Parametric modelling of visual cortex at multiple scales

Patrick Mineault

Integrated Program in Neuroscience

McGill University

Montreal, Quebec, Canada

February 2014

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

for my father

pour mon père

Acknowledgements

I would first like to thank labmates and fellow graduate students who helped me along the way: Michael Oliver, Jeremy Freeman, Marino Pagan, Corey Zembia, Vargha Talebi, Alby Richard, James Tsui, Michael Waterston, Frédéric Simard, Jachin Monteon, Liu Liu, Jan Churan, Nuha Jabakhanji, Faisal Naqib, and Mathieu Boulanger. I would especially like to thank Julie Coursol and Naomi Takeda for technical and administrative support.

I was inspired and challenged by professors and scientists who have critiqued my work and encouraged me: Mike Langer, Dan Guitton, Kathy Cullen, Geoff Boynton, Greg Horwitz, Stefan Treue, Adam Kohn, Greg DeAngelis, David Heeger, Nicole Rust, Jim DiCarlo, Jack Gallant, Stephen David, Tony Movshon, Eero Simoncelli, Liam Paninski, Jonathan Pillow, Yoshua Bengio, Geoffrey Hinton, Dario Ringach and the late David Hubel, without whom none of this work would have possible.

I am thankful for the members of my committee, Curtis Baker, Ed Ruthazer and Erik Cook, who have given very useful feedback on my progress and steered me along the right direction.

The work presented here would have been impossible without the help of collaborators: Simon Barthelmé, who graciously assisted me in writing my first paper; Farhan Khawaja, who tirelessly did neurophysiological recordings; Dan Butts, who guided me in modelling; and Theodoros Zanos, with whom I've had several fruitful collaborations.

I would like to thank all my friends who have been patient and understanding throughout this 6-year period: Amélie, Toni, Ève, Marianne, Samuelle, Philippe, Mary-Ève, Ivan, Theo, Alby, and all my dance partners. I would especially like to thank my girlfriend Evelyne and my mother Roselyne for their encouragement.

None of this would have possible, of course without Dr. Christopher Pack, who has guided me tirelessly and with great patience through 5 manuscripts and countless projects; I cannot thank him enough.

Finally, this research was made possible by a PhD scholarship (149928) from the Fonds Québécois de Recherche, Nature et Technologies and operating grants to Dr. Pack from the Canadian Institures of Health Research (MOP-115178) and Collaborative Research in Computational Neuroscience from the National Science Foundation (IIS-0904430).

Contribution of authors

The following thesis is organized into five main chapters: an introduction, three chapters reproducing the text of three published peer-reviewed manuscripts (Mineault el a. 2009, 2012, 2013) and a discussion. This is followed by three appendices, which provide supporting information relevant to each main manuscript (Mineault et al. 2009, 2012; Zanos, Mineault & Pack 2011).

The introduction and discussion were written by myself. The manuscripts were produced under the supervision of Dr. Christopher Pack; In all cases, C.C.P. helped conceive the experimental protocols, suggested analyses, contributed to the interpretation, and helped write the manuscripts.

For the first manuscript (Mineault et al. 2009), I performed psychophysical experiments, developed the methods, analyzed data, and wrote a majority of the manuscript. Author S.B. contributed to the theoretical framework, developed methods and edited the manuscript.

For the second manuscript (Mineault et al. 2013), I designed the experimental protocol, developed the methods, analyzed the experimental data, and wrote a majority of the manuscript. Author T.Z.P. performed the experimental recordings and edited the manuscript.

For the third manuscript (Mineault et al. 2012), I developed the methods, analyzed the experimental data, performed the simulations, and wrote a majority of the manuscript. Author F.A.K performed the experimental recordings and edited the manuscript. D.A.B. helped develop the experimental protocol, provided modeling guidance, and edited the manuscript.

Appendix A, which accompanies the first main manuscript, was originally published as the supplementary information in (Mineault et al. 2009).

Appendix B, which accompanies the second manuscript, was originally published as the supplementary information in Zanos, Mineault & Pack (2011). The reproduced text, which I wrote, reflects my contribution to Zanos et al. (2011), namely the development of a signal processing method to estimate local field potentials; this method is a crucial step in the analysis presented in the second main manuscript (Mineault et al. 2013), hence it is reproduced here. T.P.Z. performed the experiments, the analysis, and wrote the main text of Zanos et al. (2011); this text is not reproduced here.

Appendix C, which accompanies the third main manuscript, was originally published as the supplementary information in (Mineault et al. 2012).

Contents

List of figures

List of tables

List of algorithms

Abbreviations and symbols

AIC	Akaike Information Criterion
ANN	Artificial neural network
ARD	Automatic relevance determination
BOLD	Blood-oxygen-level-dependent
cdf	cumulative distribution function
dsANN	deep shared artificial neural network
DS	Direction selective
DCT	Discrete cosine transform
DFT	Discrete fourier transform
EP	Expectation propagation
FPC	Fixed-point continuation
fMRI	functional magnetic resonance imaging
GPL	General Public License
GAM	generalized additive model
GLM	Generalized Linear Model
GQM	Generalized Quadratic Model
IT	Inferotemporal cortex
IRLS	Iteratively reweighted least-squares
LGN	Lateral geniculate nucleus
LNP	Linear nonlinear Poisson
LFP	Local field potential
MAP	Maximum a posteriori
ML	Maximum likelihood
MST	Medial superior temporal
MT	Middle temporal
MEA	Multi-electrode array
MUA	Multi-unit activity
NIM	Nonlinear input model
PSBF	Pesudo-Bayes factor
PSD	Power spectral density
RF	Receptive field
RGC	Retinal ganglion cells
RMS	Root-mean-squared error
STRF	Spectro-temporal receptive field
STA	Spike triggered average
STC	Spike triggered covariance
UBRE	Unbiased risk estimator

Abstract

The visual system is confronted with the daunting task of extracting behaviourally relevant visual information from noisy and ambiguous patterns of luminance falling on the retina. It solves this problem through a hierarchical architecture, in which the visual stimulus is iteratively re-encoded into ever more abstract representations which can drive behaviour. This thesis explores the question of how the computations performed by neurons in the visual hierarchy create behaviourally relevant representations.

This question requires probing the visual system at multiple scales: computation is the role of single neurons and ensembles of neurons; representation is the function of multiple neurons within an area; hierarchical processing is an emergent process which involves multiple areas; and behaviour is defined at the full scale of the system, the psychophysical observer.

To study visual processing at multiple scales, I propose to develop and apply parametric modelling methods in the context of systems identification. Systems identification seeks to establish the deterministic relationship between the input and the output of a system. Systems identification has proven particularly useful in the study of visual processing, where the input to the system can be easily controlled via sensory stimulation.

Parametric modeling, built on the theory of Generalized Linear Models (GLMs), furnishes a common framework to analyze signals with different statistical properties which occur in the analysis of neural systems: spike trains, multi-unit activity, local field potentials and psychophysical decisions.

In Chapter 2, I develop the parametric modeling framework which is used throughout this thesis in the context of psychophysical classification images. Results show that parametric modeling can infer a psychophysical observer's decision process with fewer trials than previously proposed methods. This allows the exploration of more complex, and potentially more informative, models of decision processes while retaining statistical tractability.

In Chapter 3, I extend and apply this framework to the analysis of visual representations at the level of neuronal ensembles in area V4. The results show that it is possible to infer, from multi-unit activity and local field potential (LFP) signals, the representation of visual space at a fine-grained scale over several millimeters of cortex. Analysis of the estimated visual representations reveals that LFPs reflect both local

sources of input and global biases in visual representation. These results resolve a persistent puzzle in the literature regarding the spatial reach of the local field potential.

In Chapter 4, I extend and apply the same framework to the analysis of single-neuron responses in area MST of the dorsal visual stream. Results reveal that MST responses can be explained by the integration of their afferent input from area MT, provided that this integration is nonlinear. Estimated models reveal long suspected, but previously unconfirmed receptive field organization in MST neurons that allow them to respond to complex optic flow patterns. This receptive field organization and nonlinear integration allows more accurate estimation of the velocity of approaching objects from the population of MST neurons, thus revealing their possible functional role in vergence control and object motion estimation.

Put together, these results demonstrate that with powerful statistical methods, it is possible to infer the nature of visual representations at multiple scales. In the discussion, I show how these results may be expanded to gain a better understanding of hierarchical visual processing at large.

Résumé

Le système visuel est confronté à la difficile tâche d'extraire de l'information utile au comportement à partir de motifs complexes et ambigus détectés par la rétine. Il résout ce problème grâce à une architecture hiérarchique, dans laquelle le stimulus visuel est itérativement ré-encodé dans une représentation abstraite. Ce mémoire explore la question suivante : comment les computations performées par des neurones de la hiérarchie visuelle créent-elles des représentations permettant des comportements complexes?

Cette question nécessite l'étude du système visuel à plusieurs échelles : la computation est le rôle de neurones et d'ensembles de neurones; la représentation est une fonction des neurones dans une aire du cerveau; la hiérarchie émerge de la communication entre de multiples aires du cerveau; et le comportement est défini à l'échelle du système visuel complet, l'observateur psychophysique.

Afin d'étudier le système visuel à de multiple échelles, je développe et applique des méthodes de modélisation paramétrique dans le cadre de l'identification de système. Celle-ci a pour but d'établir la relation déterministe entre l'entrée d'un système et sa sortie. L'identification de système est particulièrement utile dans l'étude de la vision, où l'entrée du système peut être facilement contrôlée par stimulation sensorielle.

La modélisation paramétrique, bâtie sur la théorie des modèles linéaires généralisés, offre un paradigme commun pour analyser des signaux ayant des propriétés statistiques disparates, souvent rencontrés dans l'étude du système nerveux: les potentiels d'action, l'activité d'ensemble de neurones, et les décisions psychophysiques.

Dans le 2ème chapitre, je développe le paradigme d'analyse par modélisation paramétrique qui sera utilisé tout au long de ce mémoire dans le contexte des images de classification psychophysiques. Je démontre qu'il est possible d'inférer, grâce à ces méthodes, le processus décisionnel d'un observateur psychophysique avec moins de données que ce qui était précédemment possible. Cette avancée permet l'exploration de modèles psychophysiques plus complexes, et potentiellement plus informatifs sur le processus décisionnel de l'observateur.

Dans le 3ème chapitre, j'applique ce paradigme à l'analyse des représentations visuelles au niveau d'ensembles neuronaux dans l'aire V4 du système visuel. Les résultats démontrent qu'il est possible, à partir de l'activité des champs de potentiel locaux (CPL), d'inférer la représentation corticale de l'espace

visuel sur une échelle de plusieurs millimètres. Je démontre ainsi que les CPL reflètent à la fois des sources synaptiques locales et des biais globaux dans la représentation visuelle. Ces résultats résolvent une controverse dans la littérature concernant l'intégration spatiale des CPL.

Dans le 4ème chapitre, j'applique ce même paradigme dans l'analyse de neurones dans l'aire MST du système visuel dorsal. Je révèle que les réponses dans MST peuvent être expliquées par l'intégration de sources afférentes provenant de l'aire MT; cependant, cette intégration se révèle nonlinéaire. Cette analyse révèle des propriétés longtemps soupçonnées mais jusqu'ici non confirmées des champs réceptifs des neurones dans MST; celles-ci leur permettent de communiquer de l'information sur les motifs de flux optique complexes. Cette organisation des champs réceptifs et l'intégration nonlinéaire permet d'extraire plus facilement la vélocité d'objets s'approchant de l'observateur à partir des réponses de la population de neurones dans MST, révélant un rôle insoupçonné de ces neurones dans l'estimation de la vélocité des objets.

Pris ensemble, ces résultats démontrent qu'à l'aide de méthodes statistiques puissantes, il est possible d'inférer la nature des représentations visuelles à de multiples échelles. Dans la discussion, je démontre comment généraliser ces résultats afin d'obtenir une meilleure compréhension des computations hiérarchiques dans le système visuel.

Chapter 1 opens with a brief overview of how the thesis is structured. I then give a broad overview of hierarchical visual processing and systems identification methods relevant to this thesis. This is followed by a more detailed treatment of parametric modelling methods which will be used throughout the rest of this manuscript. Finally, I introduce the rationale behind the research conducted in this dissertation.

1. Introduction

1.1 Organization

In the following thesis, I develop and apply a parametric systems identification framework to study how visual stimuli are represented at the single-neuron, multi-neuron, and psychophysical levels. In the introduction, I first give a broad outline of visual processing, as revealed by neurophysiological and psychophysical experiments. I then give an overview of systems identification and its uses in revealing the mechanisms by which visual stimuli are translated into behaviour. Finally, I give a technical exposition of parametric systems identification methods which will be used throughout the rest of the thesis.

Three previously published manuscripts (Mineault et al., 2009, 2012, 2013), on psychophysical systems identification, multi-neuron selectivity in area V4, and single-neuron selectivity in area MST, are then presented in the main chapters of the thesis. This is followed by a discussion in which I examine the larger issues surrounding this work and suggest new experiments to further elucidate visual representation.

Appendices A,B and C present supplementary material relevant to the technical aspects to each of the three main studies presented in this thesis.

1.2 Overview

Humans are visual creatures. Roughly a third of human cortex is dedicated to visual processing; this figure rises to 50% in macaques, a frequently used animal model of high-level visual processing (Hubel, 1995). Some types of visual behaviours are driven by relatively simple neural circuits; for example, intrinsically photosensitive retinal ganglion cells measure diurnal variation in light in the blue-UV range to entrain the neural pacemaker neurons of the suprachiasmatic nucleus (Hattar et al., 2002).

For more complex behaviours, however, the visual system is confronted with the difficult task of extracting behaviourally relevant information from a stream of noisy, ambiguous photon counts in the retina. This information is generally classified into two main classes, form and motion, which in primates are processed by two partially segregated neocortical pathways, the *what* and *where* pathways (Ungerleider and Mishkin, 1982).

At the highest levels of the visual pathways, behaviourally relevant information, encoded in the firing rates of neurons, can readily guide visual behaviours via a simple readout rule. For instance, neurons at

the highest level of form processing, in inferotemporal cortex, can reliably convey information about object identity based on visual cues (Quiroga et al., 2007). Similarly, neurons in area MST, a high-level motion area, can reliably signal the heading of an observer based on the integration of wide-field visual motion information and vestibular signals (Gu et al., 2010).

In this thesis, I study the computations performed by the visual system which hierarchically reencode latent visual information into a representation which can guide behaviour (DiCarlo and Cox, 2007). This requires studying the visual system at multiple scales (DiCarlo et al., 2012): computation is a function of single neurons and small ensembles of neurons; representation is a function of larger ensembles of neurons, i.e. a single area; hierarchical processing is a collaboration between multiple areas of the brain; and behaviour is defined at the level of the whole system, i.e. the psychophysical observer. In order to study these disparate systems, I develop and apply parametric systems identification methods, which can relate a system's output to the visual stimulus, whether this system is defined at the single-neuron, multi-neuron, visual area, or psychophysical observer scale (Victor, 2005).

1.3 Hierarchical visual processing – evidence from single neurons

1.3.1 Early visual processing

In lower-level mammals, complex, behaviourally relevant information is sometimes extracted directly at the level of the retina. In mice, for example, a complex intraretinal circuit causes a class of retinal ganglion cells (RGCs), known as PV-5, to be sensitive to dark looming stimuli (Münch et al., 2009). Such sophisticated intraretinal processing can directly drive a complex behavioural response to a triggering stimulus, e.g. an approaching predator. Indeed, the selectivity properties of PV-5 RGCs are well-matched to the stimulus parameters which trigger a rapid freezing reflex in mice (Yilmaz and Meister, 2013).

