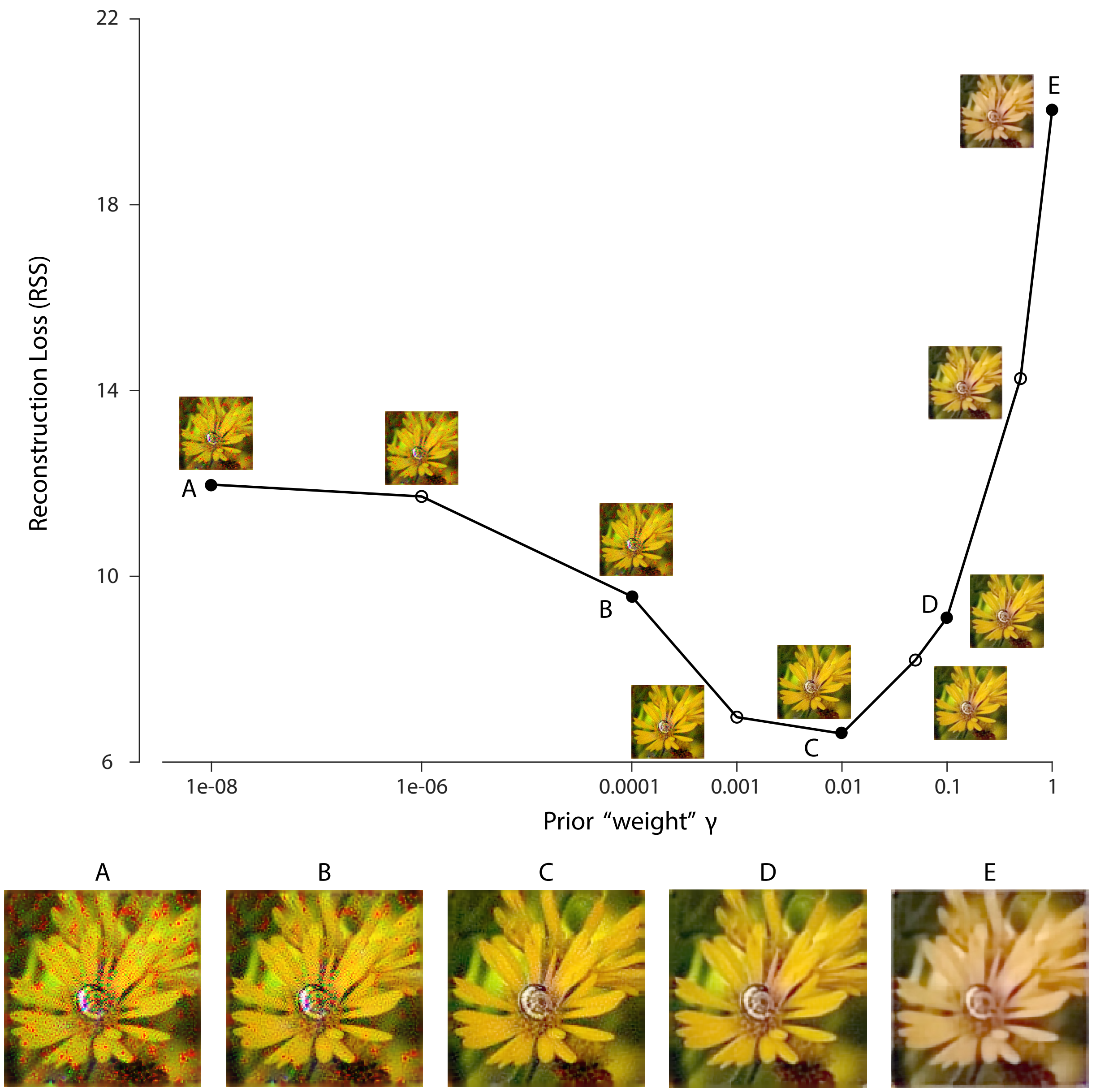

Figure 4.2: Effect of prior weight on reconstructed image. Reconstruction error for an example natural image using a 1-deg foveal mosaic and root sum of squared distance (RSS, y-axis) in the pixel space as the error metric, as a function of weight γ on the log-prior term (x-axis, see *Methods*) in the reconstruction objective function. The reconstructed image obtained with each particular γ value is shown alongside each corresponding point. Image **C** corresponds to the value of γ obtained through the cross-validation procedure (see *Methods*). The images at the bottom are magnified versions of a subset of the images for representative γ values, as indicated by the solid dots in the plot.

images that lie on the equal likelihood line with the MAP reconstruction (Fig. 4.3D): There exist an infinite set of variations in the image (stimulus) that have no effect on the value of the likelihood function. (i.e., variations within the null space of the linear likelihood render matrix, see *Methods*). Also note that one implication of the existence of the null space is that the MLE solution to the reconstruction problem is actually underdetermined, as an entire subspace of images can have the same likelihood value. In the figure we show one arbitrarily chosen MLE estimate. Thus, the cone excitations provide no information to distinguish between images that differ by such variations. However, as with the case of noise, variations inconsistent with natural images are discouraged by the prior. Other corruptions of the image, such as addition of white noise in the RGB pixel space, are countered by both the likelihood and prior (Fig. 4.3E). Lastly, for illustrative purposes, we can increase the prior probability of the reconstruction relative to the optimal by making it spatially or chromatically more uniform (Fig. 4.3F), but doing so decreases the likelihood.

4.3.2. Optimal allocation of retinal photoreceptors

Within the Bayesian reconstruction framework, the goal of the visual front end can be characterized as minimizing the average error in reconstruction across the set of natural images. In this context, we can ask how to choose various elements of the initial encoding, subject to constraints, to minimize the expected reconstruction error under the natural image prior (Levin et al., 2008; Manning and Brainard, 2009). More formally, we seek the “design" parameters θ of a visual system:

$$\theta = argmin_{\theta} \;\; E_{p(x)} \left(E_{p(m|x;\theta)} L \left(\hat{x} \left(m; \theta \right), x \right) \right), \tag{4.19}$$

where $\hat{x} \left(m; \theta \right) = argmax_{x} \;\; p \left(m|x; \theta \right) p \left(x \right)$. Here x represents individual samples of natural images, m represents instances of cone excitation (i.e., sensory measurements), and $p \left(m|x; \theta \right)$ is our model of the initial encoding (i.e., likelihood function). The particular features under consideration of the modeled visual system are indicated explicitly by the parameter vector θ. The MAP image reconstruction is indicated by $\hat{x} \left(m; \theta \right)$, and $L \left(\cdot, \cdot \right)$ is a loss function that assesses reconstruction error. In practice, the expectations are approximated by taking the average over large samples of natural images and cone excitations.

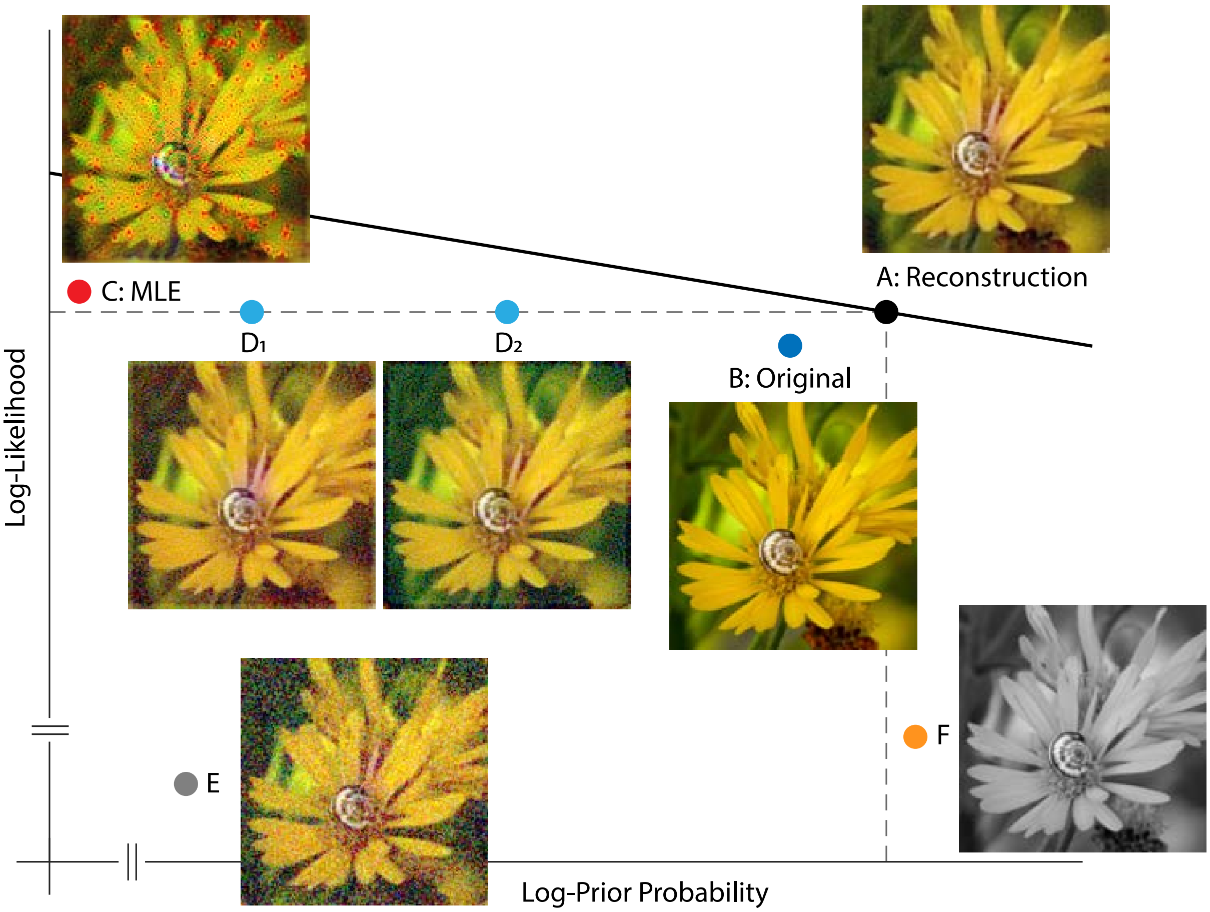

Figure 4.3: Solution space of image reconstruction. Given a particular instance of cone excitations, we can evaluate the (log-)prior probability (x-axis) and (log-)likelihood value (y-axis) for arbitrary images. Here, a few representative images are shown together with their corresponding location in a log-prior, log-likelihood coordinate system. **A)** The optimal MAP reconstruction obtained via the reconstruction algorithm. The solid line shows $\gamma x + y = c$, with the value of c evaluated at the optimal reconstruction and with the value of γ matched to that obtained through cross-validation. **B)** Original input image (ground truth). **C)** A reconstruction generated by maximum likelihood estimation (MLE, set $\gamma = 0$). Note that the maximum likelihood reconstruction shown is not unique, since adding any pattern from the null space of the likelihood matrix leads to a different reconstruction with the same maximum likelihood. Here one arbitrarily chosen MLE reconstruction is shown. **D)** Optimal reconstruction, corrupted by patterns randomly sampled from the null space of the likelihood render matrix (see *Methods*). These have the same likelihood as the optimal reconstruction, but lower prior probability. **E)** Optimal reconstruction, corrupted by white noise in RGB space. **F)** Grayscale version of the optimal reconstruction.

One intriguing design problem is the allocation of cone photoreceptor types: The maximum number of photoreceptors (cones) per unit area is bounded due to biological constraints. How should the visual system assign this limited resource across the three different types of cones? It has been observed in human subjects that there is a relatively sparse population of S cones, while large individual variability exists in the L/M cone ratio (Hofer et al., 2005). Previous research has used information-theoretical measures combined with approximations to address this question (Garrigan et al., 2010). Here, we empirically evaluated a loss function (i.e., we used root sum of squares distance in the RGB pixel space as well as the S-CIELAB space) on the reconstructed images, while systematically changing the allocation of retinal cone types (Fig. 4.4).

Interestingly, we found that large variations (nearly a 10-fold range) in the assignment of L and M cones have little impact on the average reconstruction error (Fig. 4.4A). Only when the proportion of L or M cones becomes very low is there a substantial increase in reconstruction error, as the modeled visual system approaches dichromacy. On the other hand, the average reconstruction error as a function of the proportion of S cones shows a clear optimum at a small S-cone proportion (∼10%; Fig. 4.4B).

Our results are in agreement with a previous analysis in showing that the empirically observed allocation of retinal photoreceptor type is consistent with the principle of optimal design (Garrigan et al., 2010); also see Levin et al. (2008); Manning and Brainard (2009); Tian et al. (2015). The indifference to L/M ratio can be explained by the large spatial and chromatic correlations present in natural images, together with the high overlap in L- and M-cone spectral sensitivities. This leads to a high correlation in the excitations of neighboring L and M cones in response to natural images, allowing cones of one type to be substituted for cones of the other type with little effect on reconstruction error (see the next section for additional analysis on this point). Additional analysis (Fig. B.1) revealed that the sensitivity to S cone proportion is due to a combination of two main factors: 1) chromatic aberrations, which blur the retinal image at short wavelengths and reduce the value of dense spatial sampling at these wavelengths; and 2) S cones mainly contribute to the estimation of pixel values in the B-pixel plane, whereas L and M cone contribute to both the R- and G-pixel planes (see Fig. B.1). This makes L and M cones more informative than S cones, given the particular loss functions we employ to evaluate reconstruction error. To further validate our conclusion, we have also replicated our analysis with a dataset of hyperspectral (as opposed to RGB) images (Nascimento et al., 2002; Chakrabarti and Zickler, 2011), with a loss function applied directly to the whole spectrum, and have obtained similar results (Fig. B.2, also see *Methods*).

To further study the role of statistical regularities in the optimal allocation of photoreceptor type, we repeated the L-cone proportion analysis above, but on different sets of synthetic image datasets for which the spatial and chromatic correlations in the images were manipulated explicitly (see *Methods*). The dependence of the average reconstruction error on the L-cone proportion decreases as the chromatic correlation in the signal increases (Fig. 4.5). A decrease of spatial correlation has little impact on the shape of the curves, but increases the overall magnitude of reconstruction error (Fig. 4.5; to highlight the shape, the scale of the y-axis is different across rows and columns. See Fig. B.3 for the same plot with matched y-axis scale). When both the chromatic and spatial correlation are high, there is a large margin of L-cone proportion within which the reconstruction error is close to the optimal (minimal) point (Fig. 4.5, shaded area). This analysis highlights the importance of considering visual system design in context of the statistical properties (prior distribution) of natural images, as it shows that the conclusions drawn can vary with these properties (Barlow 1961; Derrico and Buchsbaum 1991; Barlow and Földiàgk 1989; Atick, Li, and Redlich 1992; Lewis and Li 2006; Levin et al. 2008; Borghuis et al. 2008; Garrigan et al. 2010; Tkačik et al. 2010; Atick 2011; Burge 2020). Natural images are thought to have both high spatial and high chromatic correlation (Webster and Mollon 1997; Nascimento et al. 2002; Garrigan et al. 2010), making the results shown in Fig. 4.5 consistent with those in Fig. 4.4.

4.3.3. Visualization of color deficiency with image reconstruction

In addition to quantification, the reconstruction framework also provides a method for visualizing the effect of information loss in the initial visual encoding. We know that extreme values of L:M cone ratio create essentially dichromatic retinal mosaics, and from the analysis above we observed that these lead to high reconstruction error. To understand the nature of this error, we can directly

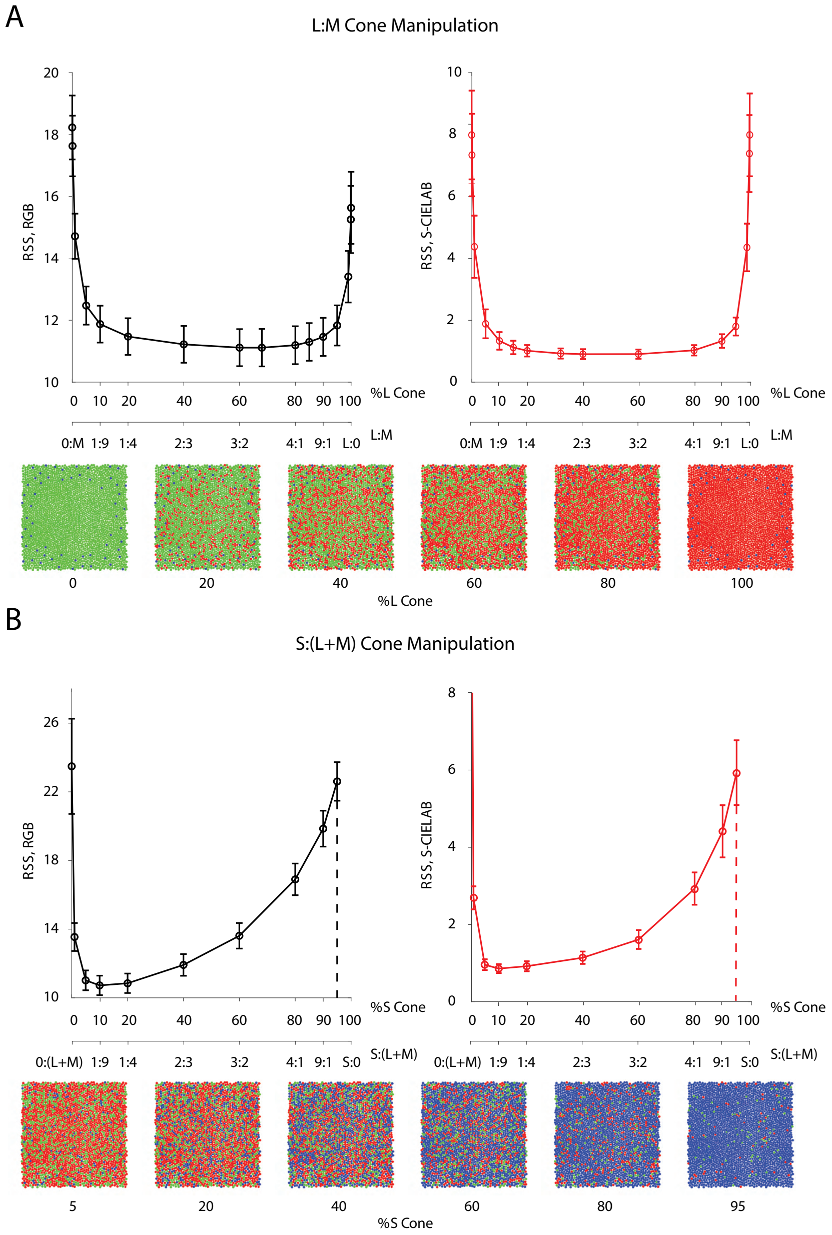

Figure 4.4: Effect of the allocation of retinal cone types on reconstruction. Average image reconstruction error from a 1-deg foveal mosaic on a set of natural images from the evaluation set, computed as root sum of squares (RSS) distance in the RGB pixel space (y-axis, left panels) and the S-CIELAB space (y-axis, right panels), as a function of different allocations of retinal photoreceptor (cone) types in the mosaic. **A)** Average (over evaluation images) reconstruction error as a function of %L cone (top x-axis), or L:M cone ratio (bottom x-axis). Example mosaics with different %L values are shown below the plot. Error bars indicate +/- 1 SEM. **B)** Average reconstruction error as a function of %S cone (top x-axis), or S:(L+M) cone ratio (bottom x-axis). Example mosaics with different %S values are shown below the plot. Error bars indicate +/- 1 SEM across sampled images. See Fig. B.2 for a replication of the same analysis with hyperspectral images.

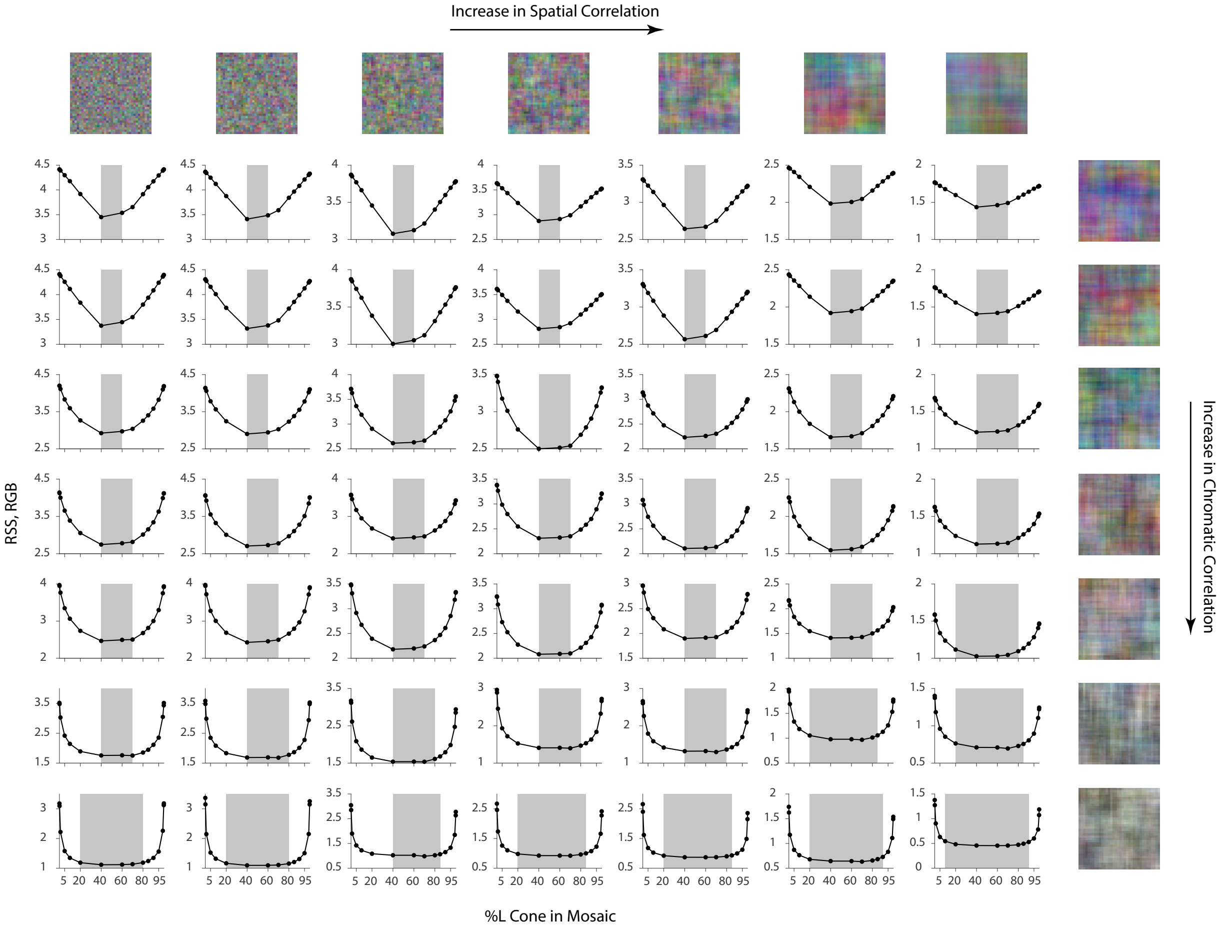

Figure 4.5: Effect of spatial and chromatic correlation on the optimal allocation of photoreceptors. Average image reconstruction error from a half-degree square foveal mosaic on different sets of synthetic images, computed as root sum of squares (RSS) distance in the RGB pixel space, as a function of %L cone (L:M cone ratio) of the mosaic (i.e., similar to Fig. 4.4A, left column). The shaded areas represent %L values that correspond to RSS values within a +0.1 RSS margin of the optimal (minimum RSS) point. Within each panel, synthetic images were sampled from a Gaussian distribution with specified spatial and chromatic correlation, as indicated by example images on the top row and rightmost column, and reconstruction was performed with the corresponding Gaussian prior (see *Methods*). The overall RSS is reduced compared to Fig. 4.4 due to the smaller image size used and the fact that the images were drawn from a different prior, as well as because the prior used in reconstruction exactly describes the images for this case. In addition, reconstruction error bars are negligible due to the large image sample size used.

visualize the reconstructed images.