In primates, however, the retina has not been shown to extract complex visual motion or form information (Field and Chichilnisky, 2007). Rather, visual information is lightly processed by an intraretinal circuit and is relayed to multiple classes of retinal ganglion cells. A photon falling on a cone causes a conformational change in an opsin, which, via a second-messenger cascade, leads to the hyperpolarization of the cone (Kandel et al., 2000). The graded signal of multiple cones is integrated in ON- and OFF- bipolar cells, which depolarize in response to an increase or decrease in luminance, respectively. Horizontal cells, which contact both cones and bipolar cells, and amacrine cells, which contact bipolar and retinal ganglion cells, laterally process this information, attenuating spatiotemporal correlations in the luminance signal (Field and Chichilnisky, 2007).

The 4 most common sub-types of retinal ganglion cells, the midget and parasol ON- and OFF- RGCs, have center-surround receptive fields (RFs) as a consequence of this intraretinal processing. That is, an ON-center retinal ganglion cell increases its firing rate in response to an increment in luminance in a localized area of the retina – the center - and to a decrement in luminance in an area surrounding the center – the surround. OFF-center cells respond similarly to the opposite stimulus polarity. This organization enhances the representation of high spatial frequency stimuli, and has been shown to lead to an efficient, decorrelated representation of natural images, which are dominated by low spatial frequencies (Atick and Redlich, 1992).

Intraretinal circuitry also participates in nonlinear luminance and contrast adaptation (Shapley et al., 1993). This is a crucial nonlinear processing stage that allows retinal ganglion cells to relay useful information about the visual world over the 9 orders of magnitude of luminance that separate the absolute threshold of vision and a typical scene in full sunlight.

The various subtypes of retinal ganglion cells – midget and parasol cells, but also small bistratified and the other less well characterized subtypes – differ in their contrast, temporal, spatial, and spectral sensitivity (Field and Chichilnisky, 2007). Thus, while not a passive processing stage, the primate retina primarily reencodes visual stimuli into multiple parallel representations as normalized, spatially localized differences in luminance in space and time (Field and Chichilnisky, 2007).

The primary thalamic projection of RGCs, the lateral geniculate nucleus (LGN), elaborates little on this basic representation. Indeed, the ON-OFF receptive field organization visible in RGCs is preserved in LGN neurons (Derrington and Lennie, 1984). Since LGN neurons receive input from a very limited number of RGCs – 1 to 3 (Huberman, 2007) – their ability to reencode the visual stimulus and become selective for high-level features is fundamentally limited.

1.3.2 Primary visual cortex and the canonical circuit of Hubel & Wiesel

The crux of primate visual processing appears to be undertaken in cortex. LGN neurons project to layer 4C of primary visual cortex (area V1; Thomson and Bannister, 2003). While neurons in the LGN and the retina can be strongly driven by flashed spots of light, Hubel and Wiesel (Hubel and Wiesel, 1962, 1968) found that most neurons in V1 are poorly driven by such stimuli. Strikingly, a large proportion of V1 neurons are instead responsive to oriented stimuli such as bars and gratings. An orientation-tuned V1 neuron will respond strongly to optimally oriented stimuli, with its response decreasing, in some cases below baseline, as stimuli are rotated away from the neuron's preferred orientation.

A subset of these neurons – direction selective (DS) cells - is selective for the direction of motion of oriented stimuli. DS cells respond to motion in one of the directions orthogonal to the orientation of stimuli, while being insensitive or inhibited by motion in the opposite direction (De Valois et al., 1982). It is believed that direction-selective cells form the first step in a hierarchical computation which processes motion information in the dorsal visual pathway (Shipp and Zeki, 1989). Conversely, non-direction selective cells may form the basis of form processing in the ventral visual pathway (Goodale and Milner, 1992).

Both direction and non-direction selective cells can be further characterized as being simple or complex. Simple cells are sensitive to contrast sign – for example, a vertically-tuned simple cell may respond to an edge that is dark on the left side and light on the right side. A simple cell will not respond, however, when the light and dark regions of this stimulus are reversed. When probed with a moving grating, the response of a simple cell is strongly modulated as a function of time, as the phase of the grating becomes matched or mismatched to the receptive field of the simple cell (DeAngelis et al., 1993).

Complex cells, however, are insensitive to the contrast sign of the oriented stimuli which drive them. That is, a complex cell may respond to a light bar, a dark bar, or to edges of either polarity in its receptive field, provided that these stimuli are at the preferred orientation of the cell. In response to a moving grating, a complex cell is not strongly modulated as a function of time; rather, it shows a sustained response, as it responds at all phases of the grating (Movshon et al., 1978).

Hubel and Wiesel suggested a neural implementation for simple and complex cells which remains a highly influential model of sensory processing (Hubel and Wiesel, 1962). They hypothesized that simple cells are built by the convergence of ON- and OFF- selective LGN afferents with the correct receptive field properties. For example, a neuron selective for vertical edges may be built by the convergence of excitatory inputs from OFF- cells on the left side of its receptive field and ON- cells on the right side of its RF. The threshold of the neuron then shapes the response of the neuron such that it becomes insensitive to other stimuli at non-preferred orientations or phases.

A complex cell, in turn, may be built by tiling its receptive field with simple cells selective for the same orientation, but with different preferred phases and positions. While the responses of simple cells converging to a complex cell would cancel out if they operated in a linear regime, the nonlinearity inherent in the threshold of simple cells prevents this cancellation (Adelson and Bergen, 1985).

Thus, two simple neuronal operations, feedforward synaptic integration and nonlinear spike generation, when cascaded, can form novel representations which are highly sensitive to some aspects of the stimulus, like orientation, while being insensitive to other aspects, like spatial phase (Riesenhuber and Poggio, 1999).

1.3.3 Anatomical evidence for the simple/complex cell circuit

Anatomical evidence is broadly consistent with the spirit of the circuit proposed by Hubel & Wiesel. In lower mammals, the position and polarity of LGN afferents to simple cells is predictive of their preferred orientation and phase (Lien and Scanziani, 2013; Reid and Alonso, 1995), although this has yet to be verified in macaque. Spike generation enhances the selectivity of both simple and complex cells to orientation compared to the underlying intracellular voltage (Priebe et al., 2004).

Finally, simple and complex cells are organized hierarchically, consistent with the canonical circuit. In layer 4C of primary visual cortex, which receives direct input from the LGN, a large proportion of cells are simple. These neurons project to more superficial layers of V1, where the majority of neurons are complex (Ringach et al., 2002).

The canonical simple/complex circuit of Hubel and Wiesel has been further elaborated and refined by theoretical and anatomical studies. Recurrent excitation within V1 may refine and amplify the orientation bias conferred by the feedforward integration of LGN inputs (Ringach, 2004; Somers et al., 1995). Indeed, up to two thirds of the input to simple cells comes not from the LGN but from cortico-cortical connections within V1 (Finn et al., 2007; Lien and Scanziani, 2013).

V1 neurons, including simple cells, display a broad range of nonlinear tuning properties, including contrast invariant orientation tuning, phase advance with increasing contrast, cross-orientation suppression, and surround suppression (Carandini et al., 2005). Much of these effects can be reconciled with the canonical circuit by assuming that the firing rate of a V1 neuron is divisively normalized by the output of a large pool of other V1 neurons (Carandini and Heeger, 2011; Heeger, 1992; Heeger and Carandini, 1994). Divisive normalization could be implemented by inhibitory circuits within V1 (Wilson et al., 2012). Indeed, synaptic activity in the visual cortex of awake behaving animals is dominated by inhibition (Haider et al., 2013).

Thus, the canonical circuit of Hubel & Wiesel, augmented with refinements such recurrent excitation, inhibition, normalization, and noise (Anderson et al., 2000), provides an accurate summary of the computations performed by orientation-selective V1 neurons (Carandini et al., 2005).

1.3.4 Theoretical perspectives on visual computation

The canonical simple/complex circuit illustrates how cortex can create selectivity for interesting features, like orientation, while building invariance for irrelevant features, like spatial phase. It has been hypothesized that by stacking selectivity and invariance computations in a hierarchical scheme, cortex can build selectivity for ever more complex features, while building invariance for transformations like translation, scaling, and changes in pose and illumination (Bottou et al., 1994; Fukushima, 1980; Poggio et al., 2012; Riesenhuber and Poggio, 1999).

Before we consider further the evidence for hierarchical computation in visual cortex, let us consider its potential role. From an information-theoretic perspective, the raw stream of photon counts available at the level of the retina contains all the information necessary to guide behaviour. Indeed, linear and nonlinear transformations, hierarchical or not, cannot create information de novo, only preserve or destroy information (Cover and Thomas, 2012).

This information, however, is not necessarily in a format which is easily readable by biophysically plausible mechanisms. As pointed out in David Marr's classic book on visual computation, "there is a trade-off; any particular representation makes certain information explicit at the expense of information that is pushed into the background and may be quite hard to recover" (Marr, 1982).

For example, a face will generate a large number of distinct images when projected onto the retina at different scales, rotations, and illuminations. From the perspective of a neuron downstream from an ensemble of retinal ganglion cells, each face corresponds to myriad patterns of spike counts– a cloud of points, or manifold, in a large-dimensional space (DiCarlo and Cox, 2007; Figure 1-1A).

Now, consider a downstream neuron that uses visual information about faces to guide a behavioural decision, such as approach or avoidance. It must translate the retinal projection of a face in an uncontrolled condition of illumination, pose and position into a behavioural decision. However, a neuron is limited, by biophysical principles, in the number of operations it can perform on its input: synaptic integration, normalization, thresholding, etc. (Carandini and Heeger, 2011; Koch, 1999; Kouh and Poggio, 2008).

Since it is only mildly nonlinear, it can only discriminate two different classes of faces if the manifolds corresponding to each class are approximately linearly separable (Figure 3-1-1B). Theoretical studies show that affine transformations – translation, scaling, and rotation – as well as non-affine

transformations – deformation, clutter, illumination changes, etc. – indeed generate highly curved manifolds (DiCarlo and Cox, 2007; Poggio et al., 2012; Figure 1-1C).

The downstream neuron may, however, perform its discrimination task and thus successfully guide behaviour if it receives input from a more abstract representation which has factored out, or linearized, identity-preserving transformations. This is the untangling hypothesis (DiCarlo and Cox, 2007; Poggio et al., 2012): the role of the visual, and other sensory hierarchies, is to untangle – that is, to linearize and spread out - the highly nonlinear manifolds generated by objects under a variety of pose, illumination, and other identity-preserving conditions.

The untangling hypothesis thus bridges encoding and decoding perspectives: the encoding process, which is reflected in the receptive field of visual neurons, facilitates a decoding process, which is hypothesized to exist in a high-level, non-visual area, for example prefrontal cortex (DiCarlo and Cox, 2007).

Theoretical studies show that hierarchical architectures can iteratively factor out identity-preserving transformations (Bengio, 2009; Poggio et al., 2012) with a modest number of computation elements - neurons. By contrast, shallow architectures, which perform invariant recognition by comparing the sensory pattern at hand with a number of templates corresponding to different views of the target object, require a number of computation elements which is exponential in the number of identity-preserving transformations that must be factored out (Bengio, 2009).

Thus, hierarchical architectures, built by stacking selectivity and invariance operations – simple and complex-like operations – are well-adapted to the task of building abstract representations which support complex behaviours (Poggio et al., 2012). As we will see next, both motion and form appear to be extracted by hierarchical processes in the dorsal and ventral visual streams, respectively.

1.3.5 Hierarchical computation in the ventral visual stream

Physiological evidence is broadly consistent with the idea that more abstract and complex representations are built by stacking selectivity and invariance operations. The processing of form is hypothesized to take place in the ventral visual stream, a collection of areas going from the occipital lobe to the temporal lobe (Felleman and Van Essen, 1991; Ungerleider and Mishkin, 1982).

Anatomical evidence shows that V1 has a major projection in area V2, which projects to V4, which itself projects to area IT (Felleman and Van Essen, 1991). These areas, which form the major elements of the

ventral stream, are not organized in a strictly hierarchical manner, however. Direct connections from V1 to V4 and IT exist (Felleman and Van Essen, 1991); feedback connections project from higher to lower-level areas (Felleman and Van Essen, 1991); and area IT is itself subdivided into myriad subareas (Tanaka, 1992). Nevertheless, latency, receptive field size, and receptive field complexity all increase as the visual input is processed by the successive stages of the ventral stream (Freeman and Simoncelli, 2011; Kobatake and Tanaka, 1994; Schmolesky et al., 1998).

Area V2, the largest extrastriate area, is one of the main projections of V1. Although it contains some direction-selective cells, their proportion is limited (Orban et al., 1986), and V2 is therefore considered the first extrastriate stage of the form-processing ventral visual stream.

V2 is highly heterogeneous. Under cytochrome oxidase staining, strongly stained regions of varying width – thin and thick stripes - are separated by lightly stained regions – pale stripes (Hubel and Livingstone, 1987; Livingstone and Hubel, 1988). These histologically-defined regions correlate with different functional specializations: disparity-sensitive neurons are concentrated in the thick stripes; colour-sensitive neurons in the thin stripes; while end-stopped neurons sensitive to luminance-and-contrast-defined form are concentrated in the pale regions.

Although as a whole, V2 is poorly understood, what we know about V2 is consistent with the idea that it elaborates on the processing performed by V1 neurons. The vast majority of orientation-selective V2 neurons are phase-invariant and thus classified as complex cells (Levitt et al., 1994). A subset of V2 neurons appears to be sensitive to second-order form patterns; that is, patterns which are defined not by differences in spatial luminance, but differences in contrast and orientation energy (Baker Jr, 1999; Baker et al., 2013; Li and Baker, 2012; Mareschal and Baker, 1998). Some V2 neurons have non-homogenous receptive fields, with varying preferred orientation as a function of position in the receptive field (Anzai et al., 2007; Hegdé and Van Essen, 2000). Nonlinear spatial interactions within receptive fields, whose nature can be excitatory or inhibitory, have also been reported (Anzai et al., 2007).

Furthermore, V2 appears specifically sensitive to higher-order structure in natural images (Freeman et al., 2013). Natural images are defined by both low-order structure (mean and covariance) as well as high-order correlations. Artificial textures, which capture local structural information, can be built by matching the distribution of products of complex-cell like outputs across scales, positions, and orientations to the empirical distribution of these products in a natural image (Portilla and Simoncelli,

2000). When probed with artificial textures, V2 neurons respond at a higher rate than when probed with noise stimuli with matched low-order correlations, in contrast to V1 neurons (Freeman et al., 2013).

All these properties are consistent with the idea that V2 neurons integrate complex cell information from V1. Computationally, nonhomogenous and second-order-sensitive receptive fields can be obtained by taking sums and differences of complex cells with appropriate receptive field properties (Mareschal and Baker, 1998). Nonlinear interactions and texture selectivity, on the other hand, can appear by taking products of complex cells (Anzai et al., 2007; Freeman et al., 2013).

Both sums and products can be performed by the same basic operations that are the basis of the canonical simple/complex circuit: linear, feedforward integration followed by a static spiking nonlinearity. To see that this is the case, assume that a neuron integrates from two complex cells with non-negative outputs a and b, and that its spiking nonlinearity can be approximated by half-squaring (Carandini et al., 1997). Its response is then given by:

$$(a+b)^2 = a + b + 2 \cdot a \cdot b$$

Thus, linear integration followed by a threshold can generate both additive and multiplicative interactions between inputs to the cell. The properties of V2 neurons are therefore consistent with feedforward integration of complex cells, followed by a static nonlinearity.

Both V1 and V2 send converging input to area V4 (Nakamura et al., 1993; Shipp and Zeki, 1985). Cells within V4 are selective for features of intermediate complexity, including non-Cartesian gratings (Gallant et al., 1993, 1996), 3 dimensional orientation (Hinkle and Connor, 2002) and texture conformation (Hanazawa and Komatsu, 2001). Pasupathy and Connor (2001, 2002), in particular, showed that neurons in V4 are selective for boundary conformation cues. Thus, in a given part of its receptive field, a V4 cell may be selective for a concave or convex boundary of the correct orientation.