Fig. 4.6A shows reconstructions of a set of example images from different dichromatic retinal mosaics. While the spatial structure of the original images is largely retained in the reconstructions, each type of dichromacy creates a distinct pattern of color confusions and shifts in the reconstructed color. Note that in the case where there is no simulated cone noise (as in Fig. 4.6), the original image has a likelihood at least as high as the reconstruction obtained via our method. Thus, the difference between the original images and each of the corresponding dichromatic reconstructions is driven by the image prior. On the other hand, the difference in the reconstructions across the three types of dichromacy illustrates how the different dichromatic likelihood functions interact with the prior.

One might speculate as to whether the reconstructions predict color appearance as experienced by dichromats. To approach this, we compare the reconstructions with two other methods that have been proposed to predict the color appearance for dichromats (Brettel et al., 1997; Jiang et al., 2016). To determine an image based on the excitations of only two classes of cones, any method will need to rely on a set of regularizing assumptions to resolve the ambiguity introduced by the dichromatic retinas. Brettel et al. (1997) started with the trichromatic cone excitations of each image pixel, and projected these onto a biplanar surface, with each plane defined by the neutral color axis and an anchoring stimulus identified through color appearance judgments made across the two eyes of unilateral dichromats. The resulting trichromatic excitations were then used to determine the rendered RGB values (Fig. 4.6B). Jiang et al. (2016) also adopted a reconstruction approach, but one that reconstructed the incident spectrum from the dichromatic cone excitations at each pixel. They then projected the estimated spectra onto trichromatic cone excitations, and used these to render the RGB values (Fig. 4.6C). In their method, a spectral smoothness constraint was introduced to regularize the spectral estimates, which favors desaturated spectra. In this sense, their prior is similar to ours: The sparse prior we used is centered on the average image, which is desaturated, and also encourages achromatic content due to the high correlations across color channels. One noticeable difference between our method and the other two is that ours takes into

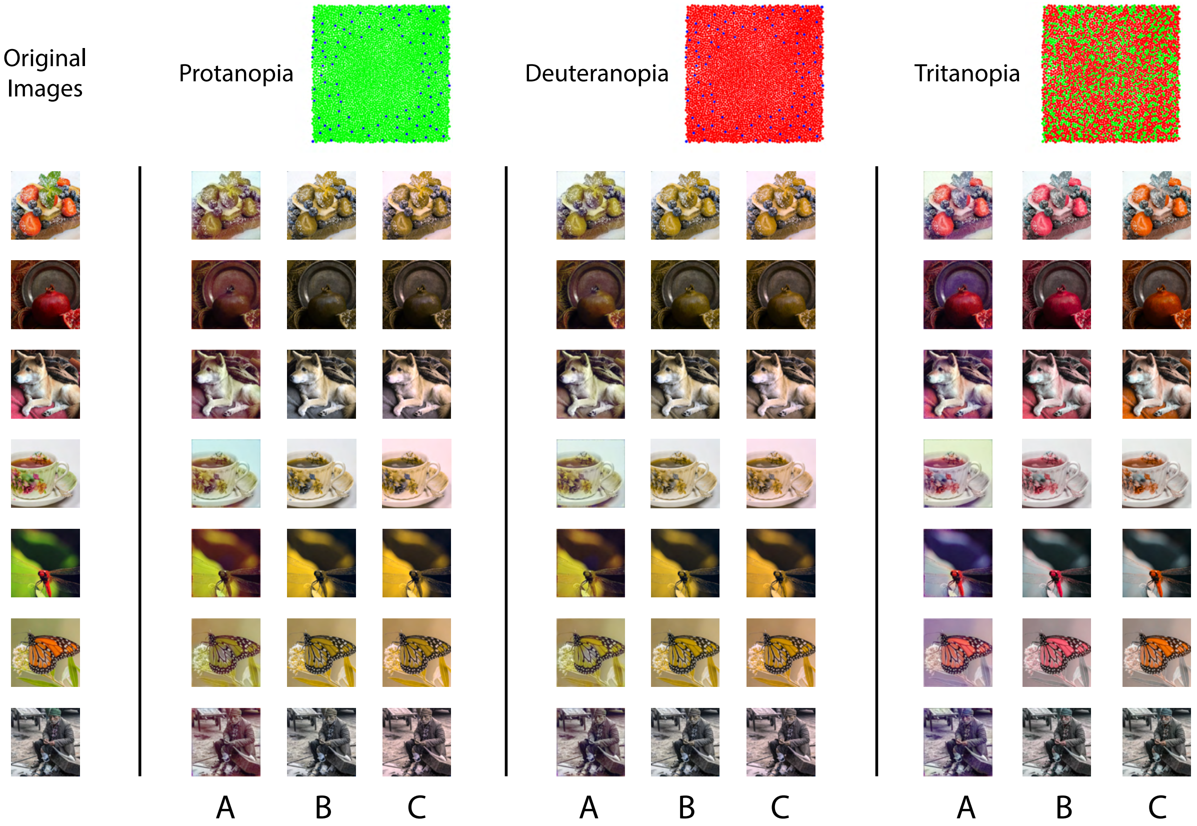

Figure 4.6: Visualization of effect of dichromacy. Reconstructions of a set of example images in the evaluation set from different types of 1-degree foveal dichromatic retinal mosaics (protanopia, deuteranopia, tritanopia) together with other previously proposed methods for predicting color appearance for dichromats. **A)** Our method; **B)** Brettel et al. (1997) **C)** Jiang et al. (2016). Cone noise was not simulated for the images shown in this figure, since the comparison methods operate directly on the input images. See *Methods* for a brief description of the implementation of the two other methods.

account the spatial structure of the image.

Interestingly, although there are differences in detail between the images obtained, in many cases the different methods produce visualizations that are quite similar. We find the general agreement between the reconstruction-based methods and the one based subject reports an encouraging sign that the reconstruction approach can be used to predict aspects of appearance.

Anomalous trichromacy is another form of color deficiency that is commonly found in human observers. For example, in deuteranomaly, the spectral sensitivity of the M cones is shifted towards that of the L cones (Fig. 4.7B). Since the three cone spectral sensitivity functions are linearly independent of each other, in the absence of noise we should be able to obtain a trichromatic reconstruction from the excitations of the deuteranomalous mosaic. However, in the presence of noise, we expect that the high degree of overlap between M and L spectral sensitivities will result in a lower signal-to-noise ratio (SNR) in the difference between M- and L-cone excitations, compared to that of a normal trichromatic observer, and thus to worse reconstructions. We performed image reconstructions for a normal trichromatic (with a peak spectral sensitivity of M cone at 530 nm) and a deuteranomalous (with a peak spectral sensitivity of M cone at 550 nm) 1-deg foveal mosaic at different overall light intensity levels (Fig. 4.7). Due to the nature of Poisson noise, the higher the light intensity, the higher the SNR of the cone excitations. At high light intensities, the reconstructions are similar for the normal and deuteranomalous mosaics (first row). At lower intensities, however, the deuteranomalous reconstruction lacks chromatic content still present in the normal reconstruction (second and third row). The increase in noise also reduces the amount of spatial detail in the reconstructed images, due to the denoising effect driven by the image prior. Furthermore, a loss of chromatic content is also seen for the reconstruction from the normal mosaic at the lowest light level (last row). This observation may be connected to the fact that biological visual systems that operate at low light levels are typically monochromatic, potentially to increase the SNR of spatial vision at the cost of completely disregarding color (e.g., the monochromatic human rod system; see Manning and Brainard (2009)for a related and more detailed treatment; also see Wald (1944); Rushton (1962); Van Hateren (1993); Land and Osorio (2003).

4.3.4. Effect of physiological optics and mosaic spatial sampling

So far, our visualizations have focused on chromatic information loss due to a reduced number of cone types or a shift in cone spectral sensitivity. However, imperfections in the physiological optics, combined with the spatial sampling of retinal mosaic, also introduces significant loss of information. Furthermore, the interleaved nature of the mosaic means that color and pattern are entangled at the very initial stage of visual processing (Brainard, 2019).To highlight these effects, we reconstructed natural images from 1-deg patches of mosaics at different retinal eccentricities across the visual field, with 1) changes in optical aberrations (Polans et al., 2015); 2) increases in size and

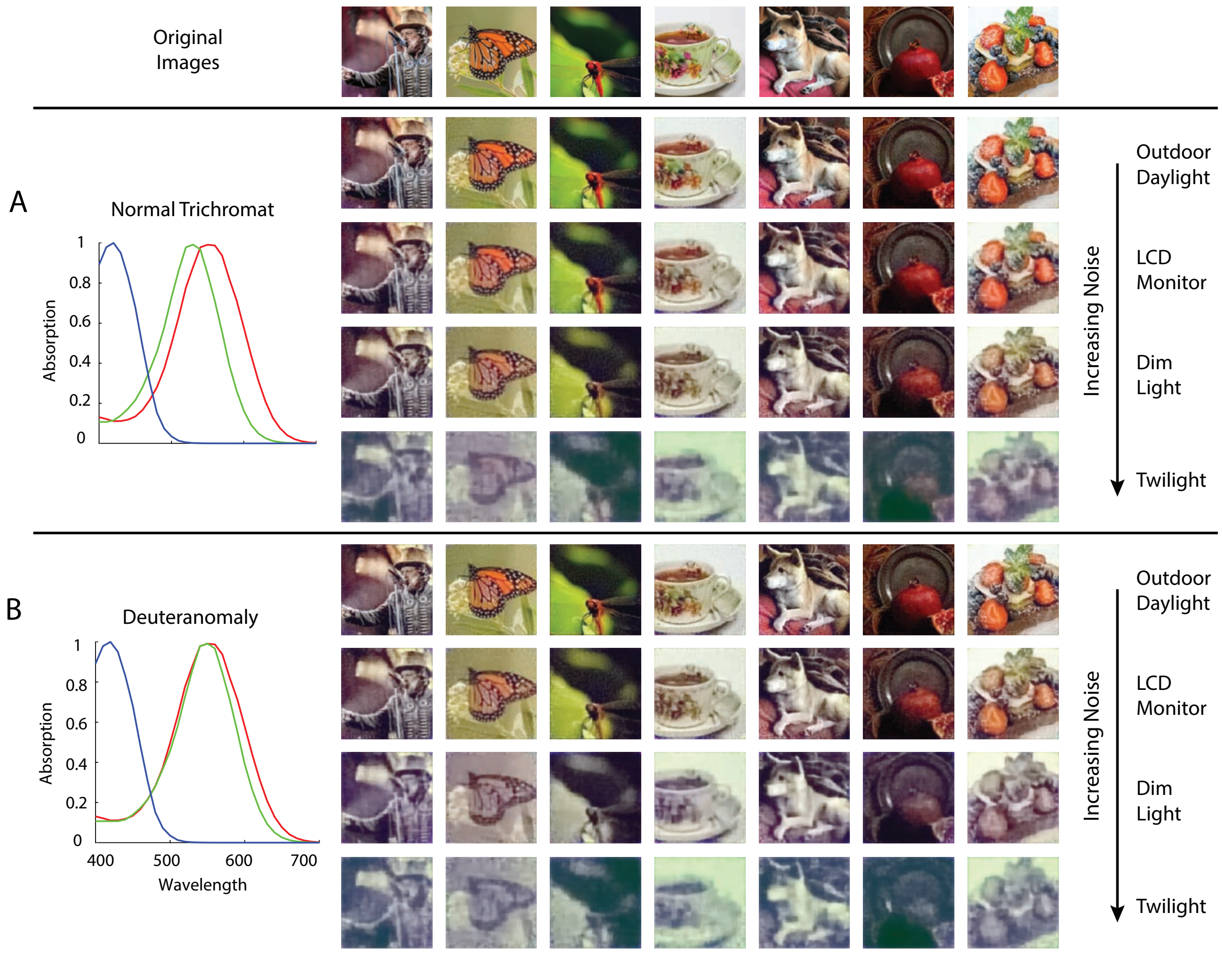

Figure 4.7: Comparison of normal and deuteranomalous observers at varying light intensities. Image reconstructions for a set of example images in the evaluation set from 1-degree, foveal **A)** normal trichromatic and **B)** deuteranomalous trichromatic mosaics at four different overall light intensity levels that lead to different Poisson signal-to-noise ratios in the cone excitations. The average excitations (photo-isomerizations) per cone per 50 ms integration time is chosen to be approximately 10^4 for *Outdoor Daylight*, 10^3 for *LCD Monitor*, 10^2 for *Dim Light*, and 10^1 for *Twilight* (Lewis and Li 2006; Stockman and Sharpe 2006). The prior weight parameter in these set of simulations was set based on a cross-validation procedure that minimizes RMSE ($\lambda = 0.05$). To highlight interaction between noise and the prior, we have also included a set of reconstructions with the prior weight set to a much lower level ($\lambda = 0.001$), see Fig. B.4.

decreases in density of the photoreceptors (Curcio et al., 1990); and 3) decreases in the density of the macular pigment (Nolan et al., 2008; Putnam and Bland, 2014). The degradation in the quality of the reconstructed images can be clearly observed as we move from the fovea to the periphery (Fig. 4.8; See Fig. B.5 for an enlarged view of the mosaic and optics). For some retinal locations, the elongated point-spread function (PSF) also introduces a salient directional blur (Fig. 4.8E, F). For a simple quantification of the average reconstruction error as a function of visual eccentricity, see Fig. B.6.

The consequences of irregular spatial sampling by the cone mosaic have been previously studied with the framework of signal processing (Snyder et al., 1977; Yellott Jr, 1983). Our results highlight that optimizing the initial visual encoding depends in rich ways on the interplay between the cone sampling and the optics. While less information (i.e., at more eccentric locations) does lead to overall lower quality reconstructions (Fig. B.6), exactly which aspects of the reconstructions are incorrect can vary in subtle ways. Concretely, in Fig. 4.8, we observe a trade-off across visual eccentricity between spatial and chromatic vision. In the image of the dragonfly, for example, the reconstructed colors are desaturated at intermediate eccentricities (e.g., Fig. 4.8C, D), compared with the fovea (Fig. 4.8A) and more eccentric locations (Fig. 4.8E, F). The desaturation is qualitatively consistent with the literature that indicates a decrease in chromatic sensitivity at peripheral visual eccentricities, at least for the red-green axis of color perception and for some stimulus spatial configurations (Virsu and Rovamo, 1979; Mullen and Kingdom, 1996); but see Hansen et al. (2009). To further elucidate this richness, in an additional analysis, we systematically varied the size of the PSF for a fixed peripheral retinal mosaic. This revealed that (Fig. B.7): 1) A larger PSF does lead to better estimate of chromatic content, albeit eventually at the cost of spatial content. 2) In general, an appropriate amount of optical blur is required to achieve the best overall image reconstruction performance, presumably due to its prevention of aliasing. We will treat the issue of spatial aliasing further in the next section.

Lastly, to emphasize the importance of the natural image prior, we performed a set of maximum likelihood reconstructions with no explicit prior constraint, which resulted in images with less co-

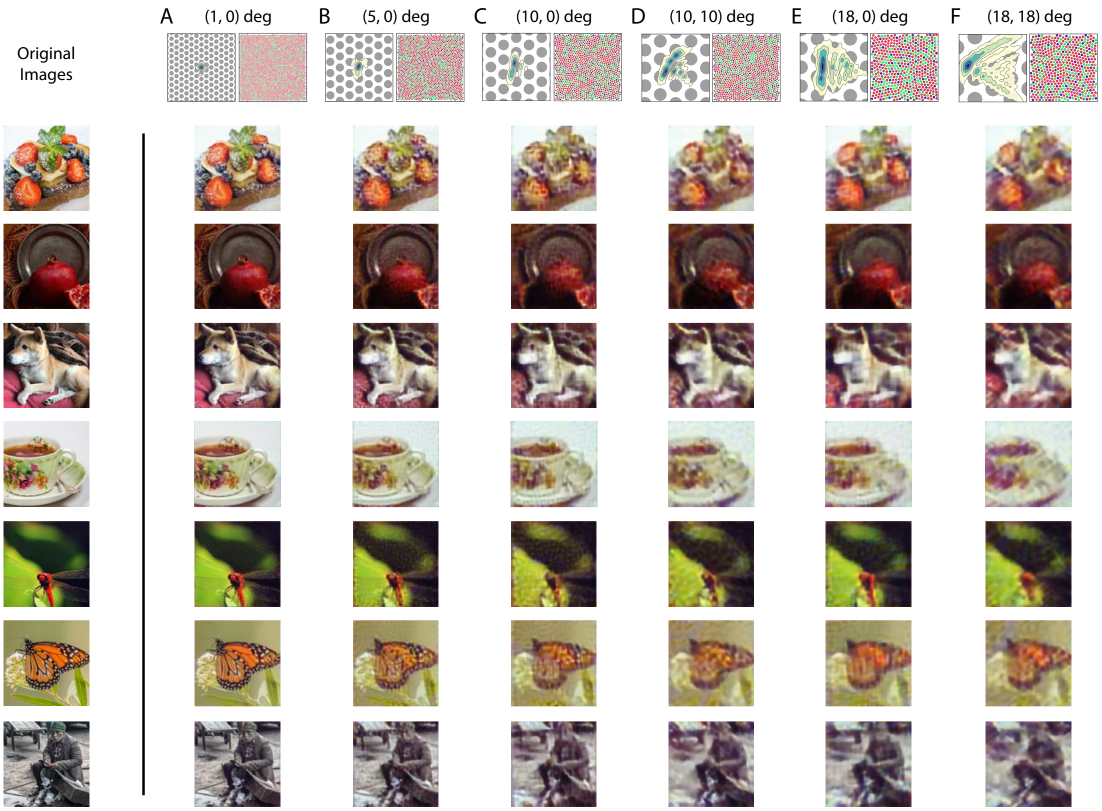

Figure 4.8: Image reconstruction with across retinal eccentricities. Image reconstructions for a set of example images in the evaluation set from 1-degree patches of mosaic at different retinal eccentricities. The coordinates at the top of each column indicate the horizontal and vertical eccentricity of the patch used for that column. The image at the top left of each column shows a contour plot of the point-spread function relative to an expanded view of the cone mosaic used for that column, while the image at the top right of each column shows the full 1-degree mosaic (see Fig. B.5 for an enlarged view of the mosaic and optics).

herent spatial structure and lower fidelity color appearance (Fig. B.8). Thus, the prior here is critical for the proper demosaicing and interpolation of the information provided by the sparse cone sampling at these peripheral locations.

4.3.5. Spatial aliasing

As we have alluded to above, the retinal mosaic and physiological optics can also interact in other important ways: Both in humans and other species, it has been noted that the optical cut-off of the eye is reasonably matched to the spacing of the photoreceptors (i.e., the mosaic Nyquist frequency), enabling good spatial resolution while minimizing spatial aliasing due to discrete sampling (Williams, 1985; Snyder et al., 1986; Land and Nilsson, 2012). In contrast to our work, these analyses did not take into account the fact that the cone mosaic interleaves multiple spectral classes of cones (but see Williams et al. (1991); Brainard (2015)), and here we revisit classic experiments on spatial aliasing for a trichromatic mosaic using our reconstruction framework.

Experimentally, it has been demonstrated that with instruments that *bypass* the physiological optics and present high contrast grating stimuli directly on the retina, human subjects can detect spatial frequencies up to 200 cyc/deg (Williams, 1985). For foveal viewing, subjects also report having a percept resembling a pattern of “two-dimensional noise" and/or “zebra stripes" when viewing those high spatial frequency stimuli (Williams, 1985). For peripheral viewing, high frequency vertical gratings can be perceived as horizontal (and vice-versa; Coletta and Williams (1987)). We explored these effects within our framework as follows: We reconstructed a set of vertical chromatic grating stimuli from the cone excitations of a foveal and a peripheral mosaic. To simulate the interferometric experimental conditions of (Williams, 1985), we used diffraction-limited optics with no longitudinal chromatic aberration (LCA), allowing high-frequency stimuli to reach the cone mosaic directly. For gratings that are above the typical optical cut-off frequency, we obtained reconstructions that 1) are quite distinct from a uniform field, which would allow them to be reliably detected in a discrimination protocol; and 2) lack the coherent vertical structure of the original stimulus (Fig. 4.9). Concretely, the reconstructions recapitulate the “zebra stripe" percept reported at approximately 120 cyc/deg in the fovea (Fig. 4.9A); as well as the orientation-reversal effect at an appropriate spatial frequency in the periphery (Fig. 4.9B). Both results corroborate previous theoretical analysis and psychophysical measurements (Williams, 1985; Coletta and Williams, 1987), but now taking the trichromatic nature of the mosaic into account. On the other hand, with full optical aberrations, the reconstructed images became mostly uniform at these high spatial frequencies (Fig. B.9). Since our method accounts for trichromacy, we have also made the prediction that for achromatic grating stimuli viewed under similar diffraction-limited conditions, while the spatial aliasing pattern will be comparable, additional chromatic aliasing should be visible (Fig. B.10; also see Williams et al. (1991); Brainard (2015)).

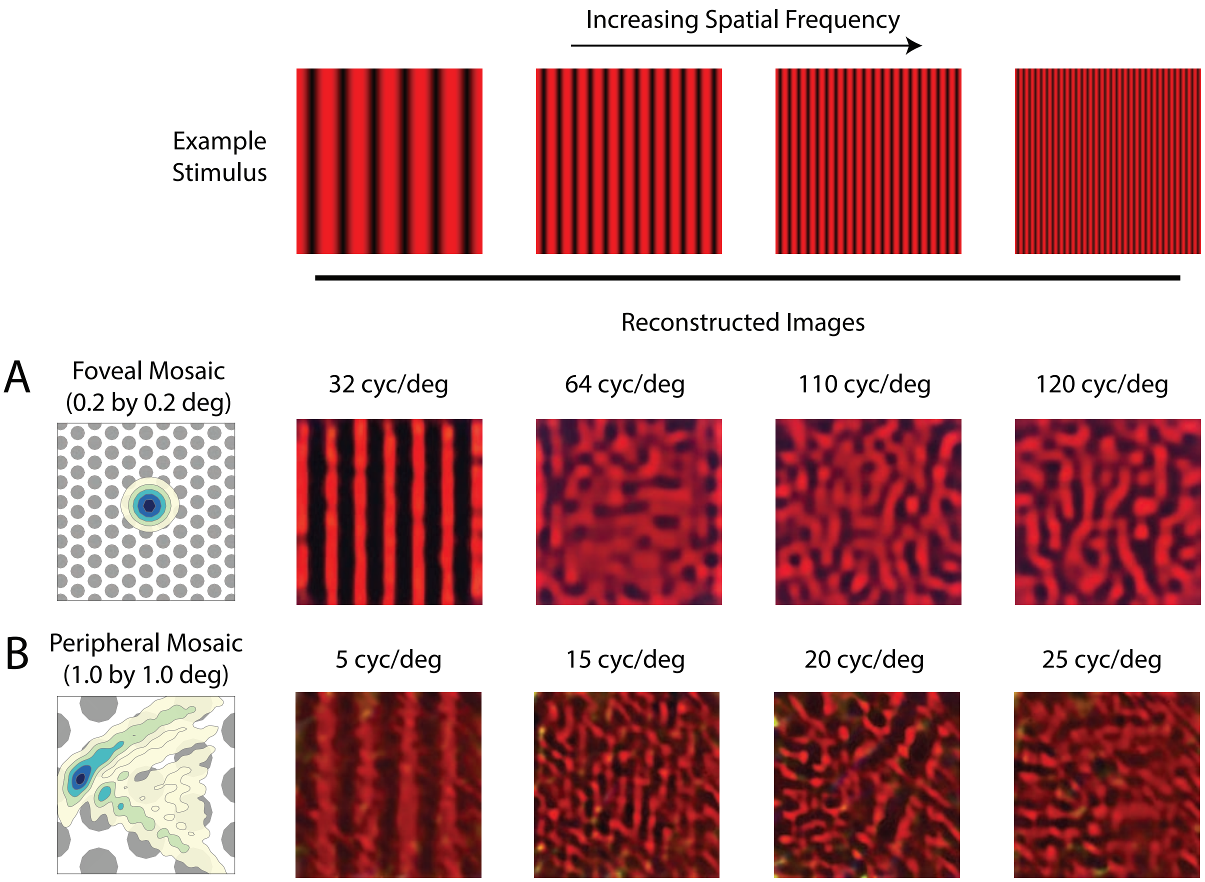

Figure 4.9: Reconstruction of chromatic grating stimuli without optical aberrations. Image reconstruction of chromatic grating stimuli with increasing spatial frequency from **A)** a 0.2-deg foveal mosaic and **B)** a 1-deg peripheral mosaic at (18, 18) degree retinal eccentricity, using diffraction-limited optics without LCA. The leftmost images show an expanded view of the cone mosaic relative to a contour plot of a typical point-spread function at that eccentricity. Images were modulations of the red channel of the simulated monitor, to mimic the 633 nm laser used in the interferometric experiments. The exact frequency of the stimuli being used for each condition is as denoted in the figure. For a more extended comparison between reconstructions with and without optical aberrations, see Fig. B.9 and Fig. B.10.

4.3.6. Contrast sensitivity function

Our framework can also be adapted to perform ideal observer analysis for psychophysical discrimination (threshold) tasks, which have been used previously to evaluate the information available in the initial encoding. Here we use the reconstructed images as the basis for discrimination decisions. This is potentially important since even the early post-receptoral visual representation (e.g., retinal ganglion cells), on which downstream decisions must be based, is likely shaped by the regularities of our visual environment (Atick et al., 1992; Borghuis et al., 2008; Karklin and Simoncelli, 2011; Atick, 1992). Our method provides a way to extend ideal observer analysis to incorporate these statistical regularities.

Concretely, we predicted and compared the diffraction-limited spatial contrast sensitivity function (CSF) for gratings with a half-degree spatial extent (see *Methods*). First, we applied the classic signal-known-exactly ideal observer to the Poisson distributed excitations of the simulated cone mosaic. We computed CSFs for both achromatic (L+M) and chromatic (L-M) grating modulations, with matched cone contrast measured as the vector length of the cone contrast vector. As expected, the ideal observer at the cone excitations produces nearly identical CSFs for the contrast-matched L+M and L-M modulations; also, as expected, these fall off with spatial frequency, primarily because of optical blur (Fig. 4.10A).

Next, we reconstructed images from the cone excitations produced by the grating stimuli. A template-matching observer based on the noise-free reconstructions was then applied to the noisy reconstructions (see *Methods*). The image-reconstruction observer shows significant interactions between spatial frequency and chromatic direction. Sensitivity in the L+M direction is relatively constant with spatial frequency. Sensitivity in the L-M direction starts out higher than L+M at low spatial frequencies, but drops significantly and is lower than L+M at high spatial frequencies (Fig. 4.10B). We attribute these effects to the role of the image prior in the reconstructions, which leads to selective enhancement/attenuation of different image components. In support of this idea, we also found that an observer based on maximum likelihood reconstruction without the explicit prior term produced CSFs similar in shape to the Poisson ideal observer (Fig. B.11).

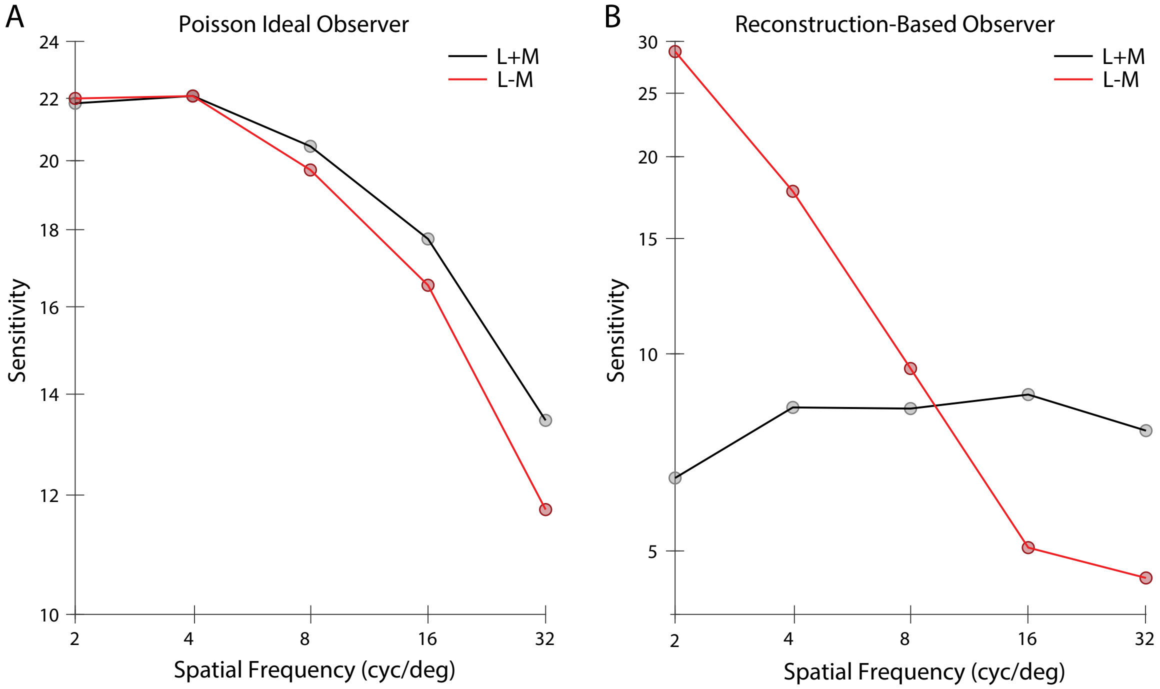

Figure 4.10: Contrast sensitivity functions. Contrast sensitivity, defined as the inverse of threshold contrast, for **A)** a Poisson 2AFC ideal observer, and **B)** an image reconstruction-based observer (see *Methods*), as a function of the spatial frequency of stimulus in either the L+M direction (black) and L-M cone contrast direction (red). Contrast was measured as the vector length of the cone contrast vector, which is matched across two color directions.

It is intriguing that the CSFs from the reconstruction-based observer show substantially higher sensitivity for L-M than for L+M modulations at low spatial frequencies (with equated RMS cone contrast), but with a more rapid falloff such that the sensitivity for L+M modulations is higher at high spatial frequencies. Both of these features are characteristic of the CSFs of human vision (Mullen, 1985; Anderson et al., 1991; Chaparro et al., 1993; Sekiguchi et al., 1993). A more comprehensive exploration of this effect and its potential interaction with other decision rule used in the calculation awaits future research.

4.4. Discussion

We developed a Bayesian image reconstruction framework for characterizing the initial visual encoding, by combining an accurate image-computable forward model together with a sparse coding model of natural image statistics. Our method enables both quantification and visualization of information loss due to various factors in the initial encoding, and unifies the treatment of a diverse set of issues that have been studied in separate, albeit related, ways. In several cases, we were able to extend previous studies by eliminating simplifying assumptions (e.g., by the use of realistic, large cone mosaics that operate on high-dimensional, naturalistic image input). To summarize succinctly, we highlight here the following novel results and substantial extensions of previous findings: 1) When considering the allocation of different cone types on the human retina, we demonstrated the importance of the spatial and spectral correlation structure of the image prior; 2) As we examined reconstructions as a way to visualize information loss, we observed rich interactions in how the appearances of the reconstruction vary with mosaic sampling, physiological optics, and the SNR of the cone excitations; 3) We found that the reconstructions are consistent with empirical reports of retinal spatial aliasing obtained with interferometric stimuli, adding an explicit image prior component and extending consideration of the interleaved nature of the trichromatic retinal cone mosaic relative to the previous treatment of these phenomena; 4) We linked image reconstructions to spatio-chromatic contrast sensitivity functions by applying a computational observer for psychophysical discrimination to the reconstructions. Below, we provide an extended discussion of key findings, as well as of some interesting open questions and future directions.

First, we cast retinal mosaic design as a "likelihood design" problem. We found that the large natural variations of L- and M-cone proportion, and the relatively stable but small S-cone proportion, can both be explained as an optimal design that minimizes the expected image reconstruction loss. This is closely related to an alternative formalism, often termed "efficient coding", which seeks to maximize the amount of information transmission (Barlow et al., 1961; Karklin and Simoncelli, 2011; Wei and Stocker, 2015; Sims, 2018). In both cases, the optimization problem is subject to realistic biological constraints and incorporates natural scene statistics. Previous work Garrigan et al. (2010) conducted a similar analysis with consideration of natural scene statistics, physiological optics, and cone spectral sensitivity, using an information maximization criterion. One advance enabled by our work is that we are able to fully simulate a 1-deg mosaic with naturalistic input, as opposed to the information-theoretical measures used by Garrigan et al. (2010), which became intractable as the size of the mosaic and the dimensionality of the input increased. In fact, Garrigan et al. (2010)

approximated by estimating the exact mutual information for small mosaic size ($N = 1 \dots 6$ cones) and then extrapolated to larger cone mosaics using a scaling law (Borghuis et al., 2008). The fact that the two theories corroborate each other well is reassuring and suggests that the results are robust to the details of the analysis.

Our approach could be applied to analyzing the retinal mosaic characteristics of different animals. Adult zebrafish, for example, feature a highly regular mosaic with fixed 2:2:1:1 R:G:B:U cone ratios (Engström, 1960). Since our analysis has highlighted the importance of prior statistics in determining the optimal design, one might speculate whether this regularity results from the particular visual world of zebrafish (i.e., underwater, low signal-to-noise ratio), which perhaps demands a more balanced ratio of different cone types to achieve the maximum amount of information transmission. Further study that characterizes in detail the natural scene statistics of the zebrafish's environment might help us to better understand this question (Zimmermann et al., 2018; Cai et al., 2020). It would also be interesting to incorporate into the formulation an explicit specification of how the goal of vision might vary across species. One extension to the current approach to incorporate this would be to specify an explicit loss function for each species and find the reconstruction that minimizes the expected (over the posterior of images) loss (Berger, 2013), although implementing this approach would be computationally challenging. Related is the task-specific accuracy maximization analysis formulation (Burge and Geisler, 2011).

Second, we applied our framework to cone excitations of retinal mosaics with varying degrees of optical quality, photoreceptor size, density, and cone spectral sensitivity. The reconstructed images reflect accurately the information loss in the initial encoding, including spatial blur due to optical aberration and mosaic sampling, pixel noise due to Poisson variability in the cone excitations, and reduction of chromatic contrast in anomalous trichromacy. Although we have mainly focused on visualization of these effects in our current paper, it would be possible to perform quantitative analyses. In fact, our reconstruction algorithm could provide a natural "front-end" extension to many image-based perceptual quality metrics, such as spatial CIELAB (Zhang et al., 1997; Lian, 2020), structural similarity (Wang et al., 2004), low-level feature similarity (FSIM; Zhang et al. (2011)), or neural network-based approaches (Bosse et al., 2018). Doing so would incorporate factors related to the initial visual encoding explicitly into the resulting image quality metrics.

In addition, when SNR is high, we found that we are able to fully recover color information even from an anomalous trichromatic mosaic. As SNR drops, this becomes less feasible. Although our analysis does underestimate the amount of total noise in the visual system (i.e., we only consider noise at cone excitations, but see Angueyra and Rieke (2013) for a detailed treatment of noise in the retinal), this nonetheless suggests that a downstream circuit that properly compensates for the shift in cone spectral sensitivity can, in principle, maintain relatively normal color perception in the low noise regime (Tregillus et al., 2021). This may potentially be related to some reports of less than expected difference in color perception between anomalous trichromats and color normal observers (Bosten, 2019; Lindsey et al., 2020).

Third, we speculate that image reconstruction could provide a reasonable proxy for modeling percepts in various psychophysical experiments. We found that images reconstructed from dichromatic mosaics resemble results generated by previously proposed methods for visualizing dichromacy, including one that uses explicit knowledge of dichromatic subjects' color appearance reports (Brettel et al., 1997). We have also reproduced the "zebra stripes" and "orientation reversal" aliasing patterns when reconstructing images from cone excitations to spatial frequencies above the mosaic Nyquist limit, similar to what has been documented experimentally in human subjects (Williams, 1985; Coletta and Williams, 1987). In a similar vein, previous work has used a simpler image reconstruction method to model the color appearance of small spots light stimulus presented to single cones using adaptive optics (Brainard et al., 2008). Our method could also be applied to such questions, and also to a wider range of adaptive optics (AO) experiments (e.g., Schmidt et al. (2019); Neitz et al. (2020)), to help understand the extent to which image reconstruction can capture perceptual behavior. More speculatively, it may be possible to use calculations performed within the image reconstruction framework to synthesize stimuli that will maximally discriminate between different hypothesis about how the excitations of sets of cones are combined to form percepts, particularly with the emergence of technology that enables precise experimental control over the stimulation of individual cones in human subjects (Harmening et al., 2014; Sabesan et al., 2016; Schmidt et al., 2019).

Last, we showed that our method can be used in conjunction with analysis of psychophysical discrimination performance, bringing to this analysis the role of statistical regularities of natural images. In our initial exploration, we found that the image-reconstruction based observer exhibits significant interaction between spatial frequency and chromatic direction in its contrast sensitivity function, a behavior distinct from its Poisson ideal observer counterpart, and is more similar to the human observer. Future computations will be needed to understand in more detail whether the reconstruction approach can account for other features of human psychophysical discrimination performance that are not readily explained by ideal-observer calculations applied to the cone excitations.

Our current model only considers the representation up to and including the excitations of the cone mosaic. Post-excitation factors (e.g., retinal ganglion cells), especially in the peripheral visual field, are likely to lead to additional information loss. In this regard, we are eager to incorporate realistic models of retinal ganglion cells into the ISETBio pipeline. Nevertheless, the value of the analysis we have presented is to elucidate exactly what phenomena can or cannot be attributed to factors up to the cone excitations, thus helping to dissect the role of different stages of processing in determining behavior. For example, we found there is desaturation of chromatic content in reconstructed images in the periphery, with the details depending on interactions between the physiological optics, cone mosaic sampling, macular pigment density, and the model of natural image statistics. This is in contrast to more traditional explanations of the decrease in peripheral chromatic sensitivity, which often consider it in the context of models of how different cone types are wired to retinal ganglion cells (e.g., Lennie et al. (1991); Mullen and Kingdom (1996); Hansen et al. (2009); Field et al. (2010); Wool et al. (2018)). Whether the early vision factors are sufficient to account for the full variation in chromatic sensitivity awaits a more detailed future study, but the fact that early vision factors can play a role through their effect on the available chromatic information is a novel insight that should be incorporated into thinking about the role of post-excitation mechanisms.

More generally, we can consider the locus of the signals analyzed in the context of the encoding-decoding dichotomy of sensory perception (Stocker and Simoncelli, 2006; Rust and Stocker, 2010). Here we reconstruct images from cone excitations, thus post-excitation processing may be viewed as part of the brain's implementation of the reconstruction algorithm. When we apply such an algorithm to, for example, the output of retinal ganglion cells, we shift the division. Our view is that analyses at multiple stages are of interest, and eventual comparisons between them are likely to shed light on the role of each stage.

Our current model also does not take into account fixational eye movements, which displace the retinal image at a time scale shorter than the integration period we have used (Martinez-Conde et al., 2004; Burak et al., 2010). It has been shown that these small eye movements can increase psychophysical visual acuity relative to that obtained with retinal-stabilized stimuli (Rucci et al., 2007; Ratnam et al., 2017). An intuition behind this is that fixational eye movements can increase the effective cone sampling density, if the visual system can sensibly combine information obtained across multiple fixation locations. This intuition is supported by computational analyses that integrate information across fixations while simultaneously estimating the eye movement path (Burak et al., 2010; Anderson et al., 2020). In their analysis, Burak et al. (2010) showed the effectiveness of their algorithm depended both on the integration time of the sensory units whose excitations were processed, and also on the receptive field properties of those units. In addition, consideration of the effects of fixational eye movement might also benefit from an accurate model of the temporal integration that occurs within each cone, as a consequence of the temporal dynamics of the phototransduction cascade (Angueyra and Rieke, 2013). ISETBio in its current form implements a model of the phototransduction cascade as well as of fixational eye movements (see Cottaris et al. (2020)). Future work should be able to extend our current results through the study of dynamic reconstruction algorithms within ISETBio.

Since our framework is centered on image reconstruction, one may naturally wonder whether we should have applied the more "modern" technique of convolutional neural networks (CNNs), which have become the standard for image processing-related tasks (Krizhevsky et al., 2012). For our scientific purposes, the Bayesian framework offers an important advantage in its *modularity*, namely, the likelihood and prior are two separate components that can be built independently. This allows us to easily isolate and manipulate one of them (e.g., likelihood) while holding the other constant (e.g., prior), something we have done throughout this paper. In addition, building the likelihood function (i.e., render matrix R, see *Methods*) is a forward process that is computationally very efficient. Performing a similar analysis with the neural network approach (or supervised learning in general) would require re-training of the network with a newly generated dataset (i.e., cone excitations paired with the corresponding images) for *every* condition in our analysis.

However, the ability of neural networks to represent more complex natural image priors (see Ulyanov et al. (2018); Kadkhodaie and Simoncelli (2021) for examples) is of great interest. Currently, we have chosen a rather simple, parametric description of natural image statistics, which leads to a numerical MAP solution. Previous work has proposed methods that alternate, within each iteration, between regularized reconstruction and denoising, which effectively allow for transfer of the prior implicit in an image denoiser (e.g., a deep neural network denoiser) to be applied to any other domain with a known likelihood model (Venkatakrishnan et al., 2013; Romano et al., 2017). More recently, Kadkhodaie and Simoncelli (2021) developed a related but more explicit and direct technique to extract the image prior (a close approximation to the gradient of the log-prior density, to be precise) from a denoising deep neural network, which could be applied to our image reconstruction problem. We think this represents a promising direction, and in the future plan to incorporate more sophisticated priors, to evaluate the robustness of our conclusions to variations and improvements in the image prior.

To conclude, we believe our method is widely applicable to many experiments (e.g., adaptive optics psychophysics) designed for studying the initial visual encoding, for modeling the effect of changes of various components in the encoding process (e.g., in clinical cases), and for practical applications (e.g., perceptual quality metric) in which the initial visual encoding plays an important role.

CHAPTER 5

OPTIMAL LINEAR MEASUREMENTS FOR NATURAL IMAGE RECONSTRUCTION USING THE PRIOR IMPLICIT IN A DENOISER

The work presented in this chapter is performed in collaboration with Zahra Kadkhodaie, Eero P. Simoncelli, and David H. Brainard. Part of the work in this chapter was previously presented as: Zhang et al. (2022b). I contributed to the conceptualization, formal analysis, methodology, validation, software, visualization, and writing of this work.

Abstract

We developed a method for finding the optimal set of linear measurements given an ensemble of natural images. We define the optimal measurements as those that, when combined with a prior model of images, minimize the mean squared error of the Bayesian image reconstruction. For simple prior models, the solution to this problem is closely related to techniques including principal component analysis (PCA) and compressed sensing (CS). Here, we utilize a complex prior over images, more specifically, the prior implicit in a convolutional neural network trained to perform image denoising. This prior is combined with the linear measurements in a coarse-to-fine projected gradient ascent procedure to reconstruct high-probability images. We find the optimal measurements by performing stochastic gradient descent in the space of orthogonal matrices with respect to the estimation loss. Our method discovered measurement subspaces that are distinct from and outperform traditional methods such as PCA and CS. Furthermore, we found that our method is also able to generalize to different loss functions beyond squared error, such as structural similarity index measure (SSIM). Our method will have important implications for both designing image processing pipelines and understanding biological visual systems.

5.1. Introduction

When working with natural images in various applications, it is often necessary to make linear measurements of the original signal. For example, the optical elements of a camera can be modeled as a linear system; In medical imaging, both x-ray and ultrasound produce linear measurements from the tissue; For the human visual system, the initial sensory encoding by the photoreceptors is also approximately linear.

We consider a general form of the linear inverse problem (Tropp and Wright, 2010). Concretely, given the original signal $x \in \mathcal{R}^n$ and a measurement matrix $M \in \mathcal{R}^{n \times c}$, the linear measurements are $m = M^T x$. Typically, we have $c < n$, and we are interested in an estimate $\hat{x}$ based on the measurements m. We can formulate this inverse problem in Bayesian terminology. Denote our measurement model as $p(m|x)$. When there is no measurements noise, $p(m|x)$ is a delta function of the form $\delta(m - M^T x)$, but it is also easy to construct $p(m|x)$ with realistic noise models such as Gaussian and Poisson. We also need to specify a prior distribution $p(x)$ that characterizes the statistical regularities of x. The estimate $\hat{x}$ that minimizes the mean squared error is the posterior mean $\hat{x} = \int x \cdot p(x|m)\ dx$, where $p(x|m) = p(m|x)p(x)/(\int p(m|x)p(x)dx)$.

We are interested in the optimal linear measurement matrix M. That is, the measurement matrix that minimizes the Bayesian estimation loss for a given number of measurements m. In camera and sensory design, this is known as an optimal design problem (Levin et al., 2008), while in sensory neuroscience, it is considered a form of the linear efficient coding problem (Atick, 1992). The solution to M is known for simple forms of $p(x)$. For example, when $p(x)$ is Gaussian distributed, a linear projection that maximizes total variance also minimizes the mean squared error, thus simply performing a principal component analysis (PCA) is sufficient (Abdi and Williams, 2010). On the other hand, the compressed sensing (CS) literature (Donoho, 2006; Davenport et al., 2012) suggests that when $p(x)$ satisfies certain sparsity properties, random linear projections actually allow for near-perfect estimation of the original signal. However, natural images have complex structures that neither the Gaussian nor sparse models can sufficiently describe, and thus it is possible to find better linear measurements that outperform PCA and CS (Weiss et al., 2007; Chang et al., 2009).

Previously, Kadkhodaie and Simoncelli (2021) developed a coarse-to-fine projected gradient ascent algorithm for Bayesian image reconstruction by drawing high-probability samples of natural images that conform to a set of linear constraints defined by the measurements. They achieve this by combining the linear measurement model with the prior implicit in a convolutional neural network

(CNN) denoiser (Mohan et al., 2019). In this paper, we develop a method for finding the optimal measurement matrix M for this more expressive prior model. Concretely, we perform stochastic gradient descent in the space of all orthonormal basis by taking derivatives of M with respect to the estimation loss through the reconstruction algorithm.

We validated our method by showing that for simple linear projections, we correctly recover the PCA solution. We then applied our method to a dataset of celebrity face images (Liu et al., 2015). By incorporating the corresponding denoiser implicit prior, we discovered measurement subspaces that are superior to and distinct from traditional methods like PCA and CS. Additionally, we demonstrated the generalizability of our approach to other loss functions, including the structural similarity index measure (SSIM, Wang et al. (2004)). Our methods offers a unified perspective of both PCA and CS as optimized linear measurements for a specific prior distribution of the signal. Moreover, our results highlight the potential for improved measurements with improved statistical characterization of the signal itself.

5.2. Methods

5.2.1. Problem formulation

We are interested in taking linear measurements of some signal $x \in \mathcal{R}^n$, represented by a measurement matrix of the form $M \in \mathcal{R}^{n \times m}$ (typically m < n). Our formulation is general, but we focus on x as natural images in the current paper. Given some measurements $m = M^T x$, we would like to estimate the original x as accurately as possible. We use $\hat{x} = h(m; M)$ to denote a generic estimator function h that produces an estimate $\hat{x}$ given m, based on the measurement model M. The expected mean squared error (MSE) of this estimator is as follows:

$$l(M) = E_{x \sim p(x)}[\ ||h(M^T x; M) - x||_2^2\ ] \tag{5.1}$$

The optimal linear measurement is the M that minimizes the expected MSE, $M = \text{argmin}_M\ l(M)$. When there is no noise in the measurement, without loss of generality, we can consider a subset of M where the column vectors of M are orthonormal, that is, $M^T M = I$. We will expand on the issue of noise in future work.