Cadieu et al. (2007) reanalyzed the results of the Pasupathy and Connor (2001) dataset with a hierarchical model of the ventral visual stream: a simple and a complex cell layer are followed by a selectivity layer operating on complex cells and a final invariance layer integrating the outputs of the second selectivity layer across space. Cadieu et al. showed that this model can account for much of the boundary conformation selectivity properties of V4 neurons, reporting excellent cross-validated model predictions. Thus, the alternation of selectivity and invariance operations, in a hierarchical

generalization of the Hubel & Wiesel canonical circuit, can account for the properties of neurons in intermediate visual cortex.

The same principles appear to apply in higher-level cortex as well. Inferotemporal cortex (IT), which receives converging input from V1, V2, and V4, is a heterogeneous complex which includes a large number of functional subdivisions (Felleman and Van Essen, 1991). Here, for the first time, receptive fields span both ipsilateral and contralateral sides of the visual field, and responses are influenced not only by the visual stimulus, but also by memory traces (Pagan et al., 2013).

Most IT neurons are selective for complex stimuli defined by multiple component features (Kobatake and Tanaka, 1994), thereby building on the selectivity provided by V4. Remarkably, this selectivity is accompanied by a large degree of invariance, in particular to scaling, such that relative stimulus preferences can be preserved over 10-fold changes in stimulus size (Riesenhuber and Poggio, 1999).

The evidence that these selectivity and invariance properties arise hierarchically within IT itself is most convincing in the case of face recognition. A network of 6 patches within IT has been shown to be specifically sensitive to faces (Freiwald et al., 2009). Stimulation experiments revealed a partial hierarchical structure within this network; middle face patches ML & MF have strong feedforward connections to anterior face patches AL & AM.

Mapping face selectivity profiles within each patch revealed a functional specialization consistent with the anatomy (Freiwald and Tsao, 2010); neurons in the middle face patches are tuned to a single view of a face; neurons in AL are frequently tuned to two, mirror-symmetric views of the same face, thus achieving partial invariance to pose; and neurons in AM achieve almost full invariance to pose. Selectivity for particular face identities increased concomitant with this increase in invariance.

Thus, the highest levels of the ventral stream can support invariant recognition of faces, in the manner hypothesized by DiCarlo and Cox (2007). The broad strokes of this computation are consistent with the stacking of selectivity and invariance operations in a hierarchical fashion.

1.3.6 Hierarchical computation in the dorsal visual stream

The dorsal visual stream, which stretches from the occipital to the parietal lobe, is hypothesized to be specialized for the processing of motion and space. Direction-selective cells of primary visual cortex project to the middle temporal (MT) area of the dorsal visual stream (Shipp and Zeki, 1989). The vast majority of MT cells are direction and speed selective (Albright, 1984). Microstimulation of MT biases

motion perception, showing that they are causally involved in the discrimination of motion (Salzman et al., 1992).

They elaborate on the processing of V1 direction-selective cells in a number of ways. Most MT neurons show weak selectivity for static stimuli, and have RFs which are larger by a factor 10 than V1 direction-selective cells at comparable eccentricities (Albright, 1984). A subset of MT neurons is selective for pattern motion as opposed to component motion. That is, when presented with sums of gratings with multiple motion components – plaids – these neurons respond to the coherent motion component implied by the subcomponents, rather than responding to each component individually, as V1 cells do (Movshon et al., 1985).

In a related computation, some MT neurons can signal the speed of motion independent of spatial frequency (Priebe et al., 2003). Finally, MT neurons are frequently suppressed by motion in their surround (Born and Bradley, 2005). These observations are consistent with the idea that MT neurons integrate, in a nonlinear fashion (Cui et al., 2013; Rust et al., 2006; Simoncelli and Heeger, 1998), the outputs of V1 direction-selective cells to form a distributed representation of velocity which is invariant to static stimulus features.

The medial superior temporal (MST) area receives the majority of its input from area MT (Boussaoud et al., 1990). Again, MST neurons build on the selectivity of MT neurons. They have 10-fold larger receptive fields than in MT (Duffy and Wurtz, 1991a, 1991b), spanning up to a quarter of the visual field, and often include both the fovea and parts of the ipsilateral visual field. Compared to MT neurons, they are still less sensitive to static and flashed stationary stimuli, and respond more to global than to component motion (Khawaja et al., 2013).

Strikingly, a subset of MST neurons is highly selective for complex optic flow stimuli, in particular expansion, contraction, and rotation patterns (Duffy and Wurtz, 1995; Graziano et al., 1994; Tanaka et al., 1986). This selectivity for complex motion patterns is often invariant to changes in other stimulus features (Duffy and Wurtz, 1995; Geesaman and Andersen, 1996), including the spatial position and the visual contents of the patterns carrying the motion.

Optic flow processing has been hypothesized to be important for the decoding of ego-motion from visual information (Koenderink, 1986; Warren et al., 2001). Consistent with this idea, neurons in MST integrate from another source of ego-motion information, vestibular input, and heading information can reliably be decoded from the population of MST neurons (Gu et al., 2010).

It is reasonable to assume that the selectivity for complex optic flow patterns in MST neurons arises in part from the spatial arrangement of MT neurons that feed into a given MST cell, much like the selectivity of V1 simple cells for orientation arises from the specific spatial arrangement of LGN cells in the Hubel & Wiesel canonical circuit. Indeed, many models of MST have posited that MST receptive fields are derived by the spatial arrangement of direction-tuned MT subunits (Royden et al., 1994; Tohyama and Fukushima, 2005; Zemel and Sejnowski, 1998; Zhang et al., 1993).

To date, however, these models have yet to be fit to neural data. I attack this difficult problem in Chapter 4 of this thesis, and show that MST responses are indeed consistent with a feedforward integration of MT inputs, provided that this integration in nonlinear (Mineault et al., 2012).

What occurs after MST is less clear. Many areas of parietal cortex, in particular, LIP, VIP and v7a, receive input from MST and are also selective for complex optic flow (Orban, 2008). These areas appear to elaborate on the processing of earlier areas to form a unified representation of space (Colby and Goldberg, 1999). In v7a, for example, neurons integrate both optic flow and gaze-position signals, an important step in computing self-motion in a body-centered, as opposed to retina-centered coordinate frame (Siegel and Read, 1997). In area VIP, inputs from multiple sensory modalities, including visual and somatosensory inputs, are integrated in an approximately body-centered coordinate frame (Graziano and Gross, 1995). Finally, in area LIP, saliency of visual stimuli is encoded, and such a signal may be used to guide eye movements (Gottlieb et al., 1998).

Together, later stages of the dorsal visual stream cannot be said to be strictly dedicated to the encoding of higher-level motion features (Wolpert et al., 1998). In natural environments, motion is consistently associated with transient changes in the environment which require immediate action: unexpected wide-field optic flow may result from a loss of balance, or a deviation from a locomotion plan (Royden et al., 1994); small field motion may signal an approaching predator or a new salient stimulus to which gaze should be directed (Folk et al., 1994; Kusunoki et al., 2000).

Conversely, motor programs which result in changes in body, head and eye position create strong patterns of motion on the retina (Royden et al., 1994). Thus, motion is a call to action, and action causes motion (Gibson, 1986; Noë, 2004). Later stages of the dorsal visual stream thus appear to integrate visual motion, sensory cues from other modalities and efferent motor signals to create invariant spatial representations which may then be used to guide actions (Pouget and Sejnowski, 1997).

Consistent with this idea, parietal visual areas have strong reciprocal connections to premotor and motor areas (Andersen et al., 1990; Gallant et al., 1996), and lesions to parietal visual areas lead to complex behavioural deficits in guiding action and attention, in particular visual hemi-neglect (Kandel et al., 2000).

1.4 Hierarchical visual processing at multiple scales

Our discussion so far has focused on visual processing at the level of single neurons. Yet visual and other cortical areas are organized along multiple, interconnecting scales. Studying these various levels of organizations gives insights about how single neurons processing visual stimuli as part of a larger network of activity.

Neurons are organized spatially along the cortical surface according to strict rules which relate to their function. A common organization principle in early and intermediate visual cortex is that of retinotopy: neurons with similar receptive field positions are clustered at similar locations on the cortical surface (Das and Gilbert, 1997; Engel et al., 1997; Hubel and Wiesel, 1962). Furthermore, retinotopy in early and intermediate visual areas varies smoothly within an area along a gradient of preferred eccentricity and an orthogonal gradient of preferred angle (Dumoulin and Wandell, 2008; Engel et al., 1997).

Other maps coexist with the retinotopic map within the same area. A prominent example of this is the orientation map in V1: neurons with similar preferred orientations are clustered spatially within columns of limited spatial extent, roughly 250 µm (Das and Gilbert, 1997; Kaschube et al., 2010; Vanduffel et al., 2002). Preferred orientation varies smoothly as a function of cortical position in primary visual cortex except at pinwheel centers, where multiple orientation columns representing different orientations meet (Nauhaus et al., 2008).

Topographic organization allows neurons with similar receptive field properties to be clustered together, which has been hypothesized to facilitate developmental processes and make optimal use of limited cortical space (Chklovskii and Koulakov, 2004). In particular, neurons with similar orientation and position preferences preferentially make corticocortical synapses to each other (Gilbert and Wiesel, 1983); this particular organization of orientation and retinotopy has been shown in simulations to minimize the wiring length of corticocortical connections (Koulakov and Chklovskii, 2001). Furthermore, retinotopic organization of LGN inputs can create an orientation bias in V1 neurons which can be refined through visual experience (Paik and Ringach, 2012; Ringach, 2004).

Topographic organization can be studied with methods which look at larger spatial windows than single electrode recordings, including multi-electrode array (MEA) recordings, intrinsic signal optical imaging, and functional magnetic resonance imaging (fMRI).

Multi-electrode arrays consist of multiple regularly spaced electrodes which record voltage signals. One-dimensional configurations with multiple electrode contacts along the z-dimension have proven useful in examining laminar organization of cortical processing (Xing et al., 2009). Two-dimensional configurations, in particular the Utah array (Maynard et al., 1997; Rousche and Normann, 1998), where multiple electrodes are organized along a regular lattice spanning a few millimetres of cortex, are useful for examining topographic organization (Rousche et al., 1999).

1.4.1 Multi-unit activity

With fixed MEAs, it becomes difficult to isolate spikes from single neurons in each electrode, as electrodes cannot easily be moved closer to neuron bodies (Gray et al., 1995). Multi-unit activity (MUA), which measures the density of threshold crossings of the voltage signals measured in each electrode, can be measured in cases where isolation is problematic. MUA reflects the density of action potentials in a small volume around the electrode (~100 microns), typically pooling the action potentials of a small number of neurons (Buzsáki, 2004). As such, it is generally straightforward to interpret and to relate to single neuron activity.

1.4.2 Local field potentials

A particularly useful signal in the study of cortical maps is the local field potential (LFP). Neurophysiology experiments commonly measure action potentials from single neurons or MUA via electrodes. These electrodes measure differences in electric potential between the electrode tip and a reference position – a voltage trace. While action potentials are fast, stereotyped events which generate a characteristic, high-frequency trace lasting a few milliseconds, the voltage trace also measures slower oscillations with a characteristic 1/f power spectrum (Bedard et al., 2006).

This is the local field potential (LFP), a measure of the electric potential in the extracellular space (Buzsáki et al., 2012). Local field potentials can easily be recorded with multi-electrode arrays, even under conditions where measuring single neuron action potentials or MUA is problematic. This makes it a valuable tool for probing topographic organization (Katzner et al., 2009; Mineault et al., 2013).

The LFP reflects a multitude of sources, the largest of which is correlated synaptic activity (Buzsáki et al., 2012; Einevoll et al., 2013). In conditions where both the local field potential and spikes can be reliably

measured, the LFP offers the potential to relate presynaptic activity – the input to neurons, measured on a coarse spatial scale - to the spiking activity of the same neurons – their output.

For example, testing the selectivity of V1, MT and MST motion-sensitive single units and LFPs to plaids, Khawaja et al. (2009) found that selectivity to pattern motion increased along the hierarchy for both signals. However, local field potentials were less selective for pattern motion than single units in the same area; rather, the selectivity of LFPs in a given area was well-matched to the selectivity of single units in the preceding area. Thus, the LFP gives precious information about the mean synaptic input underlying single neuron activity, valuable insight in the study of hierarchical processing.

Furthermore, the LFP can offer insight on lateral processing of visual inputs, which is difficult in traditional single-electrode experimental paradigms. For example, Nauhaus et al. (2009) found that spikes in primary visual cortex trigger a slowly travelling wave of activity, as measured by the spike-triggered average of the local field potential. The spread of this activity is modulated by stimulus contrast: at low contrast, the activity of a spike spreads further than during high contrast stimulation.

This suggests an interesting refinement of the Hubel and Wiesel canonical circuit: the relative strength of the feedforward and lateral contributions to simple and complex cell tuning is modulated by the contrast of the stimulus. Under low contrast conditions, the lateral contribution is larger, thus amplifying weak signals by increased spatial integration. Thus, the local field potential can offer interesting insight into hierarchical processing.

The use of local field potentials in the study of visual and cortical processing in general is however limited by its heterogeneous origin (Buzsáki et al., 2012). When LFPs are recorded on the same electrode as action potentials, as is often the case, any measure of correlation between spikes and LFPs can be exaggerated by spurious action potential traces in the voltage signal. Indeed, under common analysis scenarios, LFP-spike association metrics can often be exaggerated by a factor of 50% or more (Zanos et al., 2011a). This signal processing artifact can be corrected by using a Bayesian generative model to separate out action potential and local field potential contributions to the voltage signal; I present simulations and the corrective signal processing algorithm in the appendix (Zanos et al., 2011a).

With this technical issue resolved, in Chapter 3, I examine at the representation of visual space in intermediate visual area V4 at the multi-unit and local field potential levels (Mineault et al., 2013). As low-level and intermediate visual areas form coherent maps of visual space, I use this opportunity to estimate the spatial scale of the integration of the local field potential.

Indeed, to infer the effect of correlated subthreshold membrane fluctuations on visual processing, it is essential to understand the spatial extent of the electrical activity reflected in the LFP. Estimates of the integration radius of the LFP have been widely divergent, ranging from 250 microns to several millimeters (Katzner et al., 2009; Kreiman et al., 2006; Xing et al., 2009). Using parametric modeling methods covered later in the introduction, I show that the local field potential reflects both local and global sources of input, and that analytical choices can selectively enhance either one of these sources (Mineault et al., 2013). This resolves the divergence in earlier estimates, and shows that the local field potential can be used to probe both local visual representation and large-scale biases in visual representation.

1.4.3 Visual representations in humans - fMRI

Much of our knowledge of the human visual system is based on invasive studies of other primates, in particular macaques. Invasive neurophysiological recordings in humans have been limited to cases where brain activity must be monitored intracranially for medical reasons, in particular in epileptic patients (Quiroga et al., 2005). Yet, the non-invasive study of human vision offers significant opportunities for understanding cognitive aspects of vision, thus completing the picture of visual processing at single neuron, neural ensemble, cortical area, and cognitive scales.

Whole brain activity recordings can be performed by functional magnetic resonance imaging (fMRI). fMRI measures a blood-oxygen-level dependent (BOLD) signal which relates to the energy consumption, and ultimately the activity, of coarsely sampled volumes of neurons (Logothetis et al., 2001). fMRI has been widely used to probe visual processing (Courtney and Ungerleider, 1997), offering insight into the representation of visual stimuli at large scales in human subjects (Huth et al., 2012; Mineault and Pack, 2013).

One approach which has proven particularly useful to understand visual representations with fMRI is to reconstruct visual information from the BOLD signal in order to determine which aspects of the stimulus are represented in a given area of the brain. It has been shown that it is possible to decode image identity (Kamitani and Tong, 2005), category (Norman et al., 2006), and even the broad spatial structure of a visual stimulus (Naselaris et al., 2011) from the BOLD signal. This has been used to reveal how complex information is encoded and organized at broad spatial scales (Huth et al., 2012; Stansbury et al., 2013).