5.2.2. Bayesian image reconstruction

To fully specify the problem, we need to define the estimator function h. Since we aim to recover x based on a smaller number of measurements, the problem is underdetermined. Thus, an image prior $p(x)$ is required to properly regularize the solution. Kadkhodaie and Simoncelli (2021) developed a method for sampling high-probability natural images from a complex image prior, confined to linear constraints defined by orthogonal measurements. Here we will provide a brief overview of the methods, and refer the readers to the original article for more details.

Prior implicit in a denoiser

We start by training a convolutional neural network for blind noise removal (Mohan et al., 2019). In particular, we consider additive Gaussian noise of the form $y = x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The network $f(\cdot)$ is trained to produce a denoised image $\hat{x}$ that minimizes the expected MSE $||x - \hat{x}||_2^2$, across different σ^2. In Bayesian terms, the solution to this denoising problem is the posterior mean:

$$\hat{x} = \int x \cdot p(x|y)\ dx = \int x \cdot \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}\ dx. \tag{5.2}$$

Miyasawa (1961) showed that the estimator can be rewritten to an equivalent form. First, define the marginal distribution $p(y)$:

$$p(y) = \int p(y|x)p(x)dx. \tag{5.3}$$

Plugging in the Gaussian PDF for $p(y|x)$, and taking derivatives on both sides, we have:

$$\nabla_y p(y) = \frac{1}{\sigma^2} \int (x - y)p(y|x)p(x)dx, \tag{5.4}$$

and dividing both sides by $p(y)$ yields:

$$\sigma^2 \frac{\nabla_y p(y)}{p(y)} = \int (x - y)\frac{p(x, y)}{p(y)}dx = \int (x - y)p(x|y)dx, \tag{5.5}$$

$$\sigma^2 \nabla_y \log p(y) = \int xp(x|y)dx - \int yp(x|y)dx = \hat{x} - y. \tag{5.6}$$

Thus we have $\hat{x} = y + \sigma^2 \nabla_y \log p(y)$. This establishes a direct relationship between the residue of the denoiser network output and image prior (Kadkhodaie and Simoncelli, 2021). That is, we can use the denoiser $f(\cdot)$ to estimate the gradient of the log marginal density $p(y)$, as $\nabla_y \log p(y) = f(y) - y$.

Sample from the image prior

This gradient can be used in an annealed Langevin dynamics (Bussi and Parrinello, 2007) to sample from the prior $p(x)$ as:

$$y_{t+1} = y_t + \sigma_t^2 \nabla_{y_t} \log p(y) + \eta_t, \quad \eta \sim \mathcal{N}(0, I). \tag{5.7}$$

Kadkhodaie and Simoncelli (2021) developed a method to solve the linear inverse problem with the denoiser implicit prior. Conditioning the prior gradient on the measurement, we obtain:

$$\nabla_{y_t} \log p(y|m) = (I - MM^T) \nabla_{y_t} \log p(y) + M(m - M^T y_t). \tag{5.8}$$

The conditional gradient can be used in a procedure the same as the above for sampling high-probability natural images that also conform to the linear constraints defined by the measurements. In the current paper, we will define the estimator function $\hat{x} = h(m)$ by averaging multiple samples obtained from running this procedure at a randomly chosen starting point.

5.2.3. Optimal measurement matrix

We search for the optimal measurement matrix using stochastic gradient descent in the space of orthonormal matrices. Concretely, we parameterize the measurement matrix M using Householder product, which represents orthonormal matrices as a sequence of elementary reflections. See Shepard et al. (2014, 2015) for a detailed introduction of this parameterization:

$$Q = H_1 H_2 ... H_k, \text{ where } H_i = I - \tau v_i {v_i}^T \tag{5.9}$$

The collection of v_i's is the essential parameters ϕ of the parameterization $Q(\phi)$, which forms a lower triangular matrix of $\mathcal{R}^{n \times m}$.

We rewrite our objective function Eq. 5.1 using the parameterization $Q(\phi)$, and approximate the expectation with samples of images from the training set:

$$l(\phi) = \frac{1}{N}\sum_{i=1}^{N} ||h(Q(\phi)^T x_i;\ Q(\phi)) - x_i||_2^2. \tag{5.10}$$

We can then search for the optimal measurement matrix through stochastic gradient descent (SGD):

$$\phi_{t+1} \leftarrow \phi_t - \lambda \cdot \nabla_{\phi_t} l(\phi_t) \tag{5.11}$$

In practice, the optimization is implemented with a variant of SGD that attempts to approximate second-order (curvature) information (Kingma and Ba, 2014).

5.2.4. Training and optimization

Image dataset

All experiments reported in this paper are conducted on the CelebA dataset from Liu et al. (2015). The dataset consists of over 200K images of celebrity faces. We downsampled the images to size 50 by 40, and applied a standard inverse Gamma correction for display nonlinearity. We randomly sampled 100 images as a hold-out validation set, and used the rest as the training set.

Denoiser

We trained a bias-free CNN denoiser with an architecture similar to that reported in Mohan et al. (2019). The network consists of repeated convolutional layers and nonlinearity, with a skip connection from the input to the output layer. The model is trained to perform Gaussian noise removal on the CelebA dataset by minimizing MSE using the Adam optimizer (Kingma and Ba, 2014). See Mohan et al. (2019) for more details.

Measurement matrix

To find the optimal measurement matrix, we use the Adam optimizer, with the gradient computed as above (Eq. 5.10). Due to performance considerations, we only take one sample when reconstructing images in $h(m)$. We used a batch size of 128, and the optimization is run for 50 epochs.

5.3. Results

5.3.1. Principal component analysis

We can reconstruct images by projecting them into the measurement subspace: $\hat{x} = MM^T x$. In this case, the matrix M that minimizes the objective

$$E_{p(x)} \ ||MM^T x - x||_2^2 \tag{5.12}$$

is simply the first m principal component of the data. To validate our method, we find the optimal matrix for this linear projection problem, and compare our result with the standard PCA solution, which is computed using the singular value decomposition (SVD).

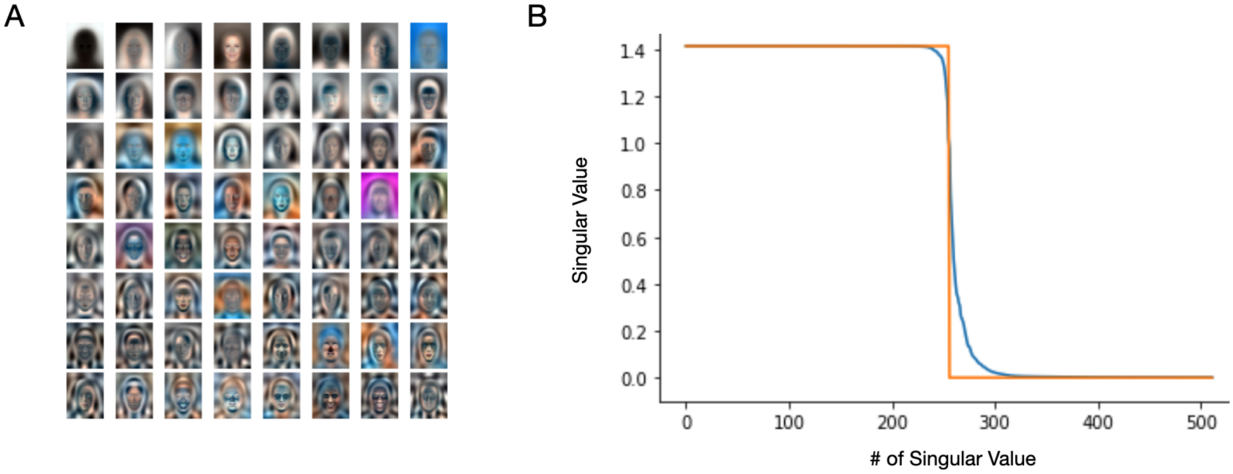

Figure 5.1: Compare the standard PCA and optimized solutions. **A)** The first 64 PCs of the CelebA dataset computed using SVD are shown as images. **B)** The singular values of a matrix with $m = 512$, constructed by concatenating two measurement matrices with $m = 256$. For the orange line, the two matrices are identical and are the first 256 PCs. For the blue line, the two matrices are the first 256 PCs, and the optimized measurement matrix computed by our method for the linear projection case (Eq. 5.12) respectively.

The PCs of the CelebA dataset resemble the eigenface basis (Turk and Pentland, 1991). When visualized as images, they appear to be a sequence of face templates with increasing spatial frequency (Fig. 5.1A). For a fixed number of measurements m, we need to confirm that the optimized matrix span the same subspace as the first m PCs. Thus, we construct a matrix by concatenating two measurement matrices, one consists of the first m PCs, and the other is the optimized measurement matrix. If the two subspaces spanned by the two matrices are the same, then only the first m singular values of the combined matrix should be non-zero. Indeed, we found that the solution found by our algorithm spans nearly the same subspace as the PCs (Fig. 5.1B). We suspect the slight amount of misalignment is due to the noise in SGD.

5.3.2. Optimal measurement matrix

The PCs are the optimal solution that minimizes MSE when the input x is Gaussian distributed. Face images, on the other hand, exhibit complex statistical regularities that are characterized by our denoiser implicit prior. Thus, we are interested in whether it is possible to find a measurement matrix that when combined with the prior, outperforms the PCs.

Below we show some example reconstructions for $m = 2$ (Fig. 5.2). Note that there are a total of $6,000$ pixels for each image, so this is an extremely small number of measurements. As expected, the two PCs can only retain very little information about the original images (Fig. 5.2B). Simply combining the image prior with the PCs already improved the visual quality of the reconstructed images, and reduced the MSE (Fig. 5.2C). More importantly, optimizing the two measurement vectors with respect to the denoiser prior further improved the quality of the reconstructions (Fig. 5.2D). And although the reconstructed images do not reflect the identities of the original ones (as we are only taking two measurements), they have a striking resemblance to real face images. Lastly, the decrease in MSE is almost universal across all images in our testing set (Fig. 5.3A). Thus, when a more complex prior model is employed, the best two linear measurements are not the PCs. We confirmed that our conclusion holds for a range of m, albeit the improvements are smaller when m gets larger. See Fig. 5.3B for a case of $m = 128$.

How different are the optimal measurements compared to the PCs? In Fig. 5.4A, we visualize the measurement vectors from an optimized matrix of $m = 64$. We found that each measurement vector has a “face template” appearance, although they all roughly have a similar spatial frequency, contrary to the PCs. We further conducted an analysis similar to that of Fig. 5.1B. We found that the optimized measurements indeed span a distinct subspace, and measure higher frequency

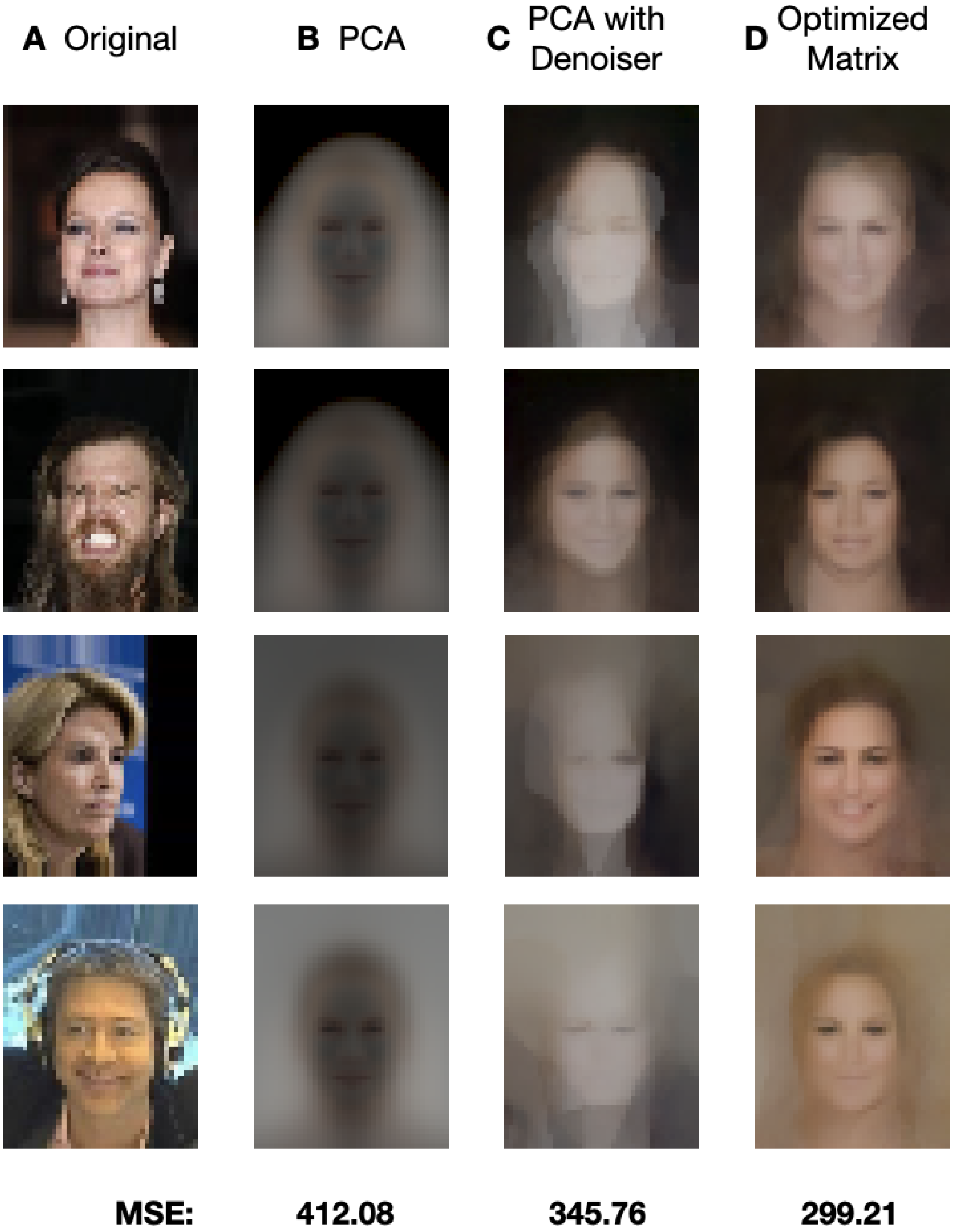

Figure 5.2: Example reconstructions for $m = 2$. From left to right, the columns of images are the **A)** the originals; **B)** images projected onto the first two PCs; **C)** images reconstructed by combining the first two PCs with the denoiser implicit prior; and **D)** images reconstructed with the linear measurements optimized for the implicit prior. The numbers at the bottom are the MSE across the test set ($N = 100$).

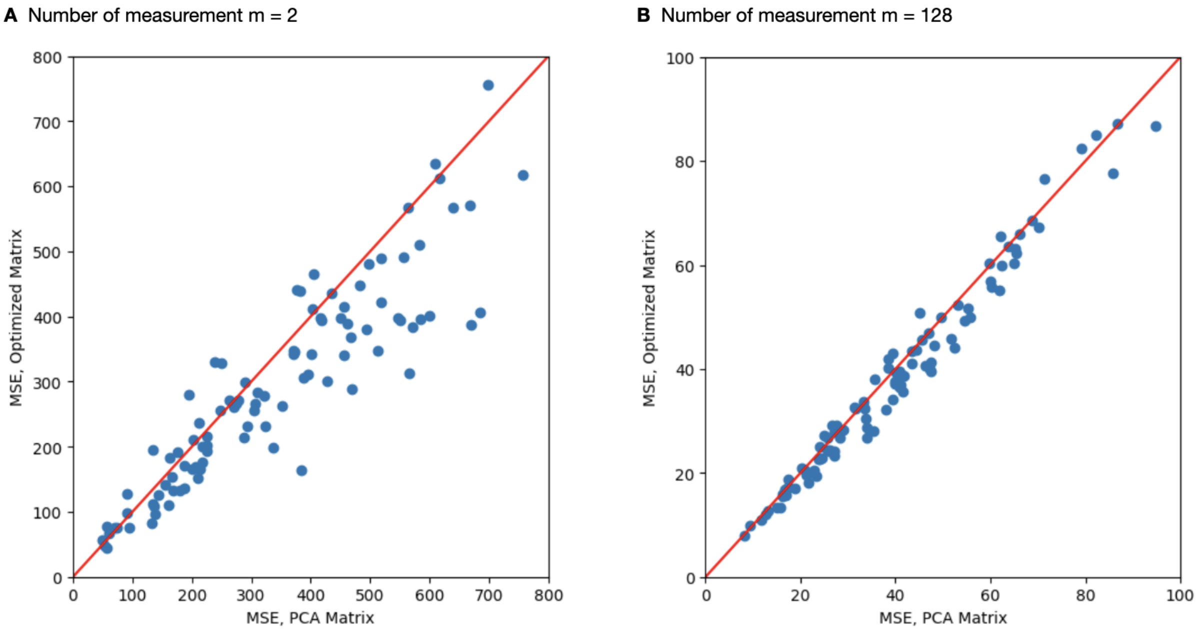

Figure 5.3: Reconstruction error of single images across two measurement matrices. Scatter plots of the sum of squared error between the original and the reconstructed images, when using the PCs combined with the implicit prior (x-axis), and the optimized measurement matrix (y-axis). **A)** Number of measurement $m = 2$. **B)** Number of measurement $m = 128$.

content, compared to its PC counterpart (Fig. 5.4B). Lastly, the fact that the measurements have a coherent template appearance, and are consistent across multiple runs of our algorithm also suggests they are not simply the random projections proposed by the compressed sensing (CS) literature (Donoho, 2006; Davenport et al., 2012). This is consistent with previous results (Weiss et al., 2007; Chang et al., 2009), which demonstrated that natural images do not satisfy the strict sparsity condition assumed in CS, and thus even PCA can outperform random projections.

5.3.3. Structural similarity index measure (SSIM)

Our optimization formulation is general, which allows us to potentially generalize our approach to other objective functions beyond minimizing MSE. In fact, it has been shown that MSE does not correlate well with how the quality of images is perceived by human observers. SSIM is developed as a perceptual quality metric that can better predict human judgments (Wang et al., 2004). Thus, here we set out to search for measurement matrices that when combined with the denoiser prior, maximize the SSIM score. Below we show some examples from a case of $m = 64$. Similar to

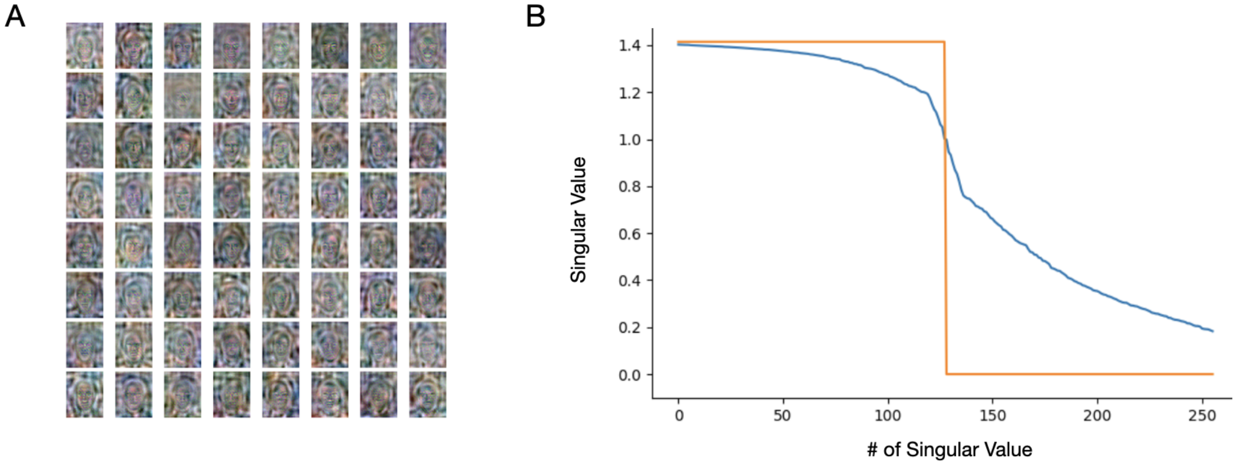

Figure 5.4: Optimal measurements. **A)** The measurement vectors from the optimized matrix with $m = 64$ are shown as images. **B)** The singular values of a matrix with $m = 256$, constructed by concatenating two measurement matrices with $m = 128$. For the orange line, the two matrices are identical and are the first 128 PCs. For the blue line, the two matrices are the first 128 PCs, and the optimized matrix with respect to the denoiser implicit prior, respectively (also see Fig. 5.1).

what we have observed previously (Fig. 5.2), simply adding the denoiser prior (Fig. 5.5C) already increases the quality of the reconstruction compared to linear projection onto the PCs (Fig. 5.5B). Optimizing the measurement matrix with respect to SSIM provides further improvements, which can be seen from both the visual quality of the images and the average SSIM score (Fig. 5.5D). Same as before, we also confirmed that the improvement is nearly universal across all images in our test set. In addition, the matrix optimized for SSIM spans a subspace distinct from both the PCs, and the matrix optimized for MSE. Thus, we have demonstrated that our method can indeed generalize to different objective functions.

5.4. Discussion

We developed a method for finding the optimal linear measurements that minimize the Bayesian image reconstruction error, given a dataset of natural images. Our method takes full advantage of the complex statistical regularities of images, characterized by a prior implicit in a CNN denoiser (Kadkhodaie and Simoncelli, 2021). We demonstrated our method on the CelebA face dataset (Liu et al., 2015). We found measurement matrices that span distinct subspaces compared to the PCs, and can drastically improve the quality of the image reconstruction when combined with the

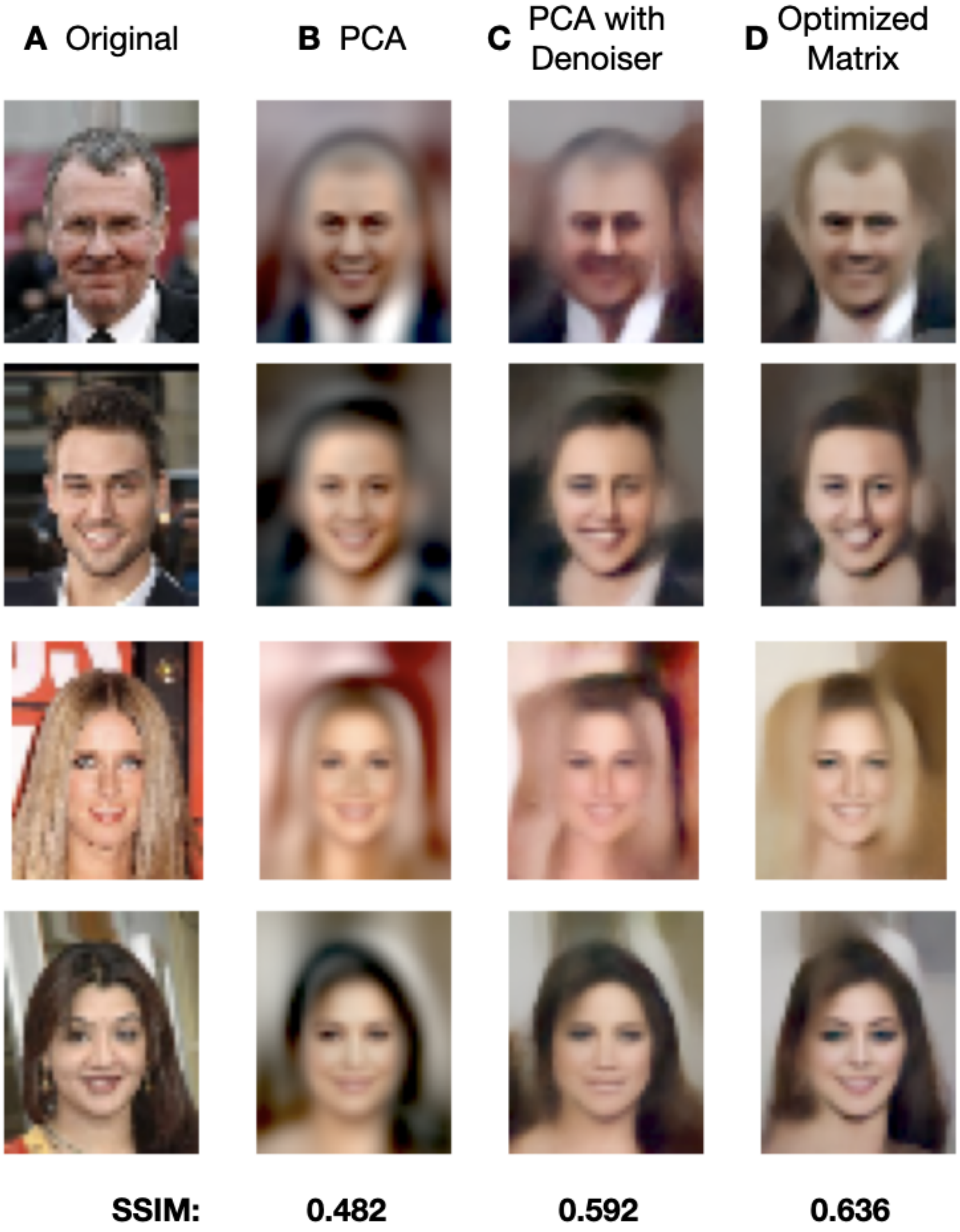

Figure 5.5: Example reconstructions for $m = 64$, with the measurement matrix optimized for SSIM. From left to right, the columns of images are the **A)** the originals; **B)** images projected onto the first two PCs; **C)** images reconstructed by combining the first two PCs with the denoiser implicit prior; and **D)** images reconstructed with the linear measurements optimized for the implicit prior, with SSIM as the objective. The numbers at the bottom are the average SSIM across the test set ($N = 100$).

denoiser prior. Our method outperforms both PCA and CS methods. Further, our method can be generalized to other loss functions such as SSIM, which better measures perceptual quality.