1.4.4 Psychophysics

In many experimental paradigms, subjects are asked to perform a psychophysical visual task while fMRI is performed, thus allowing the experimenter to relate neural activity to behaviour. Psychophysics is a broad discipline with a rich history predating the advent of modern neurophysiology by a century (Fechner, 1860). It seeks to establish relationship between visual stimulation and behavioural output.

With human psychophysics, it is feasible to probe the interactions between a stimulus, low-level sensory mechanisms and behaviour in cognitively complex tasks. For example, in Neri & Heeger (2002), the experimenters asked subjects to detect a flashed bar in spatiotemporally varying noise. Such an experimental paradigm probes both low-level sensory mechanisms and higher-level cognition. Detection of orientation energy is ultimately dependent on low-level sensory mechanisms, i.e. orientation detectors in primary visual cortex. However, the detection of a stimulus is also dependent on the computation of saliency, a higher-level visual process by which interesting stimuli are detected, for example to guide eye movements.

Indeed, the authors found evidence of both mechanisms at work in this task: the observers used a detection rule where the luminance of a bar was compared to the luminance of surrounding regions, in a fashion consistent with the operation of simple and complex cells. However, they also found that observers were most likely to detect a bar when it was preceded with high-contrast stimuli, indicating a role of saliency in driving behavioural detection. Thus, carefully designed psychophysical experiments can uncover the link between visual representations and behavioural output in cognitively complex tasks (Neri and Levi, 2006).

1.5 Systems identification overview

We have seen that in both the dorsal and ventral visual streams, a hierarchy of computations reencodes luminance information into ever more abstract representations which can support complex behaviours. In order to understand how these abstract representations are formed, then, we need to understand how the visual system, whether defined at the level of a single neuron, an ensemble of neurons, or a psychophysical observer, is related to its input. This is the subject of systems identification.

Systems identification seeks to establish a deterministic relationship between the input to a system and its output (Marmarelis and Marmarelis, 1978). Systems identification techniques are particularly useful in sensory neuroscience, where the input to the system can easily be controlled via sensory stimulation.

A classic example of neuronal systems identification (De Boer and Kuyper, 1968) illustrates an experimental paradigm that remains a fundamental building block for contemporary studies. The experimenters were interested in determining the stimulus selectivity properties of neurons in the cochlear nerve of the cat. To establish this relationship, they measured the spiking output of neurons with an electrode in the cochlear nerve while presenting, via an earphone, an experimenter-controlled stimulus – in this case, white noise.

The authors assumed that the relationship between neuronal spiking and the auditory input was linear. Mathematically, the input, codified as a vector $\mathbf{x}$, was assumed to be linearly related to the measured output of the neuron y via projection onto a weight vector $\mathbf{w}$:

$$y \propto \mathbf{x}^T \mathbf{w}$$

By taking the average of stimuli that preceded a spike – the spike-triggered average (STA) - the experimenters obtained an estimate of the weight vector $\mathbf{w}$. The STA provides an unbiased estimate of the linear filter, provided that the noise is uncorrelated in time and follows a Gaussian distribution (Schwartz et al., 2006). This method of estimating the linear filter, feasible with little else than a triggering circuit and an oscilloscope, was originally suggested by Wiener (Wiener, 1966).

The obtained reverse correlation filters revealed a rich, heretofore unsuspected structure. Filters showed numerous oscillatory modulations within in an asymmetric envelope – well-captured by gammatone functions (Johannesma, 1972) – consistent with causal, narrowband frequency tuning in cochlear neurons.

1.5.1 Systems identification in single neurons

Systems identification has proven particularly valuable in the study of visual processing, where the visual input can easily be controlled via a monitor. It has frequently been used to quantify the receptive fields of visual neurons, including in the retina (Chichilnisky, 2001; Marmarelis and Naka, 1973), in the LGN (Reid and Shapley, 1992), in primary visual cortex (DeAngelis et al., 1993; Jones and Palmer, 1987; Pack et al., 2003a; Ringach et al., 1997; Rust et al., 2005), and in area MT (Livingstone et al., 2001; Pack et al., 2003b).

Much of our mechanistic understanding of low-level visual computations such as direction selectivity can thus be traced to systems identification. Many important properties of single-neuron receptive fields are difficult to identify using traditional techniques such as stimulating receptive fields with a small

number of parametric stimuli, such as bars and gratings. Using a large set of stimuli and a model-based approach to relate the visual input to neuronal output allows one to determine a neuron's preferred stimulus with minimal assumptions (Wu et al., 2006).

In addition to determining a system's preferred stimulus, systems identification can predict its responses to non-preferred stimuli (Mineault et al., 2012), and identify its invariances to energy and identity-preserving transformations (Berkes and Wiskott, 2007). This information can be used to infer the types of computations that the system performs (Cui et al., 2013) as well as identify candidate mechanisms by which it can implement its function (Elyada et al., 2009).

An illustrative example of the use of systems identification to elucidate a neural circuit – the simple-complex circuit of primary visual cortex - is given by the work of Rust et al. (Rust et al., 2005). Probing direction and non-direction selective V1 neurons with flashed bars, systems identification revealed an unexpected diversity of filtering properties in receptive fields of both simple and complex cells.

Later work showed that this could be explained computationally by spatially tiling the receptive fields of V1 neurons with multiple, localized subunits (Lochmann et al., 2013). Typically, these subunits had similar frequency selectivity and spatial envelope, but differed in their preferred spatial position (Figure 1-2A). Complex cells differed from simple cells in that they integrated from subunits tuned to both contrast polarities (Figure 1-2B).

This example shows how systems identification can further refine our understanding of a neural circuit: both simple cells and complex cells are created by the tiling of multiple, similarly-shaped subunits. In the case of simple cells, in particular, this is computational evidence for a role of cortico-cortical connections in creating heretofore unsuspected invariance to translation and spatial frequency (Somers et al., 1995).

1.5.2 Systems identification can elucidate neural representations

We are often interested in understanding the link between single neuron computation and the representation of the visual stimulus supported by multiple neurons. Given a set of systems identification models for multiple neurons in the same area, decoding simulations, which link visual representation to task performance, can help identify the functional significance of a given computation. In the framework of (DiCarlo and Cox, 2007), hierarchical computation reencodes behaviourally relevant information to a format that is easily readable for a higher level neuron, e.g. via a linear decoding rule.

Given model fits from multiple neurons in the same area, then, decoding simulations can reveal the extent to which neurons carry readily readable information about various aspects of the stimulus, from which we can form hypotheses about the types of tasks the set of neurons may be involved in (Butts et al., 2007; Cui et al., 2013; Mineault et al., 2012). These hypotheses may in turn be used to guide more technically demanding experiments to establish causal relationships between a given set of neurons and the performance of a task.

A complementary approach is to apply systems identification at larger scales to study neuronal ensembles. Although originally formulated for the study of single neurons (De Boer and Kuyper, 1968; Marmarelis and Marmarelis, 1978), systems identification is applicable at a number of scales. We can leverage signals at larger spatial scales, like LFPs and fMRI, to understand how stimuli are represented at the neuronal ensemble level (Victor, 2005). I show in Chapter 3 how the local field potential can be characterized in a systems identification paradigm to elucidate the representation of visual space in area V4 (Mineault et al. 2013).

Evidently, we are not limited, in a single study, to the analysis of signals at a single scale. We can also consider cross-scale analyses to elucidate how large-scale representations emerge from single neuron activity (Nauhaus et al., 2009), or how ensemble activity influences single-neuron behaviour (Rasch et al., 2008).

Perhaps most surprisingly, it is possible to use systems identification to directly probe visually-guided behaviours in psychophysical observers (Ahumada and Lovell, 1971; Eckstein and Ahumada, 2002; Neri and Levi, 2006). Here, the observer is asked to perform a visual task, typically detection or discrimination. The task is made more difficult by the addition of noise to the visual stimulus. The task is usually defined such that there is an objectively optimal strategy to perform it. For example, to detect the presence or absence of a target corrupted by additive Gaussian noise, the optimal strategy is to compare the visual stimulus linearly with a template equal to the target and make a yes/no decision depending on the magnitude of the match (Bishop, 2006; Green and Swets, 1966).

By correlating the visual stimulus, including the noise, to the behavioural response of the observer, it is possible to infer the strategy that the observer is using to perform the task; this is revealed in the classification image (Eckstein and Ahumada, 2002), an estimate of the internal template used by the observer to perform the task, on the assumption that the observer uses a linear matching strategy.

A real observer is rarely optimal in performing such a task, and classification images reveal the divergence between the optimal strategy and the observer's strategy. This is reflective of the underlying neural computations which drive the behaviour. In detection tasks, for example, the classification image is typically a blurry version of the stimulus, which reflects the spatial uncertainty of the observer (Tjan and Nandy, 2006); this likely reflects the fact that the neuronal representations used to perform the tasks, i.e. simple and complex cells, are partially translation-invariant.

Indeed, in appropriate circumstances, classification images can be compared meaningfully to data obtained from single-unit recordings (Neri and Levi 2006). The classification image approach has been used widely in psychophysics, including in studies of Vernier acuity (Ahumada Jr, 1996), disparity processing (Neri et al., 1999), motion perception (Neri and Levi, 2008b), object discrimination (Olman and Kersten, 2004), and face recognition (Sekuler et al., 2004).

Systems identification, applied at a variety of different scales, is thus a powerful engine of discovery in sensory neuroscience.

1.5.3 Limitations of classical systems identification

Classical systems identification in the Wiener-Volterra framework assumes that the response of a system is related to its input by a series of kernels of increasing degree (Marmarelis & Marmarelis 1978):

$$y \propto c + \sum_i x_i w_i + \sum_{i,j} x_i x_j v_{i,j} + \sum_{i,j,k} x_i x_j x_k u_{i,j,k} + \cdots$$

This polynomial formulation, which is related to the Taylor expansion, can capture arbitrarily nonlinear computations with the addition of kernels of high degree. Many computations of interest are not well-described by a low-dimensional polynomial. For example, a simple linear-nonlinear system, $y = f(\mathbf{x}^T\mathbf{w})$, where f is, for example, a sigmoid nonlinearity, generates non-negligible high-order kernels:

$$y \propto f(0) + f'(0)\sum_i x_i w_i + \frac{1}{2}f''(0)\sum_i x_i x_i w_i w_i + \frac{1}{3!}f'''(0)\sum_{i,j,k} x_i x_i x_i w_i w_i w_i + \cdots$$

While this nonlinearity is relatively trivial, such nonlinearities accumulate along the visual hierarchy. That is, the sensory input is reformatted into a more abstract representation by the repeated application of selectivity and invariance operations in the manner of the Hubel & Wiesel canonical circuit. These accumulated processing stages are not well-captured by low-order polynomials, and they thus obscure the relationship between the input and the output of the system.

The number of parameters in a polynomial formulation increases exponentially with the degree of the kernel considered. Thus, as we encounter more nonlinear systems, as will occur when we probe higher level areas of the visual hierarchy, the parameters of our systems identification model will become more poorly constrained. In practice, the degree of the kernel considered will be limited to one or, at most, two, i.e. linear or quadratic systems (Wu et al., 2006), limiting our ability to accurately characterize higher level areas.

Even when the system under study can be approximated with low-order polynomials, we may not be able to constrain the parameters of the transfer function of such a system if the parameters are too numerous. Constraining these systems requires either performing longer experiments, which may be infeasible, or adding soft constraints to the parameters, which is not straightforward in the Wiener-Volterra framework.

Furthermore, in the Wiener-Volterra framework, the stimulus-response relationship is estimated by moment-based methods, i.e. spike-triggered averaging and spike-triggered covariance (Schwartz et al., 2006). Unbiased estimation using moments-based methods requires the use of highly constrained stimulus ensembles, i.e. white noise (Schwartz et al., 2006; although see Park et al., 2013). At high levels of the visual hierarchy, accumulated thresholding and normalization nonlinearities can prevent the system from responding appreciably to white noise .

Thus, our ability to constrain the parameters of systems which are highly nonlinear and poorly driven by white noise is limited by the constraints of classical systems identification methods. These technical issues fundamentally limit our understanding of high-level visual processing in general to phenomenological, as opposed to mechanistic descriptions.

1.6 Parametric modeling – technical aspects

In recent years, however, a variety of techniques inspired by machine learning and Bayesian statistics has been applied to great success in the context of identification of sensory systems, and in particular vision (Wu et al. 2006). These techniques can be thought of as refinements of classic Wiener-Volterra analysis: however, they can work with stimuli with arbitrary statistics, they are not limited to linear systems, and they can incorporate prior information about model parameters and the structure of the visual system. They are thus applicable to the study of complex nonlinear systems, e.g. high-level visual processing.

I start by introducing the generalized linear model, a flexible statistical model which describes a system as a linear filter acting on the stimulus followed by a static nonlinearity and non-Gaussian noise. Such a framework, which offers a unified description of signal transduction in neurons, neural ensembles and psychophysical observers, is an appropriate building block for neural systems identification.

I then describe how to extend the generalized linear model to account for nonlinear systems, by considering nonlinear representations of the stimulus. This will enable us to model the input-output relationship of sensory systems which are far removed from the input of the system, in particular high-level visual processes.

Finally, I show how the numerous parameters of the resulting models can be efficiently estimated by incorporating hard and soft constraints on the model parameters. The resulting systems identification framework, which I refer to as parametric modelling, is appropriate for determining the nonlinear relationship between a stimulus and the output of a system, whether that system is a neuron, an ensemble of neurons, or the behavioural response of a psychophysical observer.

1.6.1 The generalized linear model

Generalized linear models (GLMs; MacCullagh and Nelder, 1989) are a class of statistical models which offer a unified treatment of regression, classification, and systems identification. The GLM assumes that a system's response is given by a weighted sum of the stimulus, which is then nonlinearly transduced and corrupted by non-Gaussian noise.

Assuming that the input to the system is codified as a vector $\mathbf{x}$, the system's output is y, and the weight vector $\mathbf{w}$, we have:

$$\eta = \mathbf{x}^T\mathbf{w}$$
$$\mu = f(\eta)$$
$$y \sim Distribution(\text{mean} = \mu)$$

Here, f - referred to as the inverse link in statistical literature - is the static nonlinearity, and the distribution can be any distribution in the exponential family (MacCullagh and Nelder, 1989). These two simple modifications can be used to extend the linear model to situations where its strong assumptions – linearity and normally distributed noise - are inappropriate.

For example, a simple cell in V1 can be modeled as matching the luminance of the stimulus to an internal template (Carandini et al., 2005). In this sense, it is approximately linear. However, a neuron

cannot have negative firing rates; the number of spikes in a bin is an integer; and the variance of the firing rate of a neuron scales in proportion to its mean firing rate (Dayan et al., 2001). A more realistic approximation of a simple cell is given if we assume that its linear match is followed by a static, rectifying nonlinearity which drives a Poisson process (Carandini et al., 2005).

The linear weights of an estimated GLM are comparable to the linear receptive field derived by spike-triggered-averaging (Chichilnisky 2001; Simoncelli et al. 2004). The GLM method is applicable, however, regardless of the statistics of the stimulus (Paninski, 2004). It is also more efficient, both because the nonlinearity and non-Gaussian noise are a better reflection of the type of systems encountered in sensory neuroscience and because it allows for the specification of prior information which can reduce the variance in the receptive field estimate (Paninski, 2004).

Distributions and nonlinearities of particular interest in systems identification are:

- the normal distribution – useful for modelling continuous signals, like local field potentials and fMRI. In this case, f is often chosen to be the identity function, and the model reduces to the standard linear model. I will use this distribution for modelling local field potentials in Chapter 3.
- the binomial distribution – useful for modelling binary outcomes, like decisions in psychophysics, or spike trains at a high temporal resolution. f is often chosen to be the sigmoid or Gaussian integral function in this context. I use this distribution for modelling psychophysical observers in Chapter 2.
- the Poisson distribution – useful for modelling nonnegative count data, like spike trains at a low temporal resolution. In this case, a rectifying nonlinearity like the exponential is often chosen. I use this distribution for modelling spike trains in Chapter 4.

1.6.2 More flexible model forms

As is, the generalized linear model cannot account for highly nonlinear systems. A flexible method for resolving this issue it to consider a nonlinear projection of the stimulus (Bishop, 2006):

$$\mathbf{x}' = g(\mathbf{x})$$

Here g is a fixed multidimensional nonlinearity $g: R^M \rightarrow R^N$ which projects the stimulus from its original, e.g. luminance-based representation to another representation in which the relationship between input and stimulus is more linear. The model still takes the form of a generalized linear model in the augmented space. Typically, the new representation will be higher-dimensional than the initial representation; it will be important to further constrain the parameters with the methods outlined in the next sections.

Several choices of nonlinear transformations are possible. When studying a high-level visual area, g can be chosen to approximate the representation of the stimulus in the previous area of the hierarchy. I use this strategy in Chapter 4; to infer the processing of MST neurons, the visual input is represented in a basis approximating the processing of MT neurons. Such a strategy is feasible because MT neurons have been carefully studied quantitatively (Mineault et al., 2012).

In this context, it is possible to consider a restricted set of alternative, biologically plausible nonlinear representations. In chapter 4, in particular, I consider different MT-like representations, including or excluding normalization, center-surround interactions, and input nonlinearities. Using a validation strategy to determine which transformation best captures the response of the neurons under study, as I will discuss in the next sections, it is possible to reveal that the aspects of previous areas' processing which are critical or marginal in explaining the responses of a higher-level area.

Other choices of representation trade off biological plausibility for computational insight. In particular, one may project a stimulus into the products of its component dimensions to account for quadratic nonlinearities: $g(\mathbf{x}) = \mathbf{x}\mathbf{x}^T$. The resulting model, often termed the generalized quadratic model (Rajan et al., 2012), can be viewed as a generalization of second-order Wiener-Volterra models. This model is also closely related to the spike-triggered covariance (STC) technique (Pillow and Park, 2011). Since this type of projection can account for arbitrary quadratic nonlinearities, it is useful, for example, to characterize complex cells in primary visual cortex (Rust et al., 2005).

Finally, element-wise nonlinearities can be used to model input nonlinearities (Ahrens et al., 2008a):

$$g(\mathbf{x}) = h_j(x_i)$$

Common choices for h include half-wave or full wave rectification, squaring, sigmoids, etc. When the correct element-wise nonlinearity is unknown a priori, h can be chosen to be a localized basis, in particular a spline. In this case, the nonlinearity is known only up to several free coefficients which determine the shape of the nonlinearity applied to each stimulus dimension. Model inference then determines both the main model weights – the receptive field envelope – and the input nonlinearities themselves (Ahrens et al., 2008a). In the statistical literature, this parameterization is known as the generalized additive model (Hastie and Tibshirani, 1990; Wood, 2006).

1.6.3 Inferring model parameters with priors

The parameters of a GLM can be estimated by Bayesian inference. Given a vector of observations $\mathbf{y}$, the posterior probability of the model parameters $p(\mathbf{w}|\mathbf{y})$ is given by Bayes' rule:

$$p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$$

Here, $p(\mathbf{y}|\mathbf{w})$ is the model likelihood, which is derived from the noise distribution. For example, for a choice of Poisson noise, we have:

$$p(\mathbf{y}|\mathbf{w}) = \prod_i \frac{f((\mathbf{X}\mathbf{w})_i)^{y_i} \exp\left(-f((\mathbf{X}\mathbf{w})_i)\right)}{y_i!}$$

$p(\mathbf{w})$, on the other hand, is the prior probability of the weights, which may be derived from assumptions about the magnitude of the weights, their smoothness, sparseness, or locality, for example (Bishop 2006). In the next few sections, I will consider priors of the form:

$$p(\mathbf{w}) = \exp\left(-\lambda L(\mathbf{w})\right)$$

Under this choice of parameterization, an estimate of the model weights can then be derived from the model posterior by finding a set of weights which are highly likely according to the model posterior (Bishop 2006). In maximum a posteriori (MAP) inference, we find the mode of the posterior – its peak. The MAP estimate of the weights has excellent statistical guarantees in terms of variance and bias (Paninski 2004), and is straightforward to compute.

To find the MAP, the posterior is maximized numerically. Algorithmically, maximizing the posterior is equivalent to minimizing the negative-log of the posterior; in optimization literature, this is known as the *error function* to be minimized (Bishop 2006). GLMs have strong theoretical guarantees that make them well-suited for MAP estimation via gradient-based optimization. In particular, provided that the static nonlinearity is chosen according to a certain set of rules, the error function to be minimized is convex everywhere and has a unique minimum (Paninski 2004).

The error function, in the case Poisson noise, is given by:

$$-\log\left(p(\mathbf{w}|\mathbf{y})\right) \equiv E(\mathbf{w}) = \sum_i \left(-y_i \log f\left(X_{ij}w_j\right) + f\left(X_{ij}w_j\right)\right) + \lambda L(\mathbf{w})$$

This equation shows that minimizing the error is equivalent to finding an optimal balance between matching the model predictions to the data (first term) and matching the weights to prior expectations (second term). In the limit of infinite data, the prior term becomes negligible, and MAP inference reduces to maximum likelihood – i.e. purely data-based – inference (Gelman et al., 2003).

1.6.4 Families of priors

Two families of model priors have proven particularly useful in neuroscience applications. The first is the normal family, where the model weights are assumed to be taken from a normal distribution:

$$p(\mathbf{w}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The prior mean $\boldsymbol{\mu}$ is generally taken to be 0. $\boldsymbol{\Sigma}$, the prior covariance, can be set to a variety of different matrices to express different beliefs about the model parameters.

In the absence of structural information about the model weights, we may constrain them to be of low magnitude. This is expressed in the choice $\boldsymbol{\Sigma}^{-1} = \lambda \mathbf{I}$, which is often referred to as a weight-decay prior or Tikhonov regularization (Bishop 2006). When the design matrix $\mathbf{X}$ is well-conditioned, the weight-decay prior lowers the magnitude of all weights equally. However, when the design matrix is poorly conditioned, as is the case when using stimuli with strong low-frequency content - natural images, especially – a weight-decay prior will have the effect of enhancing the better constrained directions in probability space. In the case of natural images, this will enhance low frequencies, thus smoothing the weights (Wu et al. 2006).

Alternatively, the weights may be directly constrained to be smooth via the choice $\boldsymbol{\Sigma}^{-1} = \lambda \mathbf{D}'\mathbf{D}$. Here D is a matrix which computes the spatial or temporal derivatives of the weights. This has the effect of smoothing the weights, which is frequently warranted when estimating smooth receptive fields (Wu et al. 2006).

A second class of prior is given by the Laplacian, or L1 family of priors:

$$p(\mathbf{w}) \propto \exp\left(-\sum_j \lambda |w_j|\right)$$

Laplacian priors penalize weights proportional to their magnitude, rather than their squared magnitude as is the case for normal priors. As a consequence, small weights are relatively more penalized and large weights less penalized when compared to a normal prior. The MAP estimate under a Laplacian prior –

often termed the LASSO estimate (Tibshirani, 1996) – will have many weights which are exactly 0, as well as a few large weights. Thus, MAP estimates under a Laplacian prior are sparse.

Sparseness is a strong prior which is often appropriate in neural systems identification (Wu et al. 2006). In many systems identification scenarios, only a few of the stimulus dimensions – pixels, in an experiment on vision –drive the system under study, while the rest are irrelevant to the system. I used such an assumption in Chapter 4, where MST receptive fields were assumed to integrate from a small number of MT subunits.

The Laplace prior can be adapted when strict sparseness is inappropriate. For example, in some scenarios, weights are naturally organized into non-overlapping groups, only some of which may be non-zero; group-L1 priors may be used here (Yuan and Lin, 2006).

In chapter 2, I show that it is possible to combine sparseness and smoothness assumptions meaningfully by constraining the weights to be sparse in a smooth basis. This assumption is often warranted when considering a receptive field which is spatially restricted and smooth within its bounds (Park and Pillow, 2011). I show that such an assumption can be used to effectively estimate psychophysical receptive fields with a limited number of trials.

1.6.5 Low-rank models

A complementary approach for constraining model parameters is the use of low-rank approximations. Frequently, model parameters are organized along logical dimensions; in the content of vision, model parameters may be organized along two spatial dimensions and a temporal dimension:

$$w_i \equiv w(x, y, t)$$

In these cases, one can often specify that the weights are of low rank with respect to these logical dimensions (Adelson and Bergen, 1985). In the temporal dimension, in particular, we can specify that:

$$w(x, y, t) \approx \sum_{i}^{n} u_i(x, y) v_i(t)$$

Here *n* is a small integer which specifies the assumed rank of the weight vector. With $n = 1$, we assume that the receptive field of the system under study is separable in space and time. Such a model is much lower-dimensional than the full-rank model. I used this approach in Chapter 4, constraining

receptive fields in MST to be separable in time, and most notably in Chapter 3, where I assumed 3-fold separability in time, space and orientation domains.

Strictly speaking, such low-rank models are no longer GLMs (Ahrens et al. 2008); estimation by gradient-based optimization is complicated by the multiple local minima in the negative log-posterior of the corresponding model. In practice however, good convergence is obtained by initializing the low-rank representation using the singular value decomposition of the full-rank model weights (Ahrens et al. 2008). Alternatively, low-rank can be encouraged by adding a soft penalty on the sum of the singular values of the model parameters, in which case convexity of the error function is preserved (Pfau et al., 2013).

1.6.6 Estimating model form

Frequently, parametric models will be determined up to a small number of hyperparameters, i.e. nuisance parameters that determine the form of the prior. In the examples above, the scale constant λ determines the relative strength of the prior and the model likelihood. A simple strategy to determine hyperparameters is *k*-fold cross-validation.

Its implementation is straightforward: the data is split into *k* non-overlapping subsets. The data from *k* -1 subsets is used to fit the model for different values of the hyperparameters; the data in the *k*'th subset is predicted based on these different fits. If the prior is too strong, predictions will be poor because the weights will not be adapted to the data; if the prior is too weak, predictions will also be poor, this time because the model overfits to noise.

Thus, cross-validation can yield an estimate of the optimal hyperparameters (Wu et al. 2006). Repeating this process for the *k* different assignments of fit and validation datasets reduces the variance in the estimate of the optimal hyperparameters (Hastie et al., 2001).

A similar strategy can be used to compare the ability of different systems identification models to account for the data. While the model log-likelihood is a measure of the quality of model fits, this goodness-of-fit value cannot be used to directly compare models which have unequal numbers of free parameters. Indeed, more flexible models can account for more variance in the data simply by overadapting to accidental correlations between stimulus and response, i.e. noise; this issue is called overfitting (Bishop 2006).

It is possible, however, to compare the ability of different models to predict the response of the system to stimuli not used during model fitting. This validation strategy provides accurate measures of goodness-of-fit regardless of the dimensionality of the models under consideration (Bishop 2006).

I used *k*-fold cross-validation to estimate hyperparameters, and validation to compare different model forms in the three manuscripts presented in the main text (Wu et al. 2006).

In other scenarios, especially when there are numerous hyperparameters to estimate, it is more practical to directly estimate the hyperparameters from the data through empirical Bayes estimation (Bishop 2006). Empirical Bayes estimation is much less straightforward to implement than validation and cross-validation strategies. However, it can require much less computation, and so is effective for highly-stereotyped, very high-dimensional models (Bishop 2006). I used such a strategy for estimating the hyperparameters of a generative model of the extracellular voltage signal in the manuscript presented in the appendix.

1.6.7 Derived information from fitted models

Given an estimated GLM that is well-fit to the data, we can predict the response of the system to an arbitrary stimulus by simulating its output given the estimated model weights. Furthermore, since the GLM is a generative model for the system under study, it is possible to estimate the most probable stimulus that generated a response, given a model for the probability of each stimulus $p(\mathbf{x})$:

$$p(\mathbf{x}|y = n) \propto p(y = n|x)p(\mathbf{x})$$

Hence, the GLM, being a fully-qualified generative encoding model, implicitly defines an optimal decoding model (Nishimoto et al., 2011). While this thesis is focused on the encoding problem – how neural systems encode stimuli – encoding and decoding are intimately related, and systems identification is thus a fundamental building block for neural decoding, e.g. in neural prosthetics applications.

We can also, as we discussed earlier, consider the ability of a non-optimal – i.e. linear - decoder to read information about the visual stimulus from the output of the system (DiCarlo and Cox 2007). I used such a strategy in Chapter 4 to compare the ability of different model formulations to convey visual information relevant to a behavioural task: the estimation of object velocity.

1.7 Parametric models can elucidate visual processing at multiple scales

Hierarchical visual processing iteratively reencodes retinal signals into abstract, easily decodable representations which can drive behaviour (DiCarlo and Cox 2007). To understand how this process is carried out computationally, I propose to develop and apply parametric models to the study of visual representations at multiple scales. As we have seen in the previous sections, recently developed statistical tools can infer the response function of a system to an arbitrary ensemble of stimuli.

Generalized linear models, in particular, allow data to be used more efficiently than traditional systems identification methods by explicitly including previously known facts about the system under study. Signals with different statistical properties – spike trains, local field potentials, psychophysical decisions - can be analyzed within a coherent and theoretically sound framework. Furthermore, by parameterizing nonlinearities judiciously, parametric models can be applied to the study of highly nonlinear systems. The proposed methodology promises to help elucidate visual representations at multiple scales, and thus the mechanisms underlying the complex transformations between visual input and behavioural output.

Our understanding of high-level visual processing is fundamentally limited by methodological hurdles. In the second chapter (Mineault et al. 2009), I develop a flexible methodology for estimating parameters of a generalized linear model with a high degree of sparseness. I show that this methodology is especially well-suited to estimate psychophysical classification images, with higher efficiency than previously proposed methods.

In an example application, I show that the methodology reveals clear functional signatures of neuronal processing in a signal detection task; specifically, I show that in a detection task, the estimated decision rule of the observer is consistent with the pooling of signals from a spatially invariant representation (Tjan and Nandy, 2006). The methodology, being built on generalized linear models, is equally applicable to the analysis of visual signals at the scale of single neurons, neural ensembles, and psychophysical observers.

Building on the parametric modelling methodology developed for psychophysics, I then examine the representation of visual space at the scale of neural ensembles in Chapter 3 (Mineault et al. 2013). Comparing the selectivity of local field potentials and spikes in area V4 with multi-electrode arrays, I find that the local field potential, a reflection of subthreshold membrane fluctuations due to synaptic activity, is well tuned for space. The relationship between spikes and local field potential receptive fields

is however not constant as a function of time after stimulus onset; LFP receptive field retinotopy matches that of spikes only for a restricted set of time lags.

I explain this finding using a low-rank parametric model, where LFP receptive fields are assumed to be generated by the sum of two components: a component unique to each electrode, and another component that is shared across all electrodes. This model captures the apparent change in LFP receptive field properties as a function of time lag. The unique component matches the retinotopy of the MUA measured on the same electrode, and is thus of local origin. The shared component, which has distinct temporal and spatial properties, likely arises from large-scale biases in the representation of visual space at the level of V4.

These findings reconcile the discrepant estimates of the integration radius of the LFP: the local field potential reflects both local and global integration scales, and different signal processing methods can selectively emphasize either of these scales. Furthermore, the analysis shows that the local field potential can be a powerful tool to probe spatial representations in intermediate visual cortex, provided that the signal is carefully analyzed to separate out its components.

In Chapter 4, I then use the developed parametric modelling methodology to probe visual representations at the level of single neurons. Specifically, I will tackle the problem of estimating the receptive fields of single units in area MST (Mineault et al. 2012). As discussed previously, neurons in area MST of the dorsal visual stream are selective for wide-field, complex optic flow stimuli, a property that has eluded mechanistic description.