We can view our problem as a form of linear efficient coding: Finding the optimal linear encoding that minimizes the estimation error, subject to the constraint that there can be only a limited number of measurements. The efficient coding framework emphasizes that the solution to this problem depends strongly on the prior distribution of the signal $p(x)$. By adopting the efficient coding perspective, we can unify our approach with both PCA and CS: When x is Gaussian distributed, the PCs are indeed the optimal solution. On the other hand, when x exhibits certain sparsity properties, CS represents the most effective approach. Our finding that some measurement matrices outperform both indicates that neither Gaussian nor sparse distribution is sufficient to characterize natural face images, and the denoiser prior is in fact a better model of image statistics.

Previous work has explored combining CS with modern machine learning techniques. Bora et al. (2017) showed that substantial performance gain can be obtained by combining random Gaussian measurements with complex prior models defined by variational autoencoder and generative adversarial networks, instead of the standard sparsity prior. Although they did not try to optimize the measurements, this has already demonstrated the effectiveness of better prior models. Wu et al. (2019) optimized jointly the linear measurement together with a nonlinear reconstruction network, which eliminated the explicit optimization required when estimating the original signal. The image prior in their case is implicit in the nonlinear network. The advantage of our framework is that we can optimize the measurement matrix through a Bayesian image reconstruction procedure (Kadkhodaie and Simoncelli, 2021), with the prior model defined separately through training a denoiser. This allows us to isolate and examine the effect of the linear measurements separately from the image prior model.

As the number of measurements taken is usually smaller than the number of pixels in the images, our problem can be considered a special case of image compression, where the encoder is constrained to be linear, and the decoder is nonlinear. Superior performance can be achieved by allowing a nonlinear encoder-decoder cascade to be jointly optimized (Ballé et al., 2016; Wu et al., 2019).

However, our formalism might still be applicable if there are physical or computational constraints on the encoding process. Relatedly, our method can also be viewed as a form of linear dimensionality reduction. Although images are embedded in a linear measurement subspace in our method, the impact of this is combined with the influence of the nonlinear prior. Thus, it would be intriguing to compare our method with other nonlinear dimension reduction techniques.

Both the MSE and SSIM objectives we have explored here aim to reconstruct the original image faithfully. As our formulation is generic, we can optimize measurements toward specific tasks, for example, correctly identifying the face identities from the reconstructed images. In this regard, we are interested in exploring systematically how task demands impact the optimal measurement subspace in future work. This is closely related to accuracy maximization analysis, a linear dimensionality reduction method developed for optimal estimation of latent variables from natural signals (Burge and Jaini, 2017).

Our current model assumes the encoding process is noise-free. Our conditional sampling algorithm can be extended to accommodate Gaussian additive noise, and thus by optimizing the measurement matrix through such routine, we can find the optimal measurements that are more robust to noise. We will explore this possibility in future work. Lastly, our work will also have implications for understanding the visual system, in particular understanding the early visual encoding as optimally measuring natural images (Zhang et al., 2022a; Roy et al., 2021; Jun et al., 2021). To do so will require us to include more biologically realistic constraints such as the locality of the measurement. To summarize, our method provides a powerful and unifying approach to understanding linear measurements in relation to the statistical regularities of natural images, and has potential implications for both image processing, and biological vision.

CHAPTER 6

SUMMARY AND DISCUSSION

6.1. Summary of Contributions

Any perceptual system needs to solve the daunting challenge of representing the external world faithfully with its limited resources, despite the vast space of possibilities. Information theory suggests that to overcome this challenge, the system must exploit the statistical regularities of the signal (Huffman, 1952; Barlow et al., 1961). In this thesis, I explored how this general hypothesis can establish a quantitative connection between behavior, neural representation, and image statistics.

In Chapters 2 and 3, I demonstrate how basic stimulus statistics are reflected in the neural encoding characteristics in terms of Fisher information. Specifically, the power-law slow speed prior accounts for the logarithmic speed encoding observed in the MT cortex, while the uneven distribution of orientations in natural scenes explains the anisotropic representation of orientation in the early visual cortex. Simultaneously, the same set of stimulus priors is consistent with perceptual behavior in a speed 2AFC task and an orientation estimation task. Lastly, I have shown preliminary behavioral evidence that the same notion can be applied to understand adaptation as efficient coding based on conditional stimulus distribution.

With the full characterization of stimulus distribution, one can derive the most informative sensory encoding given appropriate constraints. I explored this direction with state-of-the-art models of natural image statistics, and computational observers based on image reconstruction. In Chapter 4, I demonstrated how this optimal design principle could explain retinal encoding features, while in Chapter 5, I developed optimal linear measurements that surpass traditional methods such as PCA and CS by leveraging the complex statistical characteristics of natural images. It is worth noting that, in Chapters 2 and 3, there is a direct link between stimulus prior and encoding through Fisher information, whereas the Chapters 4 and 5, the prior is reflected in the encoding in a rather implicit way. It is an intriguing future question to understand the connections between the two approaches, and how the FI formulation can generalize to the higher dimensional, naturalistic settings, particularly in relationship to the prior (see Yerxa et al. (2020) for an example).

To summarize, the methods developed in this thesis establish a framework to quantitatively validate the prediction of normative theories of sensory representation at both neural and behavioral levels.

6.2. General Discussions

6.2.1. Neural representation of prior

Throughout this thesis, I have adopted an encoding-decoding notion for the visual system. It is highly plausible that the clear separation might not apply to the actual neural circuits. For example, Ganguli and Simoncelli (2014) proposed that as the encoding is matched to the prior, the anisotropy in neural representation can be taken advantage of to perform Bayesian inference with the stimulus prior. Many groups have also proposed neural circuits with the temporal dynamics set to directly perform inference (see e.g., Orbán et al. (2016); Haefner et al. (2016)). In these cases, as the neural responses are representing the posterior distribution (either as families of distribution or samples), the concept of encoding does not apply anymore in a literal sense. On the other hand, see Ma et al. (2006); Walker et al. (2020) for examples where the likelihood and prior are explicitly represented by distinct neural populations.

However, the advantage provided by the encoding-decoding framework is that it allows us to isolate the two essential components of perception of the level of computation: The representation of information and the interpretation of that information in order to infer latent variables of the external world. Importantly, as it is a computational-level theory, the framework itself actually does not subscribe to a particular implementation, although it may seem to imply so. As a particularly interesting example, Lange et al. (2020) demonstrated how a neural circuit set out for posterior inference can be described by models with a separate representation of likelihood and prior. Qiu and Stocker (2021) also showed that the same notions can be used to describe artificial neural networks that by definition, are simply performing “decoding”.

A related issue is the exact meaning of probabilistic computation when used in the context of neuroscience. In the most moderate sense, as many problems faced by perception are by definition, ill-posed, some form of regularization (i.e., “prior”) must be required. It is a much stronger position, however, to assert that probabilistic quantities are explicitly represented in the circuits. It is worth noting that the Bayesian framework discussed here uses probabilistic language but does not necessarily require implementation based on Bayes’ rule or probability theory.

6.2.2. Generality of efficient coding theory

Efficient coding hypothesizes that the goal of the sensory cortex is to maximize information transmission. For the early stage of visual encoding, such as the retina, information maximization is a reasonable objective. However, it becomes increasingly difficult to choose the correct objective. In some sense, the ultimate goal of the brain is to produce decisions and actions that are advantageous for survival and reproduction, not encoding information (see Brette (2019), but see Berke et al. (2022) for an argument of why representing the external world faithfully is almost always desirable for the other objectives).

On the theoretical side, there have been many attempts to develop theories of representation by explicitly incorporating the behavioral objective (Wang et al., 2016b; Park and Pillow, 2017), including Chapter 5 of this thesis. Empirically, it has been shown that the concept of efficient representation can be expanded to include task variables (Koay et al., 2022) and state space (Tomov et al., 2020). Still, ultimately a goal needs to be assumed for neural populations to apply the normative theories, although the assumption could be proven inaccurate (Musall et al., 2019).

One potential solution is to combine the normative approach with data-driven methods that can generate these initial hypotheses from large-scale data of behavior and neural activities. Especially in the context of complex, naturalistic behavior, these methods can identify behavioral objectives that are not obvious from the experimenter’s perspective (see e.g. Johnson et al. (2020); Ashwood et al. (2022)). Thus, the combination of purely normative theories and data-driven methods that can infer biological variables has the potential to further advance our understanding of how the brain supports adaptive behavior.

APPENDIX A

SUPPLEMENTARY FIGURES FOR CHAPTER 3

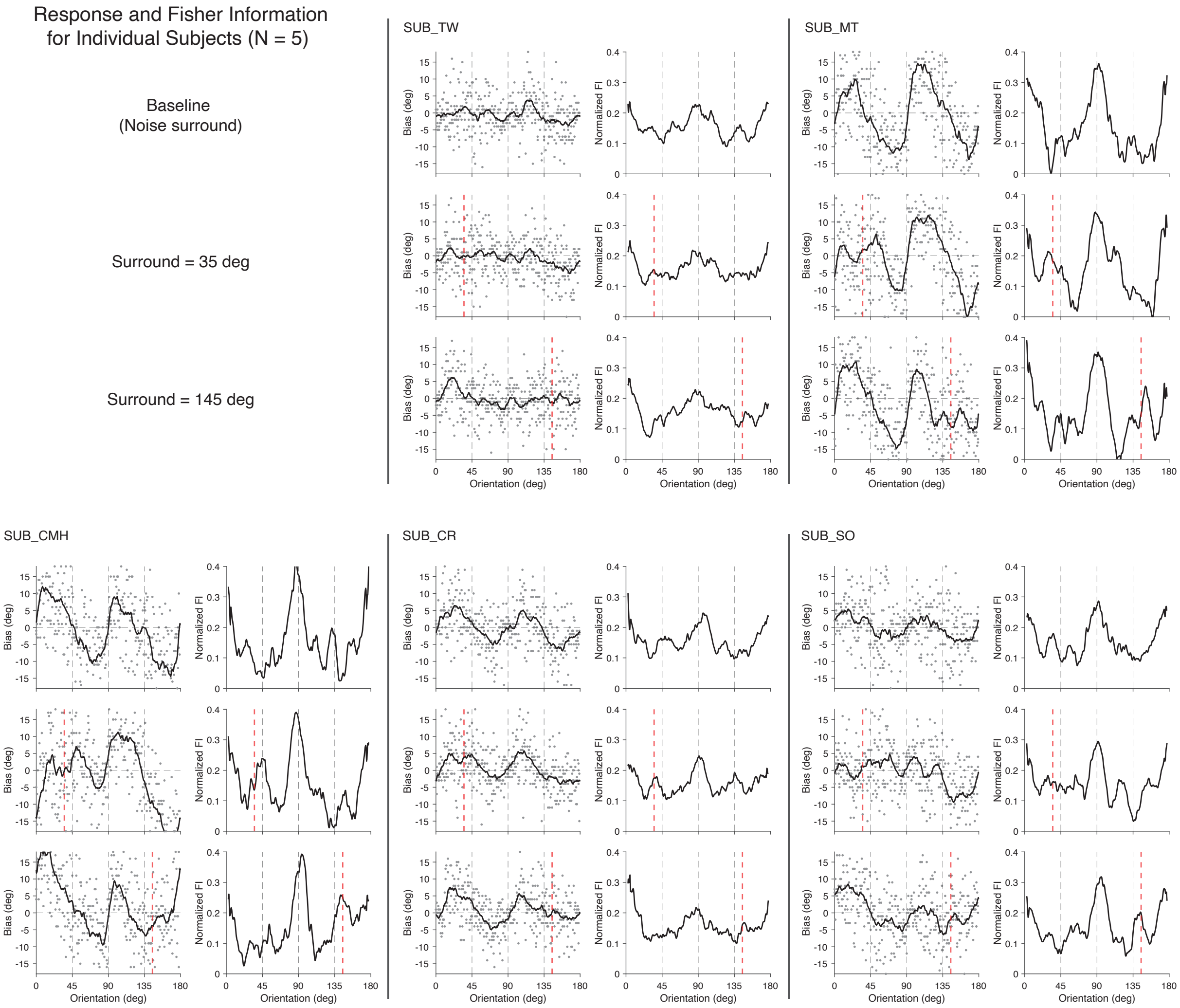

Figure A.1: Estimation bias and normalized Fisher Information (FI) for individual subjects. In each panel, the left column shows the orientation estimates as a function of the target orientation made by an individual subject. The solid line is the bias of the average estimate, computed using a sliding window. The right column shows the normalized FI as a function of orientation, computed from the response data. The three rows correspond to the baseline (noise surround), and two oriented surround conditions, respectively. See Fig. 3.1 for the same plot for the combined subject.

APPENDIX B

SUPPLEMENTARY FIGURES FOR CHAPTER 4

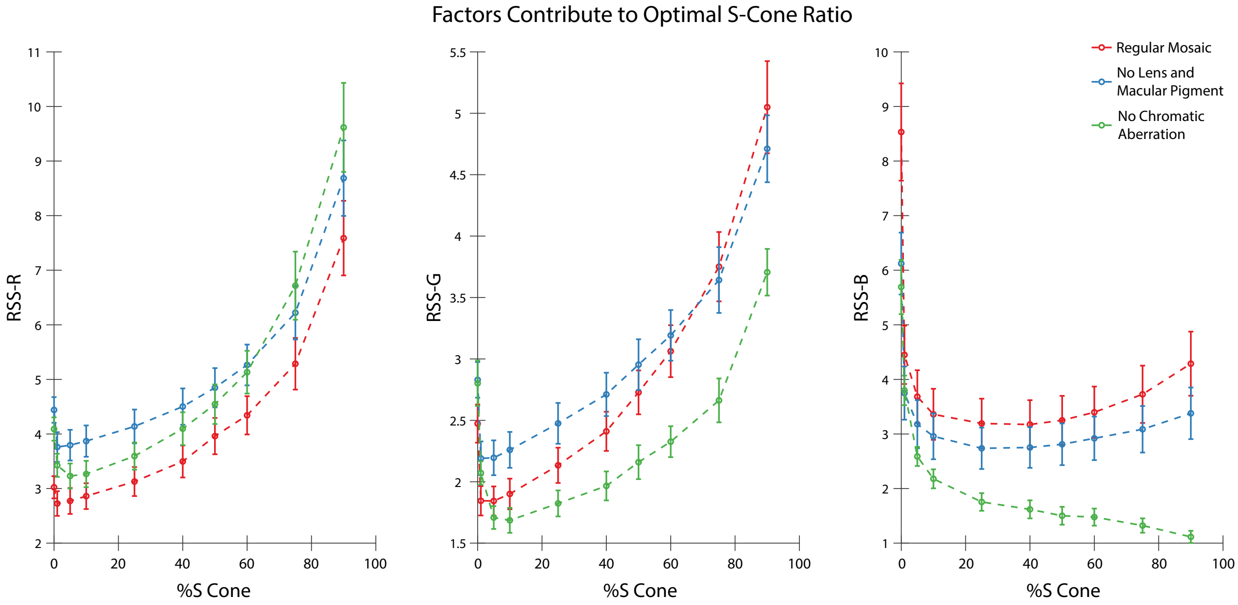

Figure B.1: Factors that contribute to optimal S Cone proportion. Average reconstruction error as a function of S-cone proportion, computed as RSS of pixel values for the R- (left), G- (middle), and B-planes (right) respectively. Under typical conditions (red), a low S-cone ratio is optimal for all three planes. Removing lens pigment and macular pigment from the simulations (blue) increases the SNR of the S cones by increasing their average quantum catch, but has little effect on the optimal S-cone proportion for any of the image planes. Correcting chromatic aberration (green) while retaining lens pigment and macular pigment greatly improves the information provided by the S cones for the B-plane, but not for the R- and G- planes. Error bars indicate ± 1 SEM.

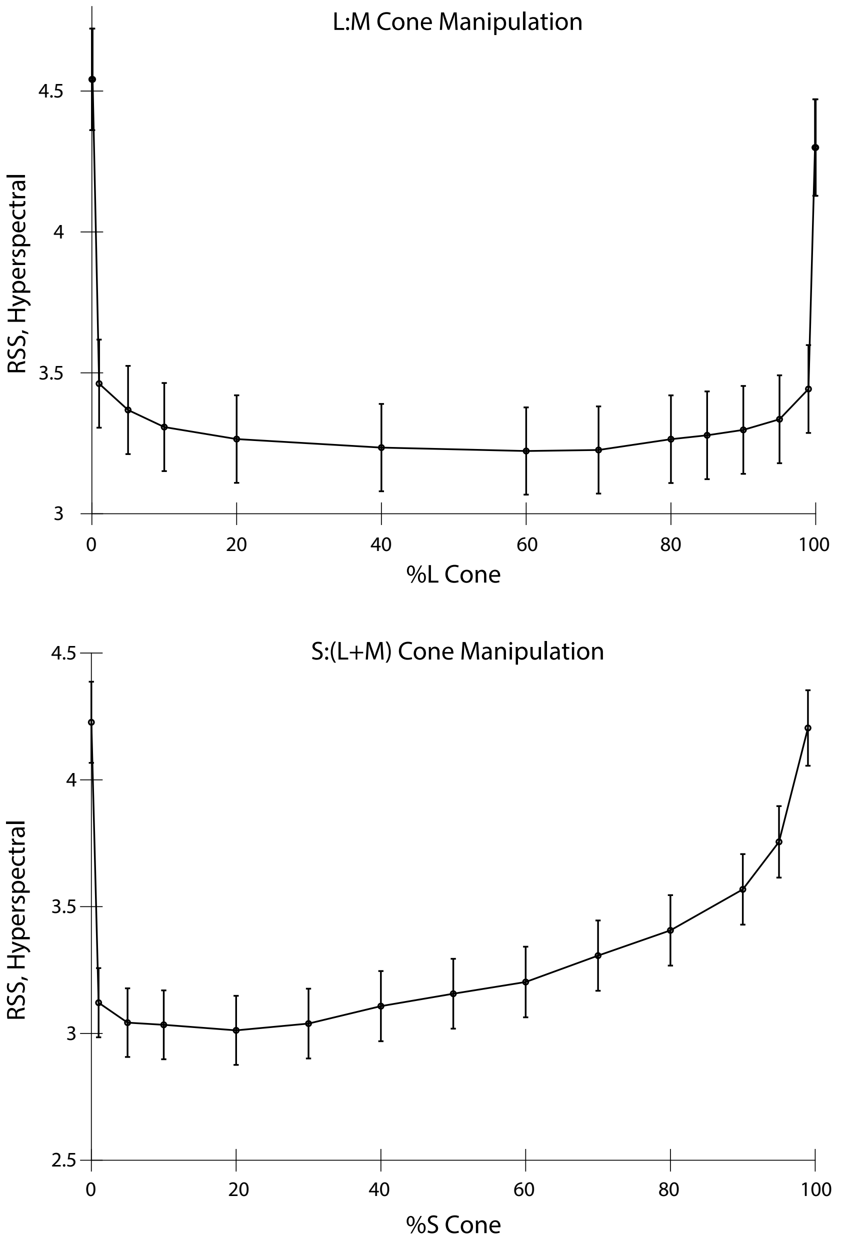

Figure B.2: Effect of the allocation of retinal cone types on reconstruction of hyperspectral images. Average reconstruction error as a function of L-cone proportion (top) and S-cone proportion (bottom), computed as RSS of pixel values over space and wavelength, for a set of evaluation hyperspectral images of size of 18*18 and 15 uniform wavelength sample between 420 nm and 700 nm (see Methods). Error bars indicate ± 1 SEM. The results corroborated our main conclusion obtained with RGB images, shown in Fig. 4.4.

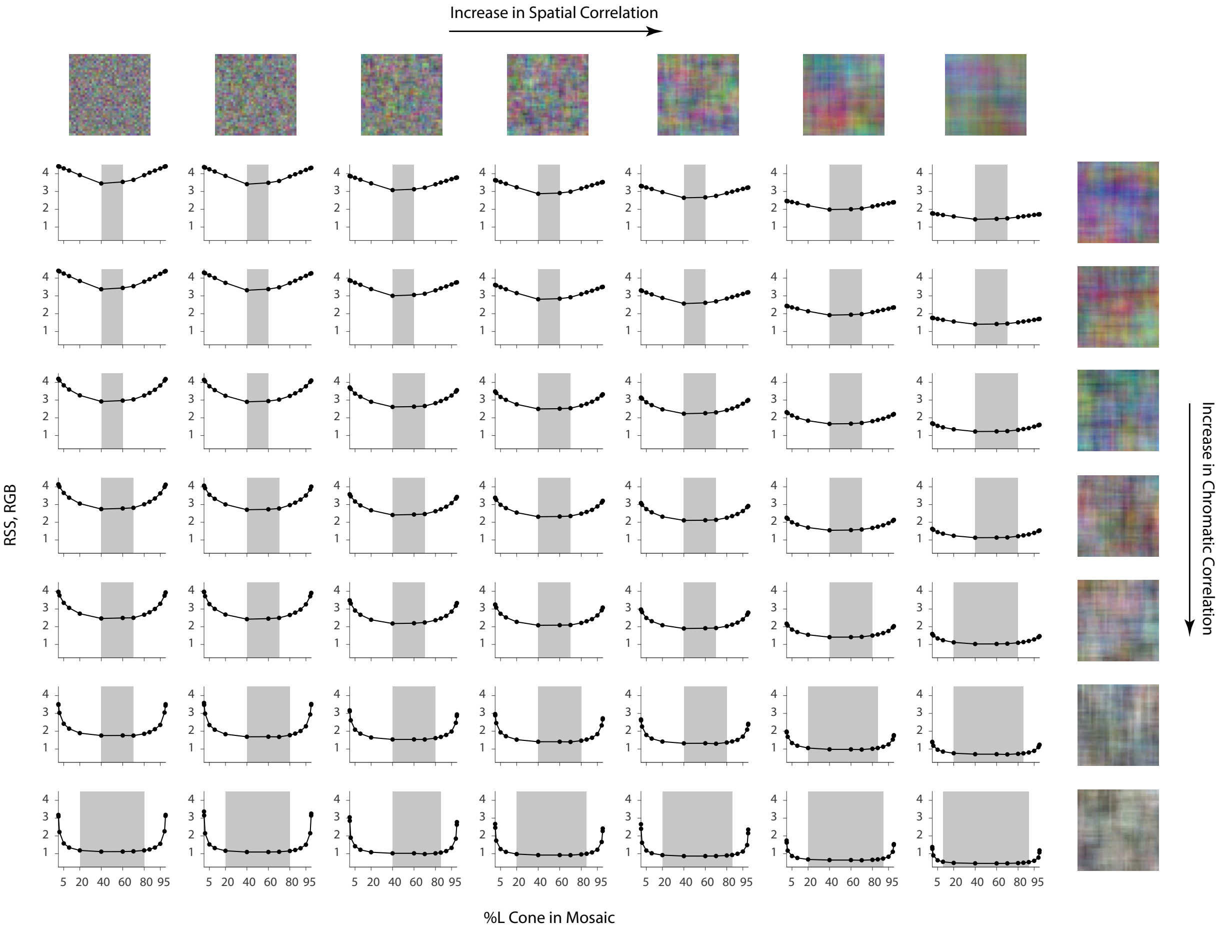

Figure B.3: Effect of spatial and chromatic correlation on the optimal allocation of photoreceptors. Same as Fig. 4.5 but with matched y-axis to highlight the overall magnitude of errors across the different conditions. Average image reconstruction error from a half-degree square foveal mosaic on different sets of synthetic images, computed as root sum of squares (RSS) distance in the RGB pixel space, as a function of %L cone (L:M cone ratio) of the mosaic The shaded areas represent %L values that correspond to RSS values within a +0.1 RSS margin from the optimal (minimum RSS) point.