While computations in MST are poorly understood, a great deal is known about the computations performed in the previous area of the dorsal stream, area MT. Thus, MST receptive fields will be assumed to integrate from the outputs of an initial processing stage that mimics the processing of MT neurons. Using the methodology developed in Chapter 2 and other parametric modeling tools, I show that MST neurons can be explained by the integration of MT neuron output, provided that the integration mechanism is nonlinear.

I then show in decoding simulations that this nonlinear integration mechanism helps MST signal, in a position-invariant fashion, the velocity of objects heading towards the observer. The uncovered computation may thus have a functionally significant role in behavioural responses to approaching objects, i.e. in the control of vergence.

Put together, these results show that visual computations can be successfully characterized at multiple scales, from the single neuron, to multi-neuron, all the way to the psychophysical level, provided that sufficiently powerful statistical methods are used. I discuss the larger implications of this body of work in the discussion.

1.8 Figures

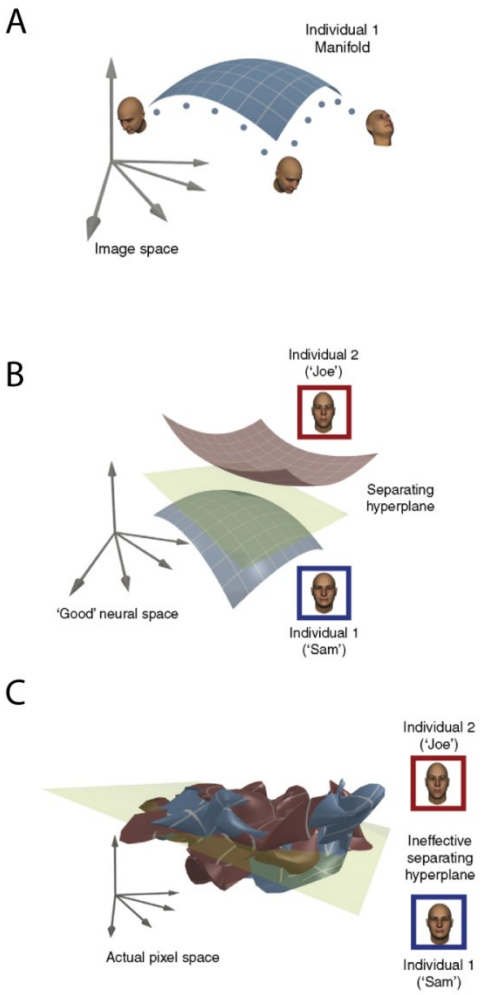


Figure 1-1: The untangling hypothesis.

Adapted from DiCarlo and Cox (2007), with permission from Elsevier (license #3321100220831).

A – Different identity preserving transformations – in this case, three-dimensional rotations – cause an object to trace out a manifold – a continuous cloud of points – in the image space.

B – The hypothesized goal of visual encoding is to form a neural representation in which manifolds corresponding to different objects are untangled, or separable using linear decoding rules.

C – Actual manifolds traced out by two different simulated objects under three-dimensional rotation. In the original, i.e. pixel-based representation, object manifolds are tangled. Untangling these manifolds is a non-trivial task that the hierarchical structure of cortex may be able to solve.

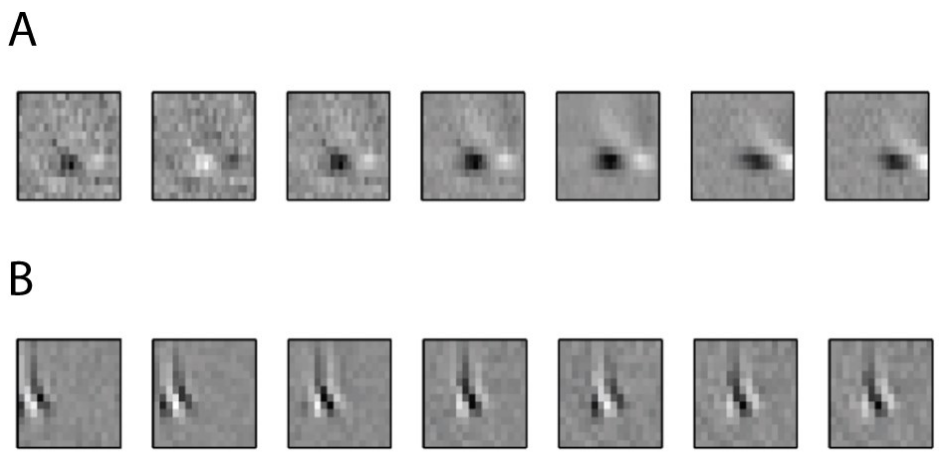


Figure 1-2: Simple and complex cell receptive field subunits.

Adapted from Lochmann et al. (2013), under a Creative Commons (CC BY 3.0) license.

A – receptive field subunits estimated from of a simple cell in macaque visual cortex. The x axis of each subunit represents space, and the y axis represents time lag. A simple cell is created by spatially tiling multiple orientation-tuned subunits with the same preferred polarity: black on the left, white on the right.

B – receptive field subunits of a directionally-selective complex cell. This complex cell, selective for motion towards the right, is created by tiling its receptive field with multiple subunits tuned to the similar orientations, spatial frequency, and direction, but different preferred spatial phase and contrast polarity.

Chapter 2 develops and applies the parametric modeling framework used through this thesis in the context of psychophysical classification images. Results show that parametric modeling can infer a psychophysical observer's decision process with more accuracy and fewer trials than previously proposed methods. This allows the exploration of more complex models of decision processes while retaining statistical tractability. I show in an example application in the context of a detection task that the resulting classification images show clear signatures that the observer used a spatially invariant detection mechanism, likely driven by a population of invariant neurons, i.e. simple and complex cells. This work was originally published in the Journal of Vision (Mineault et al. 2009). Appendix A, which accompanies this chapter and furnishes technical details about parametric modeling methods and interpretation, was originally published as the appendix of Mineault et al. (2009).

2. Improved classification images with sparse priors in a smooth basis

Classification images provide compelling insight into the strategies used by observers in psychophysical tasks. However, because of the high-dimensional nature of classification images and the limited quantity of trials that can practically be performed, classification images are often too noisy to be useful unless denoising strategies are adopted. Here we propose a method of estimating classification images by the use of sparse priors in smooth bases and generalized linear models (GLMs). Sparse priors in a smooth basis are used to impose assumptions about the simplicity of observers' internal templates, and they naturally generalize commonly used methods such as smoothing and thresholding. The use of GLMs in this context provides a number of advantages over classic estimation techniques, including the possibility of using stimuli with non-Gaussian statistics, such as natural textures. Using simulations, we show that our method recovers classification images that are typically less noisy and more accurate for a smaller number of trials than previously published techniques. Finally, we have verified the efficiency and accuracy of our approach with psychophysical data from a human observer.

2.1 Introduction

In recent years, the classification image approach has emerged as a powerful method of probing observers' strategies during psychophysical tasks. In a typical experiment, the observer is asked to indicate the presence or absence of a target signal masked with additive noise (Figure 2-1A). The resulting data are then evaluated under the assumption that the observer performs the task by linearly correlating the signal with an internal template, responding “target present” when the result exceeds a criterion, and “target absent” otherwise (Figure 2-1B). In this context, one can obtain an estimate of the observer's internal template by correlating the responses and the noise fields. The resulting *classification image* is a visual representation of the observer's strategy in the task.

The key advantage of the classification image approach over other psychophysical measures is that it is theoretically applicable when little is known about the stimulus parameters that are relevant in performing a task. The procedure is in this sense analogous to the reverse correlation technique commonly used in neurophysiology (Simoncelli, Pillow, Paninski, & Schwartz, 2004), and the results of classification image experiments can, in appropriate circumstances, be compared meaningfully to data obtained from single-unit recordings (Neri & Levi, 2006). Consequently, the approach has been used widely in psychophysics, including in studies of Vernier acuity (Ahumada, 1996), disparity processing

(Neri, Parker, & Blakemore, 1999), motion perception (Neri & Levi, 2008), object discrimination (Olman & Kersten, 2004), and face recognition (Sekuler, Gaspar, Gold, & Bennett, 2004).

2.1.1 Overcoming noise in classification images with prior assumptions

Despite the power and flexibility of the approach, the utility of classification images is limited by the amount of data that can be obtained in a given experimental task. For example, in tasks involving two spatial dimensions as well as time, even a modest stimulus resolution of 16 × 16 pixels and 16 time steps requires the estimation of over 4,000 parameters, which may require each observer to perform tens of thousands of trials. In practice, there is rarely sufficient data for the number of degrees of freedom the experimenter wishes to probe, and as a result classification images can be quite noisy. Such noise limits the interpretability and usefulness of the classification image.

One solution to this problem is to supplement observer data with prior information about the observer's strategy. Such prior information reduces the effective number of free parameters that must be estimated, leading to better estimates, assuming that the internal template conforms to the constraints. This approach has long been used implicitly in the form of post hoc smoothing and thresholding, under the assumptions that the observer's internal template is smooth and sparse, respectively. Explicit prior assumptions have recently been used successfully in classification image estimation (Knoblauch & Maloney, 2008a; Knoblauch & Maloney, 2008b; Ross & Cohen, 2009).

At first glance, the use of prior information may seem to be in contradiction with classification images' stated advantage of being applicable when little is known about the visual process being probed. However, while for a given task we may know little about the *specifics* of the visual process involved, we often have *general* exploitable knowledge derived from previous classification image experiments and from neurophysiology. To give but one example (Knoblauch & Maloney, 2008b; Thomas & Knoblauch, 2005), while we may not know exactly how a human observer detects a time-modulated luminance signal in noise, we do know that humans have limited sensitivity to high temporal frequencies. This suggests that observers' internal templates will be smooth at a certain scale for such a task. This prior knowledge of smoothness can then be used to obtain more accurate classification images on this particular task (Knoblauch & Maloney, 2008b).

Our goal in this paper is to define and explore the consequences of incorporating a powerful class of prior assumptions to estimate classification images. One approach to finding good prior assumptions is to attempt to find regularity within a set of observations, an approach that has been used with great

success in defining state-of-the-art image denoising techniques (Srivastava, Lee, Simoncelli, & Zhu, 2003; Portilla, Strela, Wainwright, & Simoncelli, 2003). An informal review of various articles (Abbey & Eckstein, 2002; Ahumada, 1996; Chauvin, Worsley, Schyns, Arguin, & Gosselin, 2005; Knoblauch & Maloney, 2008b; Levi & Klein, 2002; Mangini & Biederman, 2004; Neri & Levi, 2006, 2008; Neri et al., 1999; Sekuler et al., 2004; Tadin, Lappin, & Blake, 2006) that make use of the classification image technique shows that published classification images, regardless of exact protocol used, share a certain similarity: They appear to be well described by a small number of smooth features, such as lines and Gaussian blobs. In other words, many internal templates can be well described by a sparse sum of smooth basis functions, such as Gaussian blobs.

A second approach to finding good assumptions is to consider facts about the process *underlying* a set of observations. In the context of classification images, this means incorporating knowledge of visual physiology. The human visual system is trained on visual images which themselves are sparse in a basis of smooth, oriented filters (Srivastava et al., 2003). Conversely, a simple constraint of sparse representation of images is sufficient to reproduce many properties of the visual system, including V1 receptive fields (Olshausen & Field, 1996) and color opponency (Lee, Wachtler, & Sejnowski, 2002). It is thus tempting to conjecture that humans are naturally more efficient at representing sparse visual structure (Olshausen & Field, 2004), and that this induces sparse internal templates in classification image experiments.

A third, more pragmatic approach to finding good prior assumptions is to combine and extend proven prior assumptions in a synergistic manner. Smoothing and thresholding have proven useful in analyzing data from classification image experiments (Chauvin et al., 2005; Gold, Murray, Bennett, & Sekuler, 2000; Knoblauch & Kenneth, 2008; Mangini & Biederman, 2004; Rajashekar, Bovik, & Cormack, 2006; Tadin et al., 2006), and combining and extending the two might yield a more effective estimation technique. Sparseness in a smooth basis is a natural generalization of assumptions of smoothness and sparseness that has the potential to yield a strong class of prior assumptions.

In light of the ideas mentioned above, we propose imposing sparseness in a basis of smooth functions as a way of increasing the accuracy and efficiency with which classification images are estimated. We impose this assumption in a framework that can naturally accommodate prior information.

2.1.2 Imposing basis sparseness in generalized linear models

Our analytical framework builds upon generalized linear models (GLMs), which have been used previously to estimate classification images (Abbey & Eckstein, 2001; Knoblauch & Maloney, 2008a; Knoblauch & Maloney, 2008b; Solomon, 2002) and neuronal receptive fields (Wu, David, & Gallant, 2006). As with the linear observer model typically used in classification image experiments (Ahumada, 2002), GLMs assume that the output of a system is generated by first linearly correlating the input with a template and that the internal response is then transduced to an observed response by a fixed stochastic process (Figure 2-1B). GLMs have a number of desirable properties: they provide a unifying framework for estimating classification images, neuronal receptive fields, and functional imaging (Victor, 2005); they can be fit efficiently by maximum likelihood (ML) methods (Wu et al., 2006); they work with arbitrary stimuli; and extensions to the basic model can incorporate important input or output nonlinearities (Ahrens, Paninski, & Sahani, 2008). Most importantly here, constraints on classification images can naturally be imposed in the GLM context by the use of a prior, which assigns probabilities to different parameter values. For example, spatial smoothness is imposed by assuming that the spatial derivatives of the template follow a Gaussian distribution of a given width (Knoblauch & Maloney, 2008b).

The assumption of sparseness in a particular basis translates naturally into a prior distribution which gives higher probability to models with a small number of nonzero basis coefficients. The Laplace distribution, which is heavily peaked around zero, is the sparsest distribution for which the GLM estimation problem is tractable (Seeger, 2008). We thus propose to estimate classification images through GLMs by imposing such a sparseness-inducing prior on basis coefficients. This strategy provides a realistic and parsimonious account of the observer's strategy during a variety of psychophysical tasks.

We show by simulations and experiments with a real observer that classification images estimated with a sparse prior in a basis are less noisy and take less trials to converge than those estimated by other methods used in the literature. In addition, the sparse prior discards coefficients which do not contribute significantly to an observer's decision process, leading to classification images that are highly interpretable and, under appropriate circumstances, readily comparable to data obtained from single-unit recordings. All data sets and Matlab software used in this article are freely available at our Web site (http://apps.mni.mcgill.ca/research/cpack/sparseglm.zip) under the General Public License (GPL). A preliminary version of this work was presented previously (Mineault & Pack, 2008).

2.2 Methods: Statistical estimation of internal templates

2.2.1 The linear observer model

Consider a task in which an observer must report the presence or absence of a target. On each trial, the signal may or may not be present, and in all cases, the target or absence thereof is masked by a noise field. As in previous work (Abbey & Eckstein, 2002; Ahumada, 2002; Knoblauch & Maloney, 2008b; Murray, Bennett, & Sekuler, 2002), we assume that the observer performs the detection task by correlating the stimulus with an internal template. Trials in which the stimulus is similar to the template lead the observer to report the presence of the target. Thus, the stimulus is represented by a real vector **x** of dimension *k,* the template by a vector **w**, the internal noise by *ε,* and the observer computes an internal decision variable, *v,* by

$$v = \mathbf{x}^{\mathrm{T}}\mathbf{w} - c \quad (2\text{-}1)$$

Following each stimulus presentation, the observer gives a binary response *y* = ±1 according to whether the internal variable is larger than a criterion or offset *c*. The response *y* is thus given by

$$y = \mathrm{sign}(v + \epsilon) \quad (2\text{-}2)$$

We assume that the observer's internal noise ε is taken from a symmetric distribution with mean 0 and standard deviation σ, so that $\varepsilon \sim \Phi'(0, \sigma^2)$, where Φ is the cumulative distribution function (cdf) for the distribution in question. The internal noise represents observers' inconsistency: the same physical stimulus can elicit different responses. Other formulations (Ahumada, 2002) assume that it is the threshold *c* that is a random variable; our formulation is equivalent.