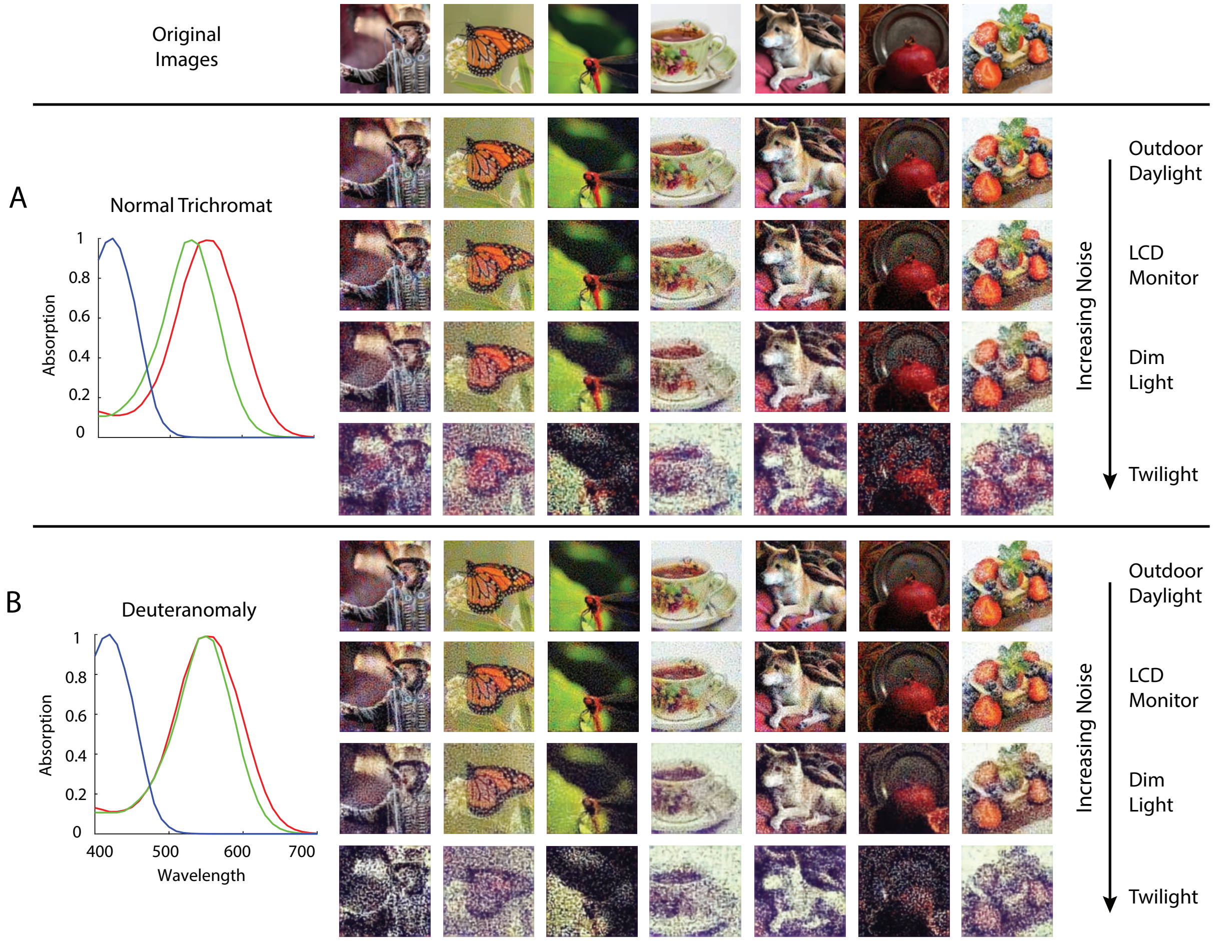

Figure B.4: Reconstruction with a weak prior across SNR levels. Image reconstructions for a set of example images in the evaluation set from 1-degree, foveal **A)** normal trichromatic and **B)** deuteranomalous trichromatic mosaics at five different overall light intensity levels that lead to different Poisson signal-to-noise ratios in the cone excitations. Same as Fig. 4.7, but to highlight the effect of noise and prior, the prior weight was set to a much lower level ($\lambda = 0.001$) than the optimal value ($\lambda = 0.05$).

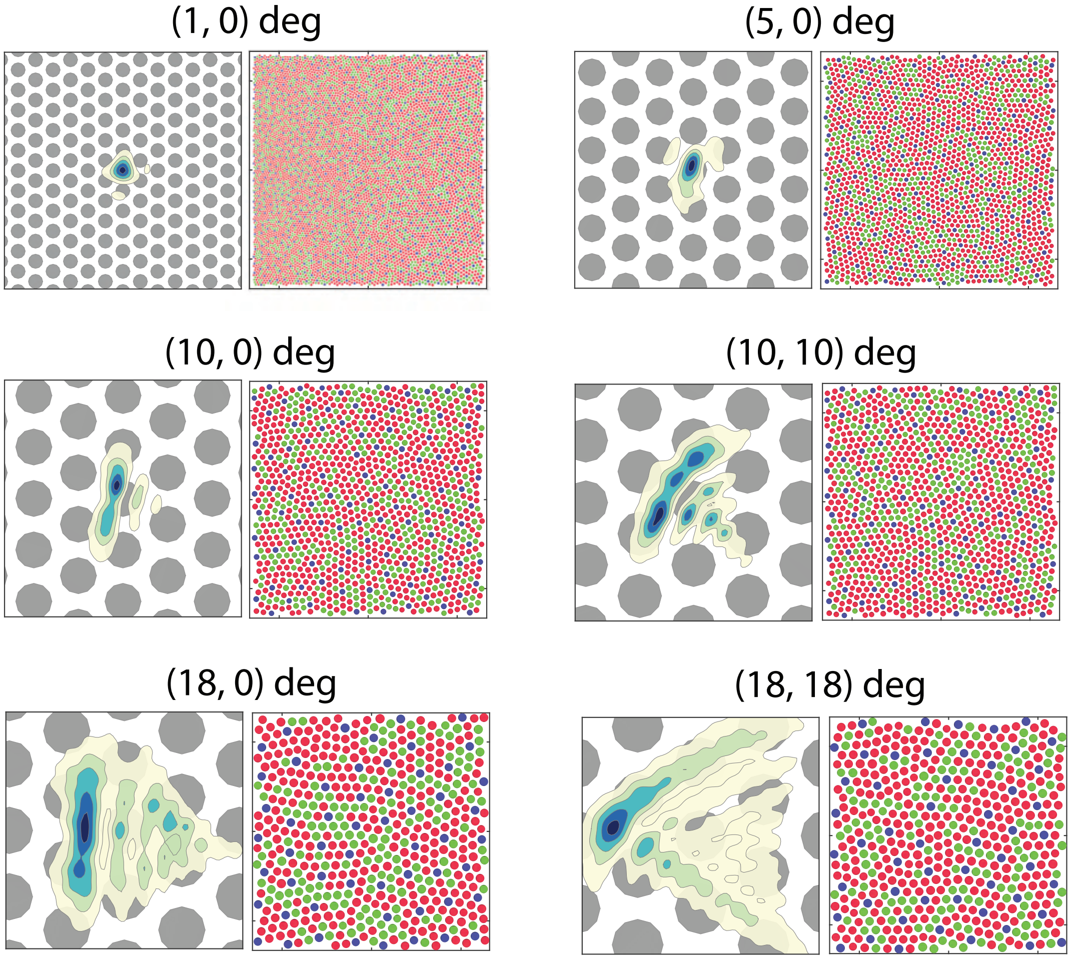

Figure B.5: Optics and cone mosaic at different retinal eccentricities. Enlarged view of the top panels of Fig. 4.8. The coordinates at the top of each pair indicate the horizontal and vertical eccentricity of the retinal patch. The left image of each pair shows a contour plot of the point-spread function relative to an expanded view of the cone mosaic, while the right image of each pair shows the full 1-degree mosaic used in the simulation.

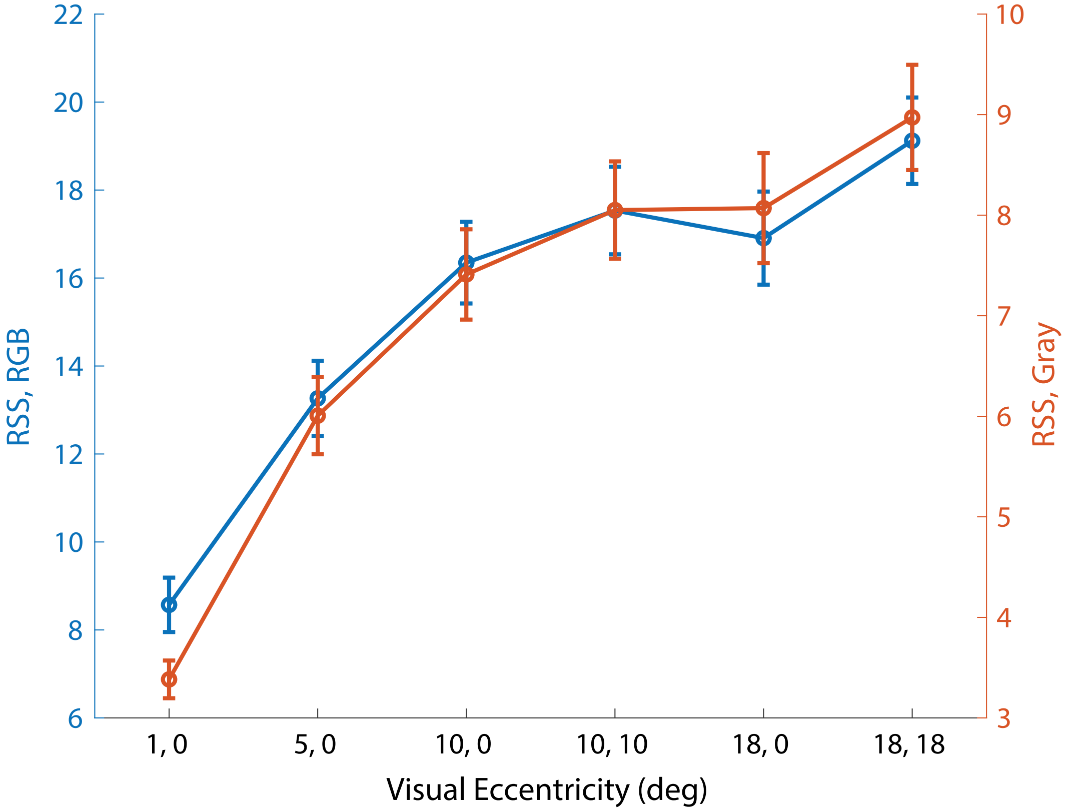

Figure B.6: Reconstruction error at different visual eccentricities. Average image reconstruction error, computed as RSS of pixel values for both the RGB images (left y-axis), and corresponding grayscale images to measure the spatial error, define as the first PC based on a PCA analysis of our image dataset ($0.57R + 0.59G + 0.56B$), right y-axis), as a function of the visual eccentricity location of a 1-deg retinal mosaic. Error bars indicate ± 1 SEM.

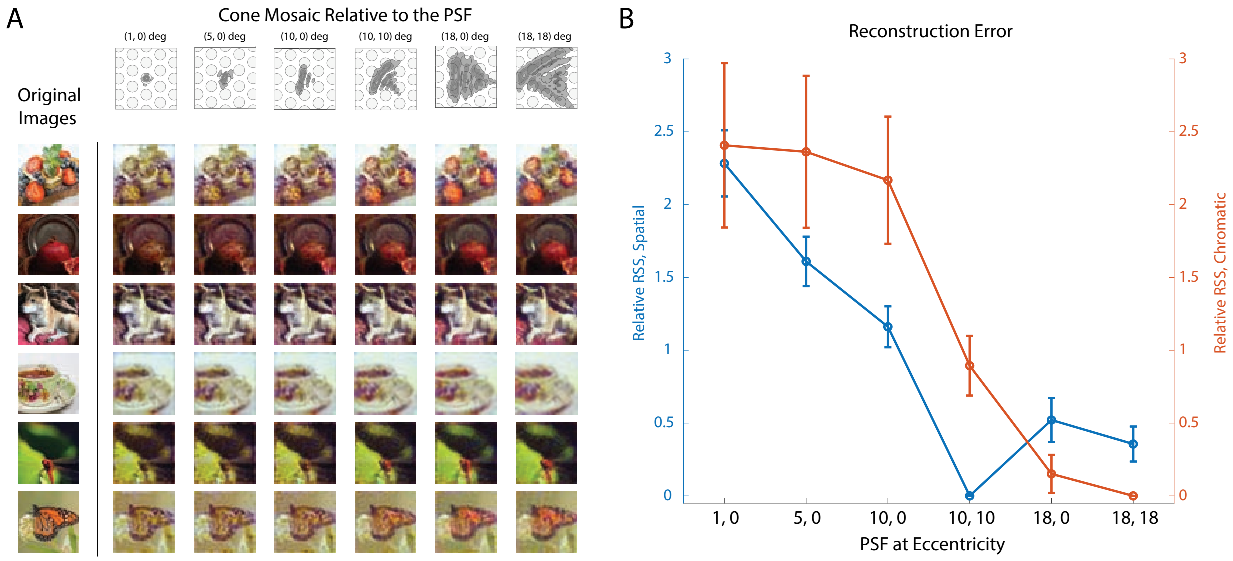

Figure B.7: Image reconstruction with different point spread functions. **A)** Image reconstructions for a set of example images in the evaluation set from 1-degree patches of mosaic at (10, 10) degree eccentricity, but with PSFs sampled from different visual eccentricities as indicated by the top panel. **B)** The average differential reconstruction error (i.e., difference in RSS compared to the lowest value obtained among the simulations) as a function of the eccentricity of the PSFs used. Error bars represent ± 1 SEM. To separate the spatial and chromatic error, we perform a PCA analysis on the RGB images. The RSS along the first PC $(0.57R + 0.59G + 0.56B)$ corresponds to the spatial error (left axis), while the RSS along the second and third PCs $(0.76R - 0.13G - 0.64B; -0.31R + 0.80G - 0.52B)$ quantify the chromatic error (right axis). With the range of PSFs in our simulation, the minimal spatial error is obtained with the PSF at (10, 10) deg (i.e., the PSF that matched to the mosaic), and the minimal chromatic error is obtained with the largest PSF, corresponding to (18, 18) deg.

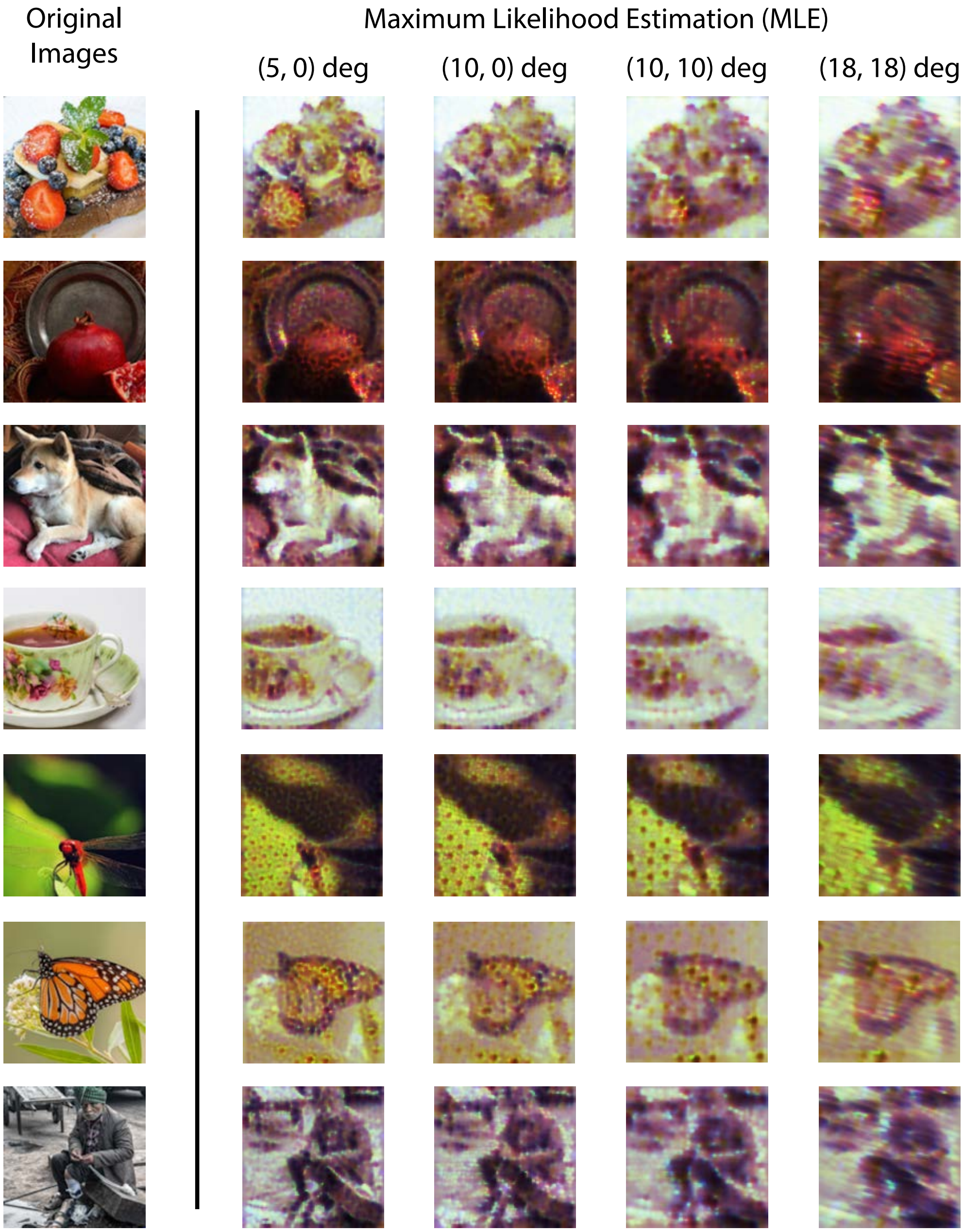

Figure B.8: Image reconstruction at peripheral eccentricities with Maximum Likelihood Estimation (MLE). Image reconstructions obtained using maximum likelihood estimation for a few example images in the evaluation set from 1-degree patches of mosaic at different retinal eccentricities, as indicated at the top of each column. Note that simulation of cone excitation noise is turned off for these reconstructions. Note also that the MLE reconstructions are not unique (see Fig. 4.3). The MLE reconstructions shown here were chosen arbitrarily as the ones converged upon by our particular numerical search algorithm.

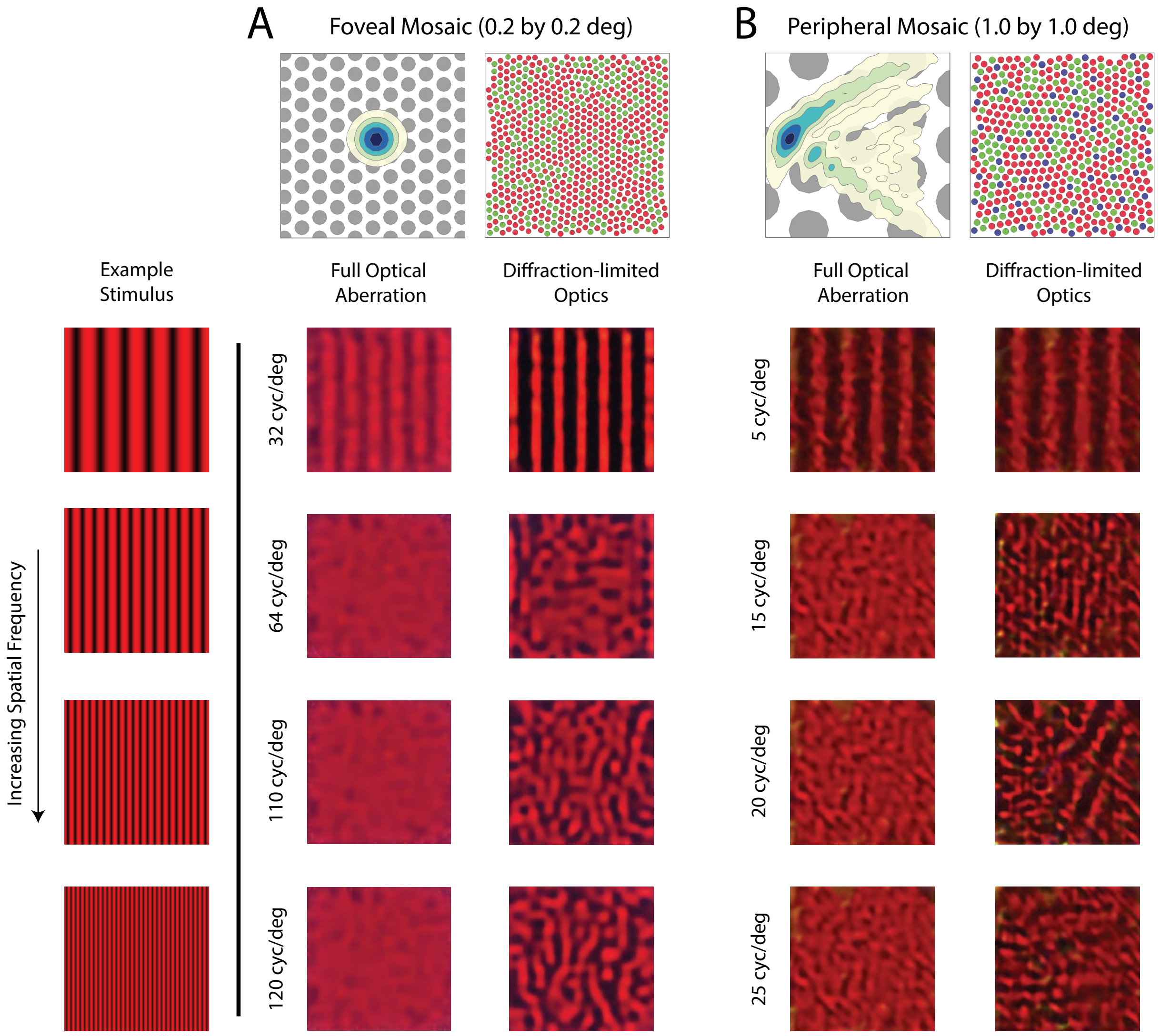

Figure B.9: Reconstruction of chromatic grating stimuli with/without optical aberrations. Image reconstruction of chromatic grating stimuli with increasing spatial frequency from **A)** a 0.2-deg foveal mosaic and **B)** a 1-deg peripheral mosaic at (18, 18) degrees retinal eccentricity with full optical aberrations (left columns) and with diffraction-limited optics (right columns). The top left images show a contour plot of the point-spread function relative to an expanded view of the cone mosaic, while the top right images show the full mosaic. Images were modulations of the red channel of the simulated monitor, to mimic the 633 nm laser used in the interferometric experiments. The exact frequency of the stimuli being used for each condition is as denoted in the figure. Note that the mottle observed in the reconstructions with full optical aberrations at high spatial frequencies match the reconstruction of a uniform field of saturated red stimulus.

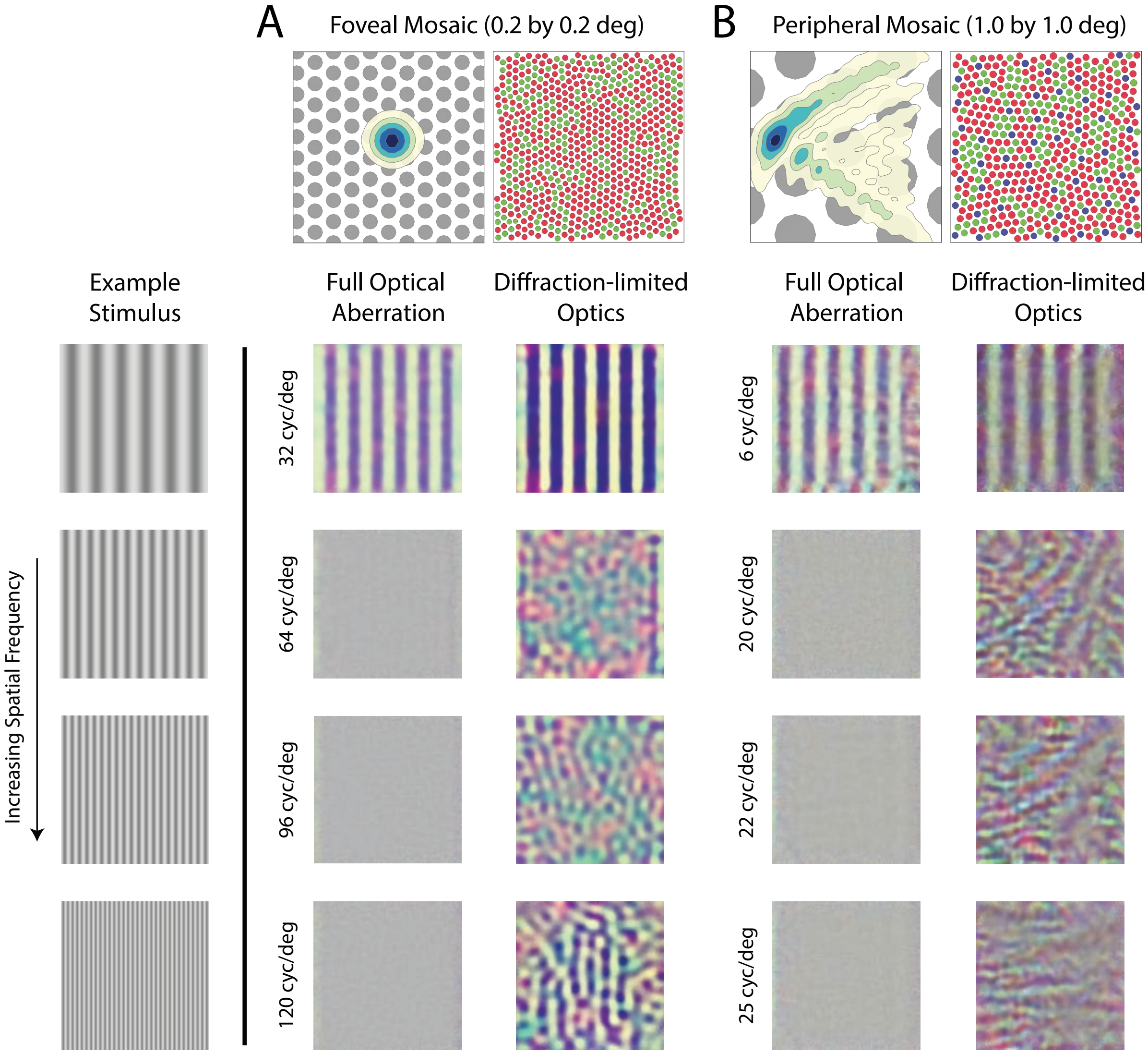
A
Foveal Mosaic (0.2 by 0.2 deg)
B
Peripheral Mosaic (1.0 by 1.0 deg)
Example Stimulus
Full Optical Aberration
Diffraction-limited Optics
Full Optical Aberration
Diffraction-limited Optics
Increasing Spatial Frequency
32 cyc/deg
64 cyc/deg
96 cyc/deg
120 cyc/deg
6 cyc/deg
20 cyc/deg
22 cyc/deg
25 cyc/deg

Figure B.10: Reconstruction of achromatic grating stimuli with/without optical aberrations. Image reconstruction of achromatic grating stimuli with increasing spatial frequency from **A)** a 0.2-deg foveal mosaic and **B)** a 1-deg peripheral mosaic at (18, 18) degrees retinal eccentricity with full optical aberrations (left columns) and with diffraction-limited optics (right columns). The top left images show a contour plot of the point-spread function relative to an expanded view of the cone mosaic, while the top right images show the full mosaic. The exact frequency of the stimuli being used for each condition is as denoted in the figure. The reconstruction shows similar spatial aliasing as in Fig. 4.9 and Fig. A.9, but shows an additional pattern of chromatic aliasing that arises because of the interleaved sampling by a mosaic of different cone types (Williams et al., 1991; Brainard et al., 2008). Whether such chromatic aliasing would actually be observed if a subject viewed achromatic gratings under diffraction-limited conditions is to our knowledge, an open question.

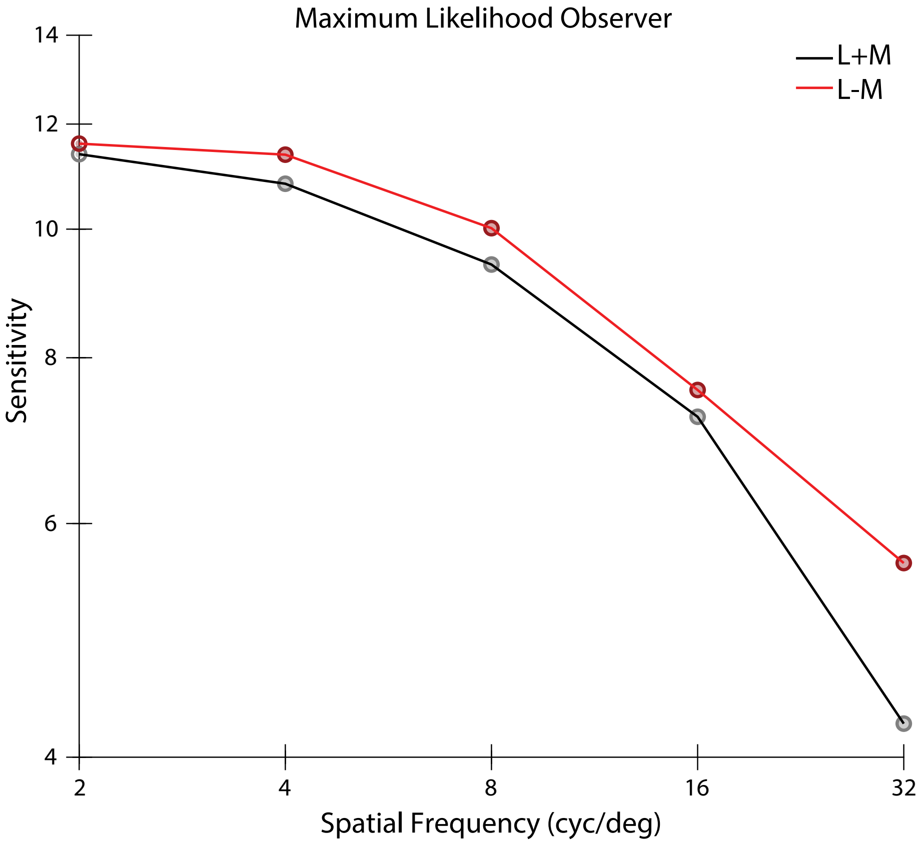

Figure B.11: Contrast sensitivity function of an MLE reconstruction observer. Contrast sensitivity, defined as the inverse of threshold contrast, for an image reconstruction-based observer without the prior term ($\lambda = 0$) as a function of the spatial frequency of stimulus in either L+M direction (black) and L-M direction (red) with equal RMS cone contrast. Note that the MLE reconstructions are not unique (Fig. 4.3). In the computations whose results are shown here, the MLE reconstructions were chosen arbitrarily as the ones converged upon by our particular numerical search algorithm.

BIBLIOGRAPHY

Larry F Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Tinku Acharya and Ajoy K Ray. *Image processing: principles and applications*. John Wiley & Sons, 2005.

Duane G Albrecht and David B Hamilton. Striate cortex of monkey and cat: contrast response function. *Journal of neurophysiology*, 48(1):217–237, 1982.

Alexander G Anderson, Kavitha Ratnam, Austin Roorda, and Bruno A Olshausen. High-acuity vision from retinal image motion. *Journal of vision*, 20(7):34–34, 2020.

Stephen J Anderson, Kathy T Mullen, and Robert F Hess. Human peripheral spatial resolution for achromatic and chromatic stimuli: limits imposed by optical and retinal factors. *The Journal of physiology*, 442(1):47–64, 1991.

Juan M Angueyra and Fred Rieke. Origin and effect of phototransduction noise in primate cone photoreceptors. *Nature neuroscience*, 16(11):1692–1700, 2013.

Stuart Appelle. Perception and discrimination as a function of stimulus orientation: the" oblique effect" in man and animals. *Psychological bulletin*, 78(4):266, 1972.

Zoe Ashwood, Aditi Jha, and Jonathan W Pillow. Dynamic inverse reinforcement learning for characterizing animal behavior. *Advances in Neural Information Processing Systems*, 35:29663–29676, 2022.

Joseph J Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

Joseph J Atick and A Norman Redlich. What does the retina know about natural scenes? *Neural computation*, 4(2):196–210, 1992.

Joseph J Atick, Zhaoping Li, and A Norman Redlich. Understanding retinal color coding from first principles. *Neural computation*, 4(4):559–572, 1992.

Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.

Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Martin S Banks, Wilson S Geisler, and Patrick J Bennett. The physical limits of grating visibility. *Vision research*, 27(11):1915–1924, 1987.

Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.

Andrea Benucci, Aman B Saleem, and Matteo Carandini. Adaptation maintains population homeostasis in primary visual cortex. *Nature neuroscience*, 16(6):724–729, 2013.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Marlene D Berke, Robert Walter-Terrill, Julian Jara-Ettinger, and Brian J Scholl. Flexible goals require that inflexible perceptual systems produce veridical representations: Implications for realism as revealed by evolutionary simulations. *Cognitive Science*, 46(10):e13195, 2022.

Mark R Blakemore and Robert J Snowden. The effect of contrast upon perceived speed: a general phenomenon? *Perception*, 28(1):33–48, 1999.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.

Bart G Borghuis, Charles P Ratliff, Robert G Smith, Peter Sterling, and Vijay Balasubramanian. Design of a neuronal array. *Journal of Neuroscience*, 28(12):3178–3189, 2008.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2018.

Jenny Bosten. The known unknowns of anomalous trichromacy. *Current Opinion in Behavioral Sciences*, 30:228–237, 2019.

Jeffrey S Bowers and Colin J Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389, 2012.

Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

David H Brainard. Color and the cone mosaic. *Annual Review of Vision Science*, 1:519–546, 2015.

David H Brainard. Color, pattern, and the retinal cone mosaic. *Current Opinion in Behavioral Sciences*, 30:41–47, 2019.

David H Brainard, David R Williams, and Heidi Hofer. Trichromatic reconstruction from the interleaved cone mosaic: Bayesian model and the color appearance of small spots. *Journal of vision*, 8(5):15–15, 2008.

Romain Brette. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 42: e215, 2019.

Hans Brettel, Françoise Viénot, and John D Mollon. Computerized simulation of color appearance for dichromats. *Josa a*, 14(10):2647–2655, 1997.

K. Britten, M. Shadlen, W. Newsome, and J.A. Movshon. Responses of neurons in macaque MT to stochastic motion signals. *Visual Neuroscience*, 10:1157–1169, 1993.

N. Brunel and J.P. Nadal. Mutual information, Fisher information, and population coding. *Neural Computation*, 10(7):1731–1757, Oct 1998.

Yoram Burak, Uri Rokni, Markus Meister, and Haim Sompolinsky. Bayesian model of dynamic image stabilization in the visual system. *Proceedings of the National Academy of Sciences*, 107 (45):19525–19530, 2010.

Johannes Burge. Image-computable ideal observers for tasks with natural stimuli. *Annual Review of Vision Science*, 6:491–517, 2020.

Johannes Burge and Wilson S Geisler. Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, 108(40):16849–16854, 2011.

Johannes Burge and Wilson S Geisler. Optimal disparity estimation in natural stereo images. *Journal of vision*, 14(2):1–1, 2014.

Johannes Burge and Wilson S Geisler. Optimal speed estimation in natural image movies predicts human performance. *Nature communications*, 6(1):1–11, 2015.

Johannes Burge and Priyank Jaini. Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS computational biology*, 13(2):e1005281, 2017.

Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Physical Review E*, 75(5):056707, 2007.

Lanya T Cai, Venkatesh Krishna, Tim C Hladnik, Nicholas C Guilbeault, Scott A Juntti, Tod R Thiele, Aristides B Arrenberg, and Emily A Cooper. Visual statistics of aquatic environments in the natural habitats of zebrafish. *Journal of Vision*, 20(11):433–433, 2020.

Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

Kenneth R Castleman. *Digital image processing*. Prentice Hall Press, 1996.

Matthew S Caywood, Benjamin Willmore, and David J Tolhurst. Independent components of color natural scenes resemble v1 neurons in their spatial and color tuning. *Journal of Neurophysiology*, 91(6):2859–2873, 2004.

Ayan Chakrabarti and Todd Zickler. Statistics of real-world hyperspectral images. In *CVPR 2011*, pages 193–200. IEEE, 2011.

Rebecca A Champion and Paul A Warren. Contrast effects on speed perception for linear and radial motion. *Vision research*, 140:66–72, 2017.

Hyun Sung Chang, Yair Weiss, and William T Freeman. Informative sensing of natural images. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 3025–3028. IEEE, 2009.

Alex Chaparro, C FIII Stromeyer, EP Huang, RE Kronauer, and Rhea T Eskew. Colour is what the eye sees best. *Nature*, 361(6410):348–350, 1993.

Samuel J Cheyette and Steven T Piantadosi. A unified account of numerosity perception. *Nature Human Behaviour*, 4(12):1265–1272, 2020.

Benjamin M Chin and Johannes Burge. Predicting the partition of behavioral variability in speed perception with naturalistic stimuli. *Journal of Neuroscience*, 40(4):864–879, 2020.

Colin WG Clifford, Anna Ma Wyatt, Derek H Arnold, Stuart T Smith, and Peter Wenderoth. Orthogonal adaptation improves orientation discrimination. *Vision research*, 41(2):151–159, 2001.

Colin WG Clifford, Michael A Webster, Garrett B Stanley, Alan A Stocker, Adam Kohn, Tatyana O Sharpee, and Odelia Schwartz. Visual adaptation: Neural, psychological and computational aspects. *Vision research*, 47(25):3125–3131, 2007.

Ruben Coen-Cagli, Adam Kohn, and Odelia Schwartz. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11):1648–1655, 2015.

Nancy J Coletta and David R Williams. Psychophysical estimate of extrafoveal cone spacing. *JOSA*

A, 4(8):1503–1513, 1987.

David M Coppola, Harriett R Purves, Allison N McCoy, and Dale Purves. The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences*, 95(7): 4002–4006, 1998.

Nicolas P Cottaris, Haomiao Jiang, Xiaomao Ding, Brian A Wandell, and David H Brainard. A computational-observer model of spatial contrast sensitivity: Effects of wave-front-based optics, cone-mosaic structure, and inference engine. *Journal of vision*, 19(4):8–8, 2019.

Nicolas P Cottaris, Brian A Wandell, Fred Rieke, and David H Brainard. A computational observer model of spatial contrast sensitivity: Effects of photocurrent encoding, fixational eye movements, and inference engine. *Journal of vision*, 20(7):17–17, 2020.

Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990.

Yang Dan, Joseph J Atick, and R Clay Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of neuroscience*, 16(10): 3351–3362, 1996.

Mark A Davenport, Marco F Duarte, Yonina C Eldar, and Gitta Kutyniok. Introduction to compressed sensing., 2012.

Karen D Davila and Wilson S Geisler. The relative contributions of pre-neural and neural factors to areal summation in the fovea. *Vision research*, 31(7-8):1369–1380, 1991.

Bart De Bruyn and Guy A Orban. Human velocity and direction discrimination measured with random dot patterns. *Vision research*, 28(12):1323–1335, 1988.

Stanislas Dehaene and Jacques Mehler. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29, 1992.

NG Deriugin. The power spectrum and the correlation function of the television signal. *Telecommunications*, 1(7):1–12, 1956.

Dawei W Dong and Joseph J Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345–358, 1995.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Valentin Dragoi, Jitendra Sharma, and Mriganka Sur. Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, 28(1):287–298, 2000.

Valentin Dragoi, Casto Rivadulla, and Mriganka Sur. Foci of orientation plasticity in visual cortex. *Nature*, 411(6833):80–86, 2001.

Ron O Dror, Alan S Willsky, and Edward H Adelson. Statistical characterization of real-world illumination. *Journal of Vision*, 4(9):11–11, 2004.

Lyndon R Duong, David Lipshutz, David J Heeger, Dmitri B Chklovskii, and Eero P Simoncelli. Statistical whitening of neural populations with gain-modulating interneurons. *arXiv preprint arXiv:2301.11955*, 2023.

Vasha DuTell, Agostino Gibaldi, Giulia Focarelli, Bruno Olshausen, and Marty Banks. The spatiotemporal power spectrum of natural human vision. *Journal of Vision*, 20(11):1661–1661, 2020.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Kjell Engström. Cone types and cone arrangement in the retina of some cyprinids. *Acta Zoologica*, 41(3):277–295, 1960.

Adrienne L Fairhall, Geoffrey D Lewen, William Bialek, and Robert R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

Fang Fang, Scott O Murray, Daniel Kersten, and Sheng He. Orientation-tuned fmri adaptation in human visual cortex. *Journal of neurophysiology*, 94(6):4188–4195, 2005.

Gustav Theodor Fechner. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel, 1860.

Gustav Theodor Fechner, Davis H Howes, and Edwin Garrigues Boring. *Elements of psychophysics*, volume 1. Holt, Rinehart and Winston New York, 1966.

Gidon Felsen, Jon Touryan, and Yang Dan. Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli. *Network: Computation in Neural Systems*, 16(2-3):139–149, 2005.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

Greg D Field, Jeffrey L Gauthier, Alexander Sher, Martin Greschner, Timothy A Machado, Lauren H Jepson, Jonathon Shlens, Deborah E Gunning, Keith Mathieson, Wladyslaw Dabrowski, et al. Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, 467(7316): 673–677, 2010.

Jason Fischer and David Whitney. Serial dependence in visual perception. *Nature neuroscience*, 17 (5):738–743, 2014.

Peter Foldiak and DM Endres. Sparse coding. *Scholarpedia*, 2008.

Matthias Fritsche, Eelke Spaak, and Floris P De Lange. A bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. *Elife*, 9:e55389, 2020.

D. Ganguli and E.P. Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. In *Adv. Neural Information Processing Systems 23*, volume 23, pages 658–666, Cambridge, MA, December 2010. MIT Press.

Deep Ganguli and Eero P Simoncelli. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural computation*, 26(10):2103–2134, 2014.

Deep Ganguli and Eero P Simoncelli. Neural and perceptual signatures of efficient sensory coding. *arXiv preprint arXiv:1603.00058*, 2016.

Patrick Garrigan, Charles P Ratliff, Jennifer M Klein, Peter Sterling, David H Brainard, and Vijay Balasubramanian. Design of a trichromatic cone array. *PLoS computational biology*, 6(2): e1000677, 2010.

Wilson S Geisler. Sequential ideal-observer analysis of visual discriminations. *Psychological review*, 96(2):267, 1989.

Wilson S Geisler. Contributions of ideal observer theory to vision research. *Vision research*, 51(7): 771–781, 2011.

Wilson S Geisler. Psychometric functions of uncertain template matching observers. *Journal of vision*, 18(2):1–1, 2018.

James J Gibson and Minnie Radner. Adaptation, after-effect and contrast in the perception of tilted lines. i. quantitative studies. *Journal of experimental psychology*, 20(5):453, 1937.

Sonya Giridhar, Brent Doiron, and Nathaniel N Urban. Timescale-dependent shaping of correlation by olfactory bulb lateral inhibition. *Proceedings of the National Academy of Sciences*, 108(14): 5843–5848, 2011.

Ahna R Girshick, Michael S Landy, and Eero P Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14(7):926–932, 2011.

Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, and Lei Zhang. Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1823–1831, 2015.

Yong Gu, Christopher R Fetsch, Babatunde Adeyemo, Gregory C DeAngelis, and Dora E Angelaki. Decoding of mstd population activity accounts for variations in the precision of heading perception. *Neuron*, 66(4):596–609, 2010.

Onur G Guleryuz. Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising-part i: theory. *IEEE Transactions on image processing*, 15(3): 539–554, 2006.

Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

Thorsten Hansen, Lars Pracejus, and Karl R Gegenfurtner. Color perception in the intermediate periphery of the visual field. *Journal of vision*, 9(4):26–26, 2009.

Wolf M Harmening, William S Tuten, Austin Roorda, and Lawrence C Sincich. Mapping the perceptual grain of the human retina. *Journal of Neuroscience*, 34(16):5667–5677, 2014.

J.H. Hedges, A.A. Stocker, and E.P. Simoncelli. Optimal inference explains the perceptual coherence of visual motion stimuli. *Journal of Vision*, 11(6):1–16, May 2011.

Hilary W Heuer and Kenneth H Britten. Contrast dependence of response normalization in area mt of the rhesus macaque. *Journal of neurophysiology*, 88(6):3398–3408, 2002.

Heidi Hofer, Joseph Carroll, Jay Neitz, Maureen Neitz, and David R Williams. Organization of the human trichromatic cone mosaic. *Journal of Neuroscience*, 25(42):9669–9679, 2005.

Mark S Horswill and Annaliese M Plooy. Reducing contrast makes speeds in a video-based driving simulator harder to discriminate as well as making them appear slower. *Perception*, 37(8):1269–1275, 2008.

Xin Huang and Stephen G. Lisberger. Noise correlations in cortical area mt and their potential impact on trial-by-trial variation in the direction and speed of smooth-pursuit eye movements. *Journal of Neurophysiology*, 101(6):3012–3030, 2009. doi: 10.1152/jn.00010.2009.

David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

F. Hürlimann, D. Kiper, and M. Carandini. Testing the Bayesian model of perceived speed. *Vision Research*, 42:2253–2257, 2002.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Arvind Iyer and Johannes Burge. The statistics of how natural images drive the responses of neurons. *Journal of Vision*, 19(13):4–4, 2019.