2.2.2 Finding estimates for the model parameters

We use statistical inference to find the most probable internal template **w**, given the data **y**. From Bayes' theorem, the posterior probability distribution of the parameters **w** is obtained from

$$p(\boldsymbol{w}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w}) \quad (2\text{-}3)$$

This equation captures the relationship between the *posterior* $p(\mathbf{w}|\,\mathbf{y})$, the *likelihood* $p(\mathbf{y}|\mathbf{w})$, and the *prior* $p(\mathbf{w})$. Our goal is to find the value of **w** that maximizes the posterior. For numerical reasons, it turns out to be simpler to find the value of **w** that minimizes the negative log of the posterior:

$$\boldsymbol{w}_{MAP} = \arg\min_{\mathrm{w}}\{-\log p(\boldsymbol{y}|\boldsymbol{w}) - \log p(\boldsymbol{w})\}$$

$$= \arg\min_{\mathbf{w}} \{L(\boldsymbol{y}, \boldsymbol{w}) + R(\boldsymbol{w})\} \quad (2\text{-}4)$$

When the prior is flat, we obtain the *maximum likelihood* (ML) estimate of **w**; otherwise, the estimate is called *maximum a posteriori* (MAP). The negative log-likelihood is denoted by *L,* while the negative log-prior, sometimes called the *regularizer,* is denoted by *R*.

2.2.3 Likelihood function for the linear observer Model

For a single stimulus **x**, the probability of observing response $y = +1$ given **w** and offset c is

$$p(y = +1|\, \boldsymbol{x}, \boldsymbol{w}, c) = \int_c^{\infty} \phi'(\mathbf{z} - \mathbf{x}^T\mathbf{w}, \sigma^2)$$
$$= \phi\left(\sigma^{-1}(\, \boldsymbol{x}^T\boldsymbol{w} - c)\right) \quad (2\text{-}5)$$

As probabilities sum to 1, $p(y = -1|\, \mathbf{x}, \mathbf{w}, c) = 1 - p(y = +1|\, \mathbf{x}, \mathbf{w}, c)$.

The experimenter's goal is to estimate the observer's internal template **w**, which is assumed to be constant throughout the duration of the experiment. We therefore wish to find a template that captures the relationship between a series of stimuli $\mathbf{X} = [\, \mathbf{x}^{\{1\}}, \mathbf{x}^{\{2\}}, \ldots, \mathbf{x}^{\{n\}}]$ and the corresponding responses $\boldsymbol{y} = (\, y_1, y_2, \ldots, y_n)$. We make the standard assumption that an observer's response on a given trial is independent of his or her responses on other trials. The likelihood function is then:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, c) = \prod_{i=1}^{N} p(y_i|\mathbf{x}, \mathbf{w}, c) = \prod_{i}^{N} \phi\left(\sigma^{-1} y_i \left(\mathbf{x}^{i\,T}\mathbf{w} - c\right)\right) \quad (2\text{-}6)$$

There is an ambiguity in this formulation, as for any given value of σ it is possible to multiply c and **w** by a constant factor such that the likelihood stays the same. We resolve this ambiguity by using the standard procedure of setting $\sigma = 1$; noisier observers are accommodated by a smaller overall magnitude for the weight vector. The negative log-likelihood is therefore

$$-\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, c) = -\sum_{i}^{N} \log \phi(y_i(\mathbf{X}\mathbf{w} - c)_i) \quad (2\text{-}7)$$

We make this formulation slightly more general by allowing for a set of auxiliary variables **U** with corresponding weights **u**. Later, we will impose a prior on **w** but not on **u**. This formulation allows for the inclusion of factors that affect the observer's responses other than the noise stimulus. For example, in Tadin et al. (2006), the question of interest is the time-varying influence of visual motion in the stimulus surround on judgments of motion direction in the center. Here, the time-varying surround stimulus would be put into the **X** matrix while the signal sign in the center would take one column of the **U** matrix. Another possibility, in detection tasks, is to include the signal sign into the **U** matrix, as in

Knoblauch and Maloney (2008b, p.5, Equation 16), rather than summing noise and signal in the **X** matrix. We found that the **U** strategy generally led to faster optimization while yielding equivalent estimates of internal templates. A final possibility is to include trial-delayed responses in the **U** matrix to model observer's tendency to respond similarly or dissimilarly on subsequent trials, as in the commonly used method of modelling refractory periods and burstiness in neurons (Pillow et al., 2008). We eliminate the constant *c* by adding a constant column to the matrix **U**. Under this new parameterization, we have the negative log-likelihood function for the linear observer model:

$$L = -\sum_i^n \log \phi(y_i(\mathbf{Xw} + \mathbf{Uu})_i) \quad (2\text{-}8)$$

The above expression is known in the statistics literature as the negative log-likelihood for a binomial *generalized linear model* (GLM) (Gelman, Carlin, Stern, & Rubin, 2003). Assuming that the internal noise has a Gaussian distribution, Φ becomes the cdf of a Gaussian, and we obtain the probit model (Knoblauch & Maloney, 2008b). If instead we assume that the internal noise has a logistic distribution, $\Phi(x) = 1 / (1 + \exp(-x))$, we obtain the logit or logistic regression model. In practice, the Gaussian and logistic distributions are very similar, and empirically it is difficult if not impossible to determine which provides a more accurate description of an observer's internal noise. For computational simplicity, we adopt the logistic regression model.

2.3 Methods: Regularization of the solution

In a classification image experiment, we typically have substantial prior expectations about the internal template **w**, as we have argued in the introduction. Such prior information can effectively narrow the space of possible classification images, thus improving the efficiency with which classification images can be recovered, as well as their accuracy and the interpretability of the results. Here we propose the use of sparse priors on smooth basis coefficients, which together impose global sparseness and local smoothness on the recovered templates.

We briefly discuss Gaussian priors for comparison. These assume that linear combinations of weights have Gaussian distributions. This can be used to impose the condition that weights are not too large, or that they vary smoothly across space. They have been discussed in detail in the context of neurophysiology in Wu et al. (2006) and are an integral component of the spline smoothing framework proposed by Knoblauch and Maloney (2008b).

2.3.1 Gaussian priors

Many priors arise from the assumption that linear transformations of weights have Gaussian distributions, $p(\mathbf{w}) = \mathcal{N}(\mathbf{w};0, (\lambda \mathbf{A}^{\mathsf{T}} \mathbf{A})^{-1})$, where λ is a scale factor. Common cases of Gaussian priors are shown in Table 2-1. When the matrix **A** is the identity matrix, we get a penalty on the magnitude of coefficients known as weight decay. When **A** is a spatial derivative operator, we get the smoothness prior, which encourages spatially smooth weights. In general, any choice of **A** for which $\mathbf{A}^T \mathbf{A}$ is positive definite generates a proper Gaussian prior. Choices other than weight decay and smoothness priors are discussed in some detail in Wu et al. (2006).

2.3.2 Sparse priors

We introduce a constraint of sparseness by way of a prior distribution over the weights **w**. If weights are sparse, then most are zero and few have large values. The corresponding prior distribution should have a peak at 0 and heavy tails. A simple prior embodying these assumptions is that of an independent sparse prior over the weights:

$$p(w_i|\lambda) \propto \exp(-\lambda|w_i|)$$

$$p(\mathbf{w}|\lambda) \propto \exp(-\lambda \textstyle\sum_i |w_i|) = \exp\left(-\lambda \left|\left|\mathbf{w}\right|\right|_1\right) \quad (2\text{-}9)$$

Here $\|\mathbf{w}\|_1$ denotes the L_1 norm of the vector **w**. The distribution $\exp(-\lambda\|\mathbf{w}\|_1)$ is known as the Laplace distribution, a term we avoid because of the potential confusion with the unrelated Laplacian pyramid which we introduce later; in the following we will refer to the prior induced by this distribution as the "sparse prior." This distribution has a sharp peak at the origin and heavy tails. In the context of Equation 2-3, the multiplication of this prior with the likelihood $p(y|w)$ has predictable effects on the resulting posterior density. For instance, if we assume a Gaussian likelihood (Figure 2-2), the resulting posterior density has a peak at the origin when the Gaussian's center is sufficiently close to 0. Hence, parameters whose presence gives only a marginal increase in the likelihood of the data will tend to be clamped at 0 and thus discarded from the model. L_1 regularization in the standard ("pixel") basis thus gives results similar to post hoc thresholding.

Of priors of the form $p(\mathbf{w}) = \exp(-\lambda \sum_i |w_i|^q)$, which includes both Gaussian and Laplace distribution priors, $p(\mathbf{w})$ is log-concave for $q \geq 1$, which leads to a convex, tractable optimization problems when estimating **w**. As sparseness increases for small q, $q = 1$ yields the sparsest distribution of this form

which leads to a tractable optimization problem (Seeger, 2008), which makes it generally preferable to alternative sparseness-inducing priors.

2.3.3 Reformulating the linear observer in terms of basis coefficients

A given model may require many weights to represent an internal template in the pixel basis yet be sparse in some other basis. We would thus like to reparameterize our problem to express our assumption that weights are sparse in an arbitrary basis **B**. Denoting the basis weights as $\tilde{\mathbf{w}}$, we re-express the negative log-likelihood (Equation 2-7) as

$$L_B = -\sum_{i=1}^{n}(\log \Phi(y_i(\mathbf{XB}\,\tilde{\mathbf{w}} + \mathbf{Uu})_i)) \quad (2\text{-}10)$$

The L_1 norm is not conserved under a rescaling of the columns of the basis matrix **B**, which means that a sparse prior will consider certain basis functions more likely if **B** is scaled unevenly. To avoid this, the matrix **B** should be normalized so that its component column vectors have norm 1. The associated MAP estimate is given by the vector $\tilde{\mathbf{w}}$ which minimizes:

$$L_B + \lambda\left|\left|\tilde{\mathbf{w}}\right|\right|_1 \quad (2\text{-}11)$$

The sparse prior is now imposed on the basis coefficients rather than on the classification image coefficients. The internal template can then be visualized by projection of the weights onto pixel space, $\mathbf{w} = \mathbf{B}\tilde{\mathbf{w}}$.

2.3.4 Choice of basis

We use the term *basis* loosely, as the rows of **B** need not span $\mathbf{R}^k$, as with the spline basis used in Knoblauch and Maloney (2008b), nor do they need to be linearly independent, as with overcomplete transforms. We can freely construct a basis matrix that embeds our assumptions for the particular classification image reconstruction problem at hand. In the case where **B** is overcomplete, that is, its columns span $\mathbf{R}^k$ but are not linearly independent, there are many equivalent ways of expressing the classification image. The sparse prior will tend to select the simplest way of expressing the classification image. The compatibility of sparse priors with overcomplete bases is advantageous as it is generally simpler to construct bases which have desirable properties when one removes the restriction of linear independence.

An assumption of smoothness can be embedded implicitly into the choice of smooth basis functions. In this regard, Gaussian basis functions are a natural choice, but for the analysis of real classification

images it is desirable to have a basis that allows for the degree of smoothness to vary over space. Because internal templates may vary among observers and among tasks an ideal basis would not require strong *a priori* assumptions about the degree of smoothness in the classification image. Both criteria can be met by the use of a Laplacian pyramid (Burt & Adelson, 1983), which consists of multiple Gaussian functions that are smooth on various spatial scales.

In constructing a Laplacian pyramid, one typically chooses Gaussian functions with widths that are powers of 2. Each set of Gaussian basis functions that share the same width is known as a level, and within a level, the spatial separation of the basis functions is proportional to the width of the Gaussians. The power of two scheme means that in *m* dimensions, the decomposition is overcomplete by a factor $2^m / (2^m - 1) \leq 2$, i.e., only mildly overcomplete. The Laplacian pyramid can be extended by using more basis functions than is standard, for example by adding half-levels, or having more basis functions than standard within a level. Such undecimated Laplacian pyramids lead to greater computational cost during model fitting but can be more powerful than standard Laplacian pyramids.

Other decompositions based on using oriented basis functions at different resolutions include the steerable pyramid transform (Simoncelli & Freeman, 1995), several families of overcomplete wavelets (Selesnick, Baraniuk, & Kingsbury, 2005), and Gabor pyramids. These correspond to an assumption of sparse, smooth, oriented structure. Finally, transformations such as the discrete cosine transform (DCT) and the Fourier transform (Ahumada & Lovell, 1971; Levi & Klein, 2002) may be used to analyze the contributions of different spatial frequencies to performance on the task.

2.4 Methods: Hyperparameter selection

Using a Gaussian or sparse prior involves the specification of a free hyperparameter λ, much like smoothing involves the choice of the width of the smoothing kernel. This hyperparameter may be selected by assessing how well a model generalizes to untrained examples using *k*-fold cross-validation (Wu et al., 2006). Generalization performance is naturally assessed using the cross-validated log-likelihood of the data.

The *k*-fold cross-validation increases the computational burden of estimating a model by roughly *km,* where *m* is the number of regularization hyperparameters considered. In typical use, *km* is on the order of 50–100. This limits the practical applicability of this form of cross-validation to models which take a minute or so to fit, such as models with Gaussian priors.

Many methods for fitting a model with a sparse prior work by successively finding minima of subproblems of the form:

$$\arg\min_{\mathbf{w}} f(\mathbf{w}) + \lambda_i \left|\left|\mathbf{w}\right|\right|_1 \quad (2\text{-}12)$$

Here λ_i is a decreasing sequence of regularization hyperparameters, and $f(\mathbf{w})$ is some convex function of $\mathbf{w}$. These algorithms therefore generate an entire regularization path for roughly the same cost as fitting the model with the smallest λ considered. The cost of k-fold cross-validation is then roughly $k + 1 \ll km$ times the cost of fitting a single model, which makes cross-validation with sparse priors feasible. We chose the fixed-point continuation algorithm (Hale, Yin, & Zhang, 2007) to estimate model parameters with sparse priors. The algorithm is outlined in Appendix A.

2.5 Methods: Simulations

We simulated a linear observer on several detection tasks. The response y on the ith trial was given by

$$v_i = (\mathbf{X}_i + \mathbf{S}_i)\mathbf{w} + c$$
$$y_i = \text{sign}(v_i + \epsilon_i) \quad (2\text{-}13)$$

The stimulus was composed of a signal $\mathbf{S}_i$ combined additively with a noise stimulus $\mathbf{X}_i$, which was chosen from the Gaussian distribution. The templates $\mathbf{w}$ varied depending on the task. In the one-dimensional task, the signal and templates were even Gabors. In the two-dimensional task, the signal was a Gaussian blob and the template was a Gaussian blob in one case and a difference of Gaussians in the other. The signal was normalized so that the observer correctly categorized the signal on 81% of the trials before the addition of internal noise. The internal noise ε_i was chosen from a Gaussian distribution $p(\varepsilon) \sim \mathcal{N}(0, \sigma^2)$, after which the performance dropped to 75%. Since Gaussians are mapped to Gaussians under linear transformations, $p(v) \sim \mathcal{N}(\mu, \sigma_r^2)$ for trials of the same signal sign and the following relation holds:

$$\sigma/\sigma_r = \sqrt{\left(\frac{\Phi^{-1}(0.81)}{\Phi^{-1}(0.75)}\right)^2 - 1} \approx 0.833 \quad (2\text{-}14)$$

This gave an observer self-consistency on repeated simulated presentations of the stimuli of 0.76, within the range of reported values of self-consistency in various classification image experiments (Murray et al., 2002; Neri & Levi, 2008). The offset or criterion term c was adjusted such that the observer was unbiased. We estimated the observer's internal template with a logistic regression GLM with a weight

decay, smoothness, and sparse prior. The optimal hyperparameter λ for each prior type was estimated through 5-fold cross-validation.