Joseph Jastrow. Studies from the university of wisconsin: on the judgment of angles and positions of lines. *The American Journal of Psychology*, 5(2):214–248, 1892.

Haomiao Jiang, Joyce Farrell, and Brian Wandell. A spectral estimation theory for color appearance matching. *Electronic Imaging*, 2016(20):1–4, 2016.

Matjaž Jogan and Alan A Stocker. Signal integration in human visual speed perception. *Journal of Neuroscience*, 35(25):9381–9390, 2015.

Robert Evan Johnson, Scott Linderman, Thomas Panier, Caroline Lei Wee, Erin Song, Kristian Joseph Herrera, Andrew Miller, and Florian Engert. Probabilistic models of larval zebrafish behavior reveal structure on many scales. *Current Biology*, 30(1):70–82, 2020.

Matt Jones and Bradley C Love. Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and brain sciences*, 34(4):169, 2011.

Na Young Jun, Greg D Field, and John Pearson. Scene statistics and noise determine the relative arrangement of receptive field mosaics. *Proceedings of the National Academy of Sciences*, 118(39): e2105115118, 2021.

Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021.

Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring fisher information accurately in correlated neural populations. *PLoS computational biology*, 11(6): e1004218, 2015.

Yan Karklin and Eero Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. *Advances in neural information processing systems*, 24, 2011.

D Knill D Kersten and A Yuille. Introduction: A bayesian formulation of visual perception. *Perception as Bayesian inference*, pages 1–21, 1996.

Seha Kim and Johannes Burge. Natural scene statistics predict how humans pool information across space in surface tilt estimation. *PLoS Computational Biology*, 16(6):e1007947, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sue Ann Koay, Adam S Charles, Stephan Y Thiberge, Carlos D Brody, and David W Tank. Sequential and efficient neural-population coding of complex task information. *Neuron*, 110(2):328–349, 2022.

Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual review of neuroscience*, 39:237–256, 2016.

B. Krekelberg, R.J.A. van Wezel, and T. Albright. Interactions between speed and contrast tuning in the middle temporal area: Implications for the neural code for speed. *Journal of Neuroscience*, 26:8988–8998, August 2006.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2012.

Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in visual path integration. *Neuron*, 99(1):194–206, 2018.

Michael F Land and Dan-Eric Nilsson. *Animal eyes.* OUP Oxford, 2012.

Michael F Land and Daniel C Osorio. Colour vision: colouring the dark. *Current Biology*, 13(3): R83–R85, 2003.

Richard D Lange, Sabyasachi Shivkumar, Ankani Chattoraj, and Ralf M Haefner. Bayesian encoding and decoding as distinct perspectives on neural coding. *BioRxiv*, pages 2020–10, 2020.

Simon Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912, 1981.

Quoc Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Ng. Ica with reconstruction cost for efficient overcomplete feature learning. *Advances in neural information processing systems*, 24, 2011.

Peter Lennie, P William Haake, and David R Williams. The design of chromatically opponent receptive fields. *Computational models of visual processing*, pages 71–82, 1991.

Anat Levin, William T Freeman, and Frédo Durand. Understanding camera trade-offs through a bayesian analysis of light field projections. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, pages 88–101. Springer, 2008.

Trisha Lian. *Vision Modeling Tools for Evaluating Next-Generation Displays.* Stanford University, 2020.

Delwin Lindsey, Lindsey Hutchinson, and Angela Brown. Unique yellow and other special colors seen by deuteranomalous trichromats. *Journal of Vision*, 20(11):1249–1249, 2020.

R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

Svein Magnussen and Tore Johnsen. Temporal aspects of spatial adaptation. a study of the tilt aftereffect. *Vision research*, 26(4):661–672, 1986.

Jeremy R Manning and David H Brainard. Optimal design of photoreceptor mosaics: why we do not see color at night. *Visual neuroscience*, 26(1):5–19, 2009.

David H Marimont and Brian A Wandell. Matching color images: the effects of axial chromatic aberration. *JOSA A*, 11(12):3113–3122, 1994.

Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. The role of fixational eye movements in visual perception. *Nature reviews neuroscience*, 5(3):229–240, 2004.

M. D. McDonnell and N. G. Stocks. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Physical Review Letters*, 101(5):058103, 2008.

Suzanne P McKee, Gerald H Silverman, and Ken Nakayama. Precise velocity discrimination despite random variations in temporal frequency and contrast. *Vision research*, 26(4):609–619, 1986.

Donald E Mitchell and Darwin W Muir. Does the tilt after-effect occur in the oblique meridian? *Vision research*, 16(6):609–613, 1976.

Koichi Miyasawa. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38(181-188):1–2, 1961.

Wiktor F Młynarski and Ann M Hermundstad. Adaptive coding for dynamic sensory inference. *Elife*, 7:e32055, 2018.

Wiktor F Młynarski and Ann M Hermundstad. Efficient and adaptive sensory codes. *Nature Neuroscience*, 24(7):998–1009, 2021.

Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *arXiv preprint arXiv:1906.05478*, 2019.

Michael Morais and Jonathan W Pillow. Power-law efficient neural codes provide general link between perceptual bias and discriminability. In *Advances in Neural Information Processing Systems 31*, pages 5071–5080. Curran Associates, Inc., 2018.

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

J Anthony Movshon and William T Newsome. Visual response properties of striate cortical neurons projecting to area mt in macaque monkeys. *Journal of Neuroscience*, 16(23):7733–7741, 1996.

Kathy T Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of physiology*, 359(1):381–400, 1985.

Kathy T Mullen and Frederick AA Kingdom. Losses in peripheral colour sensitivity predicted from "hit and miss" post-receptoral cone connections. *Vision research*, 36(13):1995–2000, 1996.

Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22 (10):1677–1686, 2019.

Sérgio MC Nascimento, Flávio P Ferreira, and David H Foster. Statistics of spatial cone-excitation ratios in natural scenes. *JOSA A*, 19(8):1484–1490, 2002.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

Alexandra Neitz, Xiaoyun Jiang, James A Kuchenbecker, Niklas Domdei, Wolf Harmening, Hongyi Yan, Jihyun Yeonan-Kim, Sara S Patterson, Maureen Neitz, Jay Neitz, et al. Effect of cone spectral topography on chromatic detection sensitivity. *JOSA A*, 37(4):A244–A254, 2020.

William T Newsome and Edmond B Pare. A selective impairment of motion perception following lesions of the middle temporal visual area (mt). *Journal of Neuroscience*, 8(6):2201–2211, 1988.

Andreas Nieder and Earl K Miller. Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37(1):149–157, 2003.

Jean-Paul Noel, Ling-Qi Zhang, Alan A. Stocker, and Dora E. Angelaki. Individuals with autism spectrum disorder have altered visual encoding capacity. *PLOS Biology*, 19(5):1–21, 05 2021. doi: 10.1371/journal.pbio.3001215.

John M Nolan, James M Stringham, Stephen Beatty, and D Max Snodderly. Spatial profile of macular pigment and its relationship to foveal architecture. *Investigative ophthalmology & visual science*, 49(5):2134–2142, 2008.

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

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

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

C.C. Pack, J.N. Hunter, and R.T. Born. Contrast dependence of suppressive influences in cortical area MT of alert macaque. *Journal of Neurophysiology*, 93:1809–1815, 2005.

Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

Steven C Panish. Velocity discrimination at constant multiples of threshold contrast. *Vision research*, 28(2):193–201, 1988.

Il Memming Park and Jonathan W Pillow. Bayesian efficient coding. *BioRxiv*, page 178418, 2017.

Nikhil Parthasarathy, Eleanor Batty, William Falcon, Thomas Rutten, Mohit Rajpal, EJ Chichilnisky, and Liam Paninski. Neural networks for efficient bayesian decoding of natural images from retinal neurons. *Advances in Neural Information Processing Systems*, 30, 2017.

Denis G Pelli. Uncertainty explains many aspects of visual contrast detection and discrimination. *JOSA A*, 2(9):1508–1532, 1985.

Megan AK Peters, Jonathan Balzer, and Ladan Shams. Smaller= denser, and the brain knows it: natural statistics of object density shape weight expectations. *PloS one*, 10(3):e0119794, 2015.

Steven T Piantadosi and Jessica F Cantlon. True numerical cognition in the wild. *Psychological science*, 28(4):462–469, 2017.

Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4):628–635, 2012.

Rafael Polania, Michael Woodford, and Christian C. Ruff. Efficient coding of subjective value. *Nature Neuroscience*, 22(1):134–142, 2019. doi: 10.1038/s41593-018-0292-0.

James Polans, Bart Jaeken, Ryan P McNabb, Pablo Artal, and Joseph A Izatt. Wide-field optical model of the human eye with asymmetrically tilted and decentered lens that reproduces measured ocular aberrations. *Optica*, 2(2):124–134, 2015.

Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40:49–70, 2000.

Javier Portilla, Vasily Strela, Martin J Wainwright, and Eero P Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image processing*, 12 (11):1338–1351, 2003.

Arthur Prat-Carrabin and Michael Woodford. Efficient coding of numbers explains decision bias and noise. *bioRxiv*, 2021. doi: 10.1101/2020.02.18.942938.

N.J. Priebe, C.R. Cassanello, and S.G. Lisberger. The neural representation of speed in macaque area MT/V5. *The Journal of Neuroscience*, 23(13):5650–5661, July 2003.

Christopher M Putnam and Pauline J Bland. Macular pigment optical density spatial distribution measured in a subject with oculocutaneous albinism. *Journal of optometry*, 7(4):241–245, 2014.

Cheng Qiu and Alan Stocker. Bayesian interpretation of artificial neural network models in perception. *Journal of Vision*, 21(9):2712–2712, 2021.

Luke Rast and Jan Drugowitsch. Adaptation properties allow identification of optimized neural codes. In *Advances in Neural Information Processing Systems*, volume 33, pages 1142–1152. Curran Associates, Inc., 2020.

Kavitha Ratnam, Niklas Domdei, Wolf M Harmening, and Austin Roorda. Benefits of retinal image motion at the limits of spatial vision. *Journal of vision*, 17(1):30–30, 2017.

D Regan and KI Beverley. Postadaptation orientation discrimination. *JOSA A*, 2(2):147–155, 1985.

Reuben Rideaux and Andrew E Welchman. But still it moves: static image statistics underlie how we see motion. *Journal of Neuroscience*, 40(12):2538–2552, 2020.

Reuben Rideaux, Rebecca K West, Dragan Rangelov, and Jason B Mattingley. Distinct early and late neural mechanisms regulate feature-specific sensory adaptation in the human visual system. *Proceedings of the National Academy of Sciences*, 120(6):e2216192120, 2023.

Bas Rokers, Jacqueline M Fulvio, Jonathan W Pillow, and Emily A Cooper. Systematic misperceptions of 3-d motion explained by bayesian inference. *Journal of vision*, 18(3):23–23, 2018.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Stefan Roth and Michael J Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1):33–50, 2007.

Suva Roy, Na Young Jun, Emily L Davis, John Pearson, and Greg D Field. Inter-mosaic coordination of retinal receptive fields. *Nature*, 592(7854):409–413, 2021.

Michele Rucci, Ramon Iovin, Martina Poletti, and Fabrizio Santini. Miniature eye movements enhance fine spatial detail. *Nature*, 447(7146):852–855, 2007.

William Albert Hugh Rushton. Visual pigments in man. *Scientific American*, 207(5):120–135, 1962.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Nicole C Rust and Alan A Stocker. Ambiguity and invariance: two fundamental challenges for visual processing. *Current opinion in neurobiology*, 20(3):382–388, 2010.

Ramkumar Sabesan, Brian P Schmidt, William S Tuten, and Austin Roorda. The elementary representation of spatial and color vision in the human retina. *Science advances*, 2(9):e1600797, 2016.

Brian P Schmidt, Alexandra E Boehm, William S Tuten, and Austin Roorda. Spatial summation of individual cones in human color vision. *Plos one*, 14(7):e0211397, 2019.

Odelia Schwartz and Eero P Simoncelli. Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8):819–825, 2001.

Odelia Schwartz, Anne Hsu, and Peter Dayan. Space and time in visual context. *Nature Reviews Neuroscience*, 8(7):522–535, 2007.

Odelia Schwartz, Terrence J Sejnowski, and Peter Dayan. Perceptual organization in the tilt illusion. *Journal of Vision*, 9(4):19–19, 2009.

G. Sclar, J. Maunsell, and P. Lennie. Coding of image contrast in central visual pathways of the macaque monkey. *Vision Research*, 30(1):1–10, 1990.

Nobutoshi Sekiguchi, David R Williams, and David H Brainard. Efficiency in detection of isoluminant and isochromatic interference fringes. *JOSA A*, 10(10):2118–2133, 1993.

Irene Senna, Cesare V. Parise, and Marc O. Ernst. Hearing in slow-motion: Humans underestimate the speed of moving sounds. *Scientific Reports*, 5(1):14054, 2015. doi: 10.1038/srep14054.

Peggy Seriès, Alan A Stocker, and Eero P Simoncelli. Is the homunculus "aware" of sensory adaptation? *Neural computation*, 21(12):3271–3304, 2009.

Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237 (4820):1317–1323, 1987.

Ron Shepard, Gergely Gidofalvi, and Scott R Brozell. The multifacet graphically contracted function method. ii. a general procedure for the parameterization of orthogonal matrices and its application to arc factors. *The Journal of Chemical Physics*, 141(6):064106, 2014.

Ron Shepard, Scott R Brozell, and Gergely Gidofalvi. The representation and parametrization of orthogonal matrices. *The Journal of Physical Chemistry A*, 119(28):7924–7939, 2015.

Eero Simoncelli. *Distributed analysis and representation of visual motion.* PhD thesis, MIT, Dept. of Electrical Engineering, Cambridge, MA, 1993.

Eero P Simoncelli. 4.7 statistical modeling of photographic images. *Handbook of Video and Image Processing*, 9, 2005.

Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings., International Conference on Image Processing*, volume 3, pages 444–447. IEEE, 1995.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Chris R Sims. Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389):652–656, 2018.

Vijay Singh, Nicolas P Cottaris, Benjamin S Heasly, David H Brainard, and Johannes Burge. Computational luminance constancy from naturalistic images. *Journal of Vision*, 18(13):19–19, 2018.

Shiva R Sinha, William Bialek, and Rob R De Ruyter Van Steveninck. Optimal local estimates of visual motion in a natural environment. *Physical review letters*, 126(1):018101, 2021.

Allan W Snyder, Simon B Laughlin, and Doekele G Stavenga. Information capacity of eyes. *Vision research*, 17(10):1163–1175, 1977.

Allan W Snyder, Terry RJ Bossomaier, and Austin Hughes. Optical image quality and the cone mosaic. *Science*, 231(4737):499–501, 1986.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Grigorios Sotiropoulos, Aaron R. Seitz, and Peggy Series. Contrast dependency and prior expectations in human speed perception. *Vision Research*, 97:16 – 23, 2014. doi: 10.1016/j.visres.2014.01.012.

A.A. Stocker. *Analog VLSI Circuits for the Perception of Visual Motion.* John Wiley & Sons Ltd., Chichester, May 2006. 242 pages. Hardcover.

A.A. Stocker and E.P. Simoncelli. Constraining a Bayesian model of human visual speed perception. In *Advances in Neural Information Processing Systems NIPS 17*, pages 1361–1368, Cambridge, MA, December 2004. MIT Press.

Alan A Stocker and Eero Simoncelli. Sensory adaptation within a bayesian framework for perception. *Advances in neural information processing systems*, 18, 2005.

Alan A Stocker and Eero P Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578, 2006.

Alan A Stocker, Najib Majaj, Chris Tailby, J Anthony Movshon, and Eero P Simoncelli. Decoding velocity from population responses in area MT of the macaque. *Journal of Vision*, 9(8):741–741, 2009.

Andrew Stockman and Lindsay T Sharpe. The spectral sensitivities of the middle-and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision research*, 40(13):1711–1737, 2000.

Leland S Stone and Peter Thompson. Human speed perception is contrast dependent. *Vision research*, 32(8):1535–1549, 1992.

Steven P Strong, Roland Koberle, Rob R De Ruyter Van Steveninck, and William Bialek. Entropy and information in neural spike trains. *Physical review letters*, 80(1):197, 1998.

R Taylor and PM Bays. Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *Journal of Neuroscience*, 2018. doi: 10.1523/JNEUROSCI.1018-18.2018.

Peter Thompson. Perceived rate of movement depends on contrast. *Vision research*, 22(3):377–380, 1982.

Qiyuan Tian, Henryk Blasinski, Steven Lansel, Haomiao Jiang, Munenori Fukunishi, Joyce E Farrell, and Brian A Wandell. Automatically designing an image processing pipeline for a five-band camera prototype using the local, linear, learned (l3) method. In *Digital Photography XI*, volume 9404, pages 18–23. SPIE, 2015.

Bosco S Tjan and Gordon E Legge. The viewpoint complexity of an object-recognition task. *Vision research*, 38(15-16):2335–2350, 1998.

Gašper Tkačik, Jason S Prentice, Vijay Balasubramanian, and Elad Schneidman. Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107(32): 14419–14424, 2010.

David J Tolhurst, Yoav Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

Momchil S Tomov, Samyukta Yagati, Agni Kumar, Wanqian Yang, and Samuel J Gershman. Discovery of hierarchical representations for efficient planning. *PLoS computational biology*, 16(4): e1007594, 2020.

Katherine EM Tregillus, Zoey J Isherwood, John E Vanston, Stephen A Engel, Donald IA MacLeod, Ichiro Kuriki, and Michael A Webster. Color compensation in anomalous trichromats assessed with fmri. *Current Biology*, 31(5):936–942, 2021.

Michel Treisman. Noise and weber's law: The discrimination of brightness and other dimensions. *Psychological review*, 71(4):314, 1964.

Joel A Tropp and Stephen J Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

Kathleen Turano and Allan Pantle. On the mechanism that encodes the movement of contrast variations: velocity discrimination. *Vision Research*, 29(2):207–221, 1989.

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3 (1):71–86, 1991.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

RS Van Bergen and JFM Jehee. Tafkap: An improved method for probabilistic decoding of cortical activity. *BioRxiv*, pages 2021–03, 2021.

Ruben S Van Bergen and Janneke FM Jehee. Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. *Journal of Neuroscience*, 39(41):8164–8176, 2019.

Ruben S Van Bergen, Wei Ji Ma, Michael S Pratte, and Janneke FM Jehee. Sensory uncertainty decoded from visual cortex predicts behavior. *Nature neuroscience*, 18(12):1728–1730, 2015.

J Hans van Hateren. Real and optimal neural images in early vision. *Nature*, 360(6399):68–70, 1992.

JH Van Hateren. Spatial, temporal and spectral pre-processing for colour vision. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 251(1330):61–68, 1993.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

V Virsu and J Rovamo. Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental brain research*, 37:475–494, 1979.

George Wald. The vertebrate eye and its adaptive radiation, 1944.

Edgar Y Walker, R James Cotton, Wei Ji Ma, and Andreas S Tolias. A neural basis of probabilistic computation in visual cortex. *Nature neuroscience*, 23(1):122–129, 2020.

Z. Wang, A.A. Stocker, and D.D. Lee. Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, Infomax and minimum L_p loss. In *Advances in Neural Information Processing Systems NIPS 25*, pages 2177–2185. MIT Press, May 2012.

Z. Wang, A.A. Stocker, and D.D. Lee. Efficient neural codes that minimize L_p reconstruction error. *Neural Computation*, 28(12):2656–2686, December 2016a. doi: 10.1162/NECO_a_00900.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhuo Wang, Alan A Stocker, and Daniel D Lee. Efficient neural codes that minimize l p reconstruction error. *Neural computation*, 28(12):2656–2686, 2016b.

Barry Wark, Brian Nils Lundstrom, and Adrienne Fairhall. Sensory adaptation. *Current opinion in neurobiology*, 17(4):423–429, 2007.

Barry Wark, Adrienne Fairhall, and Fred Rieke. Timescales of inference in visual adaptation. *Neuron*, 61(5):750–761, 2009.

Andrew B Watson. Quest+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision*, 17(3):10–10, 2017.

Michael A Webster. Adaptation and visual coding. *Journal of vision*, 11(5):3–3, 2011.

X.-X. Wei, P. Ortega, and A.A. Stocker. Perceptual adaptation: Getting ready for the future. In *Vision Science Society VSS conference*, May 2015. doi: doi:10.1167/15.12.388.

Xue-Xin Wei and Alan A Stocker. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. *Advances in neural information processing systems*, 25:1304–1312, 2012.

Xue-Xin Wei and Alan A Stocker. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.

Xue-Xin Wei and Alan A Stocker. Mutual information, fisher information, and efficient coding. *Neural computation*, 28(2):305–326, 2016.

Xue-Xin Wei and Alan A Stocker. Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 114(38):10244–10249, 2017.

Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature* *neuroscience*, 5(6):598–604, 2002.

Yair Weiss, Hyun Sung Chang, and William T Freeman. Learning compressed sensing. In *Snowbird Learning Workshop, Allerton, CA*, 2007.

A.E. Welchman, J.M. Lam, and H.H. Bülthoff. Bayesian motion estimation accounts for a surprising bias in 3D vision. *Proceeding of the National Academy of Sciences of the U.S.A.*, 105(33):12087–12092, August 2008.

David R Williams. Aliasing in human foveal vision. *Vision research*, 25(2):195–205, 1985.

David R Williams, Nobutoshi Sekiguchi, William Haake, David Brainard, and Orin Packer. The cost of trichromacy for spatial vision. *From pigments to perception: advances in understanding visual processes*, pages 11–22, 1991.

Lauren E Wool, Joanna D Crook, John B Troy, Orin S Packer, Qasim Zaidi, and Dennis M Dacey. Nonselective wiring accounts for red-green opponency in midget ganglion cells of the primate retina. *Journal of Neuroscience*, 38(6):1520–1540, 2018.

Yan Wu, Mihaela Rosca, and Timothy Lillicrap. Deep compressed sensing. In *International Conference on Machine Learning*, pages 6850–6860. PMLR, 2019.

John I Yellott Jr. Spectral consequences of photoreceptor sampling in the rhesus retina. *Science*, 221(4608):382–385, 1983.

Thomas E Yerxa, Eric Kee, Michael R DeWeese, and Emily A Cooper. Efficient sensory coding of multidimensional stimuli. *PLoS computational biology*, 16(9):e1008146, 2020.

S.M. Zeki. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *Journal of Physiology, London*, 236:549–573, 1974.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Ling-Qi Zhang and Alan A Stocker. Prior expectations in visual speed perception predict encoding characteristics of neurons in area MT. *Journal of Neuroscience*, 42(14):2951–2962, 2022.

Ling-Qi Zhang, Nicolas P Cottaris, and David H Brainard. An image reconstruction framework for characterizing initial visual encoding. *eLife*, 11:e71132, 2022a.

Ling-Qi Zhang, Zahra Kadkhodaie, Eero P Simoncelli, and David H Brainard. Image reconstruction from cone excitations using the implicit prior in a denoiser. *Journal of Vision*, 22(14):3793–3793, 2022b.

Xuemei Zhang, Brian A Wandell, et al. A spatial extension of cielab for digital color image reproduction. In *SID international symposium digest of technical papers*, volume 27, pages 731–734. Citeseer, 1997.

Mingyi Zhou, John Bear, Paul A Roberts, Filip K Janiak, Julie Semmelhack, Takeshi Yoshimatsu, and Tom Baden. Zebrafish retinal ganglion cells asymmetrically encode spectral and temporal information across visual space. *Current Biology*, 30(15):2927–2942, 2020.

Maxime JY Zimmermann, Noora E Nevala, Takeshi Yoshimatsu, Daniel Osorio, Dan-Eric Nilsson, Philipp Berens, and Tom Baden. Zebrafish differentially process color across visual space to match natural scenes. *Current Biology*, 28(13), 2018.

Ehud Zohary, Michael N Shadlen, and William T Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, 1994.