For the 1D task, the signal was sampled at a resolution of 64 pixels. The corresponding sparse prior basis for the task was a Laplacian pyramid spanning three levels, with half-levels included, overcomplete by a factor of ≈4. For the 2D task, the signal was sampled at a resolution of 17 × 17 pixels, and three bases were used: the Dirac (pixel) basis, a full Laplacian pyramid, overcomplete by a factor of ≈1.25, and a full steerable pyramid basis with two orientations, overcomplete by a factor of ≈3.25.

2.6 Methods: Real observer

We tested our procedure on a real observer (author PM), whose task was to detect the presence or absence of a Gaussian blob under conditions directly analogous to the those used in our simulations. We also used variations of this task in which the observer had to identify the null signal or four Gaussian blobs placed symmetrically around the center of the screen. The 2500 trials were performed originally for each task. Signals were masked by additive independent Gaussian noise. The noise variance was adjusted by a staircase procedure so that the observer performed at 75%. The signal and noise were sampled at a resolution of 16 × 16 pixels. The results were analyzed with a full Laplacian pyramid with half-levels, overcomplete by a factor of ≈2, in conjunction with the sparse prior. The 2500 trials were separated at random into a fit set containing 2000 trials and a validation set containing 500 trials.

2.6.1 Inference power estimation

For the four-blob task, an additional 2700 trials were collected. The goal was to estimate how many trials one must do to have sufficient power to reject a baseline hypothesis using models with different priors. This is equivalent to asking how many trials one must do for the cross-validated deviance of a given model to be smaller than that of a baseline model. We first pooled the additional trials with the original trials, for a total of 5200 trials. We took 100 samples each of lengths 500, 800, 1200, 2000, 3250, and 5100 trials, without replacement, from this pool. For each sample, we fit a variety of different models, discussed in the Real observer section. Cross-validated deviance was averaged across samples of the same length to obtain an estimate of the mean cross-validated deviance attained by each model for a certain number of trials.

2.7 Results

2.7.1 Simulated observer, one-dimensional Gabor

We first simulated an experiment in which the linear observer had to detect a one-dimensional Gabor stimulus (Figure 2-3, lower right) embedded in additive, Gaussian noise. Figure 2-3 shows estimated templates for increasing numbers of simulated trials based on the standard weighted sums formula (top row), a GLM with a smoothness prior (middle row), and a GLM with a sparse prior in a Laplacian pyramid basis (bottom row). The sparse prior template estimate is accurate even for very low numbers of trials. In terms of correlation between the real and estimated templates, the sparse prior estimate at 200 trials is comparable to the smoothness prior estimate at somewhere between 500 and 1000 trials.

The smoothness prior performs suboptimally here because the smoothness scale varies over the template. The sides of the template are flat (infinitely smooth), while the center of the template is smooth on a small spatial scale. Using a stronger smoothness prior to eliminate the noise on the side of the template would oversmooth the center of the template. The shortest characteristic length scale of the internal template acts as an upper bound on the level of smoothness that can be imposed on the template without significantly biasing parameter estimates. In contrast, the sparse prior in a Laplacian pyramid basis is highly effective here, as a Gabor can be represented sparsely in this basis.

More insight into the effect of a sparse prior can be gained by considering the weights as a function of the strength of the regularization. Figure 2-4A shows the estimated weights as a function of λ for 1000 simulated trials. For large λ, most weights are zero. As λ is decreased, more weights become active. Once a weight has become active, it tends to stay active for smaller λ.

Figure 2-4B shows the template estimates for different values of λ. For large λ, only the areas of the template which have the most influence on the observer's decision are recovered. As λ is decreased and more weights become active, the reconstruction becomes more complex and accurate. Beyond a certain point, however, the model starts fitting to noise, and the reconstruction becomes less accurate. From this point of view, the sparse prior works by determining how influential each weight is, and only allowing weights into the model which are more influential than a cutoff that is determined by the strength of the regularization λ.

The tuning of the hyperparameter is conceptually very similar to thresholding. When an overcomplete basis is used, however, the influence of a single weight, as measured by its magnitude, no longer has a

straightforward interpretation. In an overcomplete basis, there are several equivalent ways of expressing the same template, which means that the magnitude of a weight can change depending on the representation chosen. To resolve this ambiguity, the sparse prior approximately selects a best subset of model variables in a nongreedy fashion (Tibshirani, 1996). This can be viewed as an instantiation of Occam's razor: Given several models which predict the same outcome, whichever model is simplest is the preferred model. This gives a model with a sparse prior robustness against irrelevant covariates and naturally generalizes the standard practice of thresholding classification images.

2.7.2 Simulated observers, two-dimensional difference of Gaussians

We have shown that estimated templates are accurate when sparse priors are imposed on a suitable basis. Our proposed estimation method can accommodate any choice of basis, and it is this free choice of basis which underlies the power of the method. If, in the chosen basis, the observer's template cannot be represented sparsely, or if there is too much noise for coefficients to be well constrained, the sparse prior will tend to discard coefficients which have little influence on the outcome: the model is truncated. Model truncation can lead to artifacts whose nature depends on the basis considered; typically, artifacts look like the basis functions used.

A good basis for a given classification image meets two criteria: (1) plausible templates can be represented sparsely for the given problem; and (2) possible artifacts of model truncation will not bias the experimenter's interpretation of the estimated templates. For many problems, the first criterion will rule out the Dirac (pixel) basis, in which many templates are not sparse. The second criterion rules out bases whose functions are global, because artifacts of reconstruction will spread across space; this includes the Fourier basis. A good basis should be neither completely local nor completely global, which leaves a number of schemes based on using the same basis functions at different scales, such as pyramid transforms and wavelets.

We illustrate these ideas with simulated linear observers in a 2D Gaussian blob detection task. We imagine that the experimenter wants to know, in such a task, the general shape of the observer's internal template, and in particular, whether the observer's internal template has an inhibitory surround. We simulated linear observers with two different internal templates: one with an inhibitory surround (a difference of Gaussians) and one without (a single Gaussian). Figure 2-5 shows typical estimates of these templates after 2000 simulated trials for three choices of basis: Dirac basis, Laplacian pyramid, and steerable pyramid with two orientations.

The Dirac basis (Figure 2-5I) fails to meet Criterion 1 for this problem, as the template with an inhibitory surround is not sparse in pixel space. Because each pixel has a very small influence on the outcome of the model, very many have to be kept active in order to obtain a fair approximation of the template, and the sparse prior sets only 10% of the pixels to zero. The resulting templates (Figures 2-5B and 2-5F) are similar to what would be expected when using post hoc thresholding of a classification image with a low threshold.

The steerable pyramid uses Gaussian spatial derivatives at different scales as basis functions, similar to Gabors, which are appropriate for representing internal templates containing edges (Figure 2-5K). It fails to meet Criterion 2 for this problem, as the basis functions have both positive and negative regions. The latter can masquerade as inhibitory surrounds, and indeed the template shown in Figure 2-5H, contains a faint inhibitory zone around the excitatory region of the template. As the actual template has no surround, the blue region in the figure is an artifact of the choice of basis.

The Laplacian pyramid basis functions are Gaussian blobs of different sizes (Figure 2-5J). This is an appropriate basis here, as plausible templates for this task can be sparsely described in the basis, and the main artifact of reconstruction, excessive smoothing, is not critical in judging the existence of an inhibitory surround.

More generally, the right choice of basis depends on the specific problem at hand. Pyramid bases, composed of stereotyped smooth basis functions at different scales, are effective basis choices when the observer's internal template contains sparse smooth structure at an unknown and potentially spatially varying scale. The Laplacian pyramid, in particular, may be a safe choice for many classification image paradigms where smoothing is an effective denoising strategy. For problems in which the features of the classification image are oriented and localized, a wavelet-like basis such as the steerable pyramid may be useful. When frequency response is of importance, the discrete Fourier transform (DFT) basis can be a good choice, although one should not attempt to spatially reconstruct the template in such a basis because of ripple artifacts. Finally, bases that exploit the geometry of the task can be used to obtain the most power in answering specific questions. For example, to judge the existence or absence of an inhibitory surround, one could exploit the radial symmetry of the problem by using a basis composed of concentric rings blurred at different scales, which would give effects similar to radial averaging (Abbey & Eckstein, 2002).

2.7.3 Real observer

Real observers can display a number of behaviors not accounted for by a linear model: nonstationarity, lapses in attention, input nonlinearities, spatial uncertainty, and so forth. A good model should be robust to model misspecification. We therefore tested our approach with a real observer on a blob detection task similar to the one used in the simulations. In the first task, an observer was asked to indicate whether the target, a Gaussian blob, was present in the center of the screen. The target or absence thereof was masked by additive Gaussian noise. The second task was similar, but the target was now four Gaussian blobs placed around the center of the screen. Figure 2-6 shows the target stimulus (column 1), along with internal templates estimated under a GLM model with smoothness (columns 2–4) and sparse (columns 5–7) priors. The sparse prior was defined in a Laplacian pyramid basis as described in the Methods section.

For the one-blob task (top row), the sparse prior estimate shows a shape similar to that obtained with the smoothness prior, although noticeably less noisy. This increase in efficiency allows us to estimate with some fidelity two partial classification images, one for when the signal is present and one for when the signal is absent, by splitting the design matrix in two as described in Knoblauch and Maloney (2008b). The template corresponding to conditions in which the signal was present is highly spatially localized and quite similar to the signal, while the *signal-absent* template is more blurry. This is likely due to spatial uncertainty, which strongly affects recovered signal-absent templates but not signal-present templates in detection tasks (Tjan & Nandy, 2006). This effect is visible in a similar task in Figure 4 of (Abbey & Eckstein, 2002), and an analogous effect was shown in the time domain in Knoblauch and Maloney (2008b). Both the signal-present and signal-absent templates show hints of a large, weak inhibitory surround.

In the four-blob scenario (bottom row), the template estimated through the sparse prior again appears less noisy than that obtained with a smoothness prior. The partial templates show a pattern consistent with spatial uncertainty, with the estimated signal-present template looking very much like the signal, and the signal-absent template being just a blur. Again, signal-present and signal-absent templates show hints of a weak inhibitory surround. Estimated templates in both tasks are thus compatible with a spatially uncertain observer who judges the presence or absence of a blob by comparing the luminance around one or several reference points to the luminance of the surround.

It is noteworthy that the sparse prior estimates show the same kind of features which are visible, with hindsight, in the smoothness prior estimates; the sparse prior simply enhances the visibility of these

features and denoises the estimated classification images. The sparse prior achieves this denoising by setting parameters which contribute little to the model to 0, as can be seen by the large areas of pure green in the estimated classification images. This in turn permits better resolution of subtle effects, such as the difference in the size of the excitatory blob in signal-present and signal-absent conditions in the one-blob task (top row, columns 3–4 versus columns 6–7), indicative of spatial uncertainty. This enhancement in the interpretability of the results is a key advantage of the sparse prior over other choices of prior.

There are several ways of performing quantitative comparisons of different models within the GLM framework, and these are discussed at length in Appendix B. In general goodness-of-fit is measured by the *deviance,* defined as twice the negative log-likelihood. Since the deviance always decreases with increasing number of free parameters, deviance values must be corrected for number of free parameters. Here we use the cross-validated deviance (D^{CV}), the Akaike Information Criterion (AIC), and the validated deviance (D^{val}) as measures of goodness-of-fit. All three metrics measure the generalization performance of a model; the first by repeated partitions of the data, the second using an asymptotic result, and the last with a set-aside validation set. In all cases, lower values are better.

Tables 2-2 and 2-3 compare the various models according to their goodness-of-fit and estimated degrees of freedom. We have included two models that can be considered as baselines. One is the ideal observer model, in which the response is given by the signal multiplied by an undetermined constant, with a signal presence/absence effect, plus an offset, embedded in a GLM with a weight decay prior. A second baseline is a pseudo-ideal observer similar to the first, but with separate gains for the signal-present and signal-absent cases, corresponding to a signal-sign dependent efficiency. The others models considered are GLMs with a weight decay prior, a smoothness prior, and a sparse prior in the Laplacian pyramid basis, either using a single template or partial templates for the signal-present and signal-absent conditions. In all cases, 5-fold cross-validation was used to determine optimal hyperparameters.

In both the one-blob (Table 2-2) and the four-blob (Table 2-3) cases, there is a clear pattern in which the model with a weight decay prior exhibits the worst performance, while the smoothness prior performs better, and the sparse is prior better still. Furthermore, the decrease in AIC and CV deviance when considering separate templates for signal-present and signal-absent conditions is most dramatic when using a sparse prior. The reason is simple: while the observer is behaving sufficiently nonlinearly for separate templates to be warranted for the signal-present and signal-absent cases, models with weight decay and smoothness priors must expend a large number of degrees of freedom to deal with the two

conditions separately, as shown in the *df* column. The decreased deviance is thus overshadowed by a large increase in degrees of freedom.

Classification image experiments are often designed to detect deviations from ideal observer behavior; these deviations are then considered as signs of hard-wired strategies or mechanisms. In many classification experiment paradigms, a failure to detect a reliable departure from ideal observer behavior would warrant running more trials or changing the experimental paradigm. Table 2-2 shows that in the one-blob detection experiment, an ideal observer or pseudo-ideal observer null hypothesis cannot be safely rejected on the basis of the weight decay prior. In the smoothness prior case, the situation is complicated as D^{val} favors the smoothness prior over the null hypotheses, but not D^{CV} nor the AIC, which suggests that the models estimated with the smoothness prior are performing sufficiently close to the ideal and pseudo-ideal observers to warrant doing more trials. Table 2-3 shows a similar effect in the four-blob task, as neither the weight decay nor the smoothness prior is close to rejecting the null hypotheses. In contrast, the model that makes use of a sparse prior supports the idea that the observer is behaving nonideally and nonlinearly, decisively in the one-blob task and to a lesser extent in the four-blob task. Again, looking at the *df* column, it is clear that the sparse prior achieves this by aggressively searching for low-dimensional models.

These results indicate that the sparse prior yields improvements in goodness-of-fit and inference power in comparison to smoothness or weight decay priors. We next sought to determine how these improvements depend on the total number of trials by collecting additional data in the four-blob task (total of 5200 trials). For each of the eight models considered above, we computed the cross-validated deviance for different numbers of trials, using a resampling procedure described in the Methods section. Based on this, we derived an estimate of the number of trials required for models with different priors to perform significantly better than a baseline.

Figure 2-7 plots the cross-validated deviance per trial for all eight models, for different numbers of trials. Under a model with no free parameters, the CV deviance per trial should be a straight line as a function of the number of trials. An ideal observer model (left, blue line) has few free parameters, so it asymptotes fast at a high value of CV deviance per trial (roughly 1.043). In contrast, linear observer models with various priors are more flexible, but they require more data to be well constrained. With a weight decay prior, it takes more than 5000 trials to disprove the ideal observer null hypothesis; with a smoothness prior, about 3400 trials; and with a sparse prior, about 1100 trials. In other words, we can

obtain the same inference power under the sparse prior as under a smoothness prior with less than a third of the trials.

With a signal effect, which requires the estimation of twice as many parameters, the conclusion is similar; the model with a sparse prior reaches significance against the pseudo-ideal observer null hypothesis with roughly 30% of the trials required to reach the same conclusion under a smoothness prior (1400 versus 4700 trials). The GLM with a sparse prior thus uses the observer's data more efficiently, leading both to an appreciable increase in inference power for a fixed number of trials, and to an appreciable decrease in the number of trials required to reach a certain inference power.

Sparse priors are thus helpful in real experiments in two main ways. First, they yield clear, noise-reduced internal template estimates which facilitate interpretation, in the same way that thresholding does. Second, by aggressively searching for low-complexity interpretations of the data, sparse priors recover better models, which allows one to conclusively show certain properties of observer's strategies without the need to gather data from an impractically large number of trials. The improvements in model fit brought by the use of a sparse prior are appreciable in a real sense due to the high noise inherent in classification image protocols and the high dimensionality of the models involved. These results show that the proposed method can supplement traditional methods even in simple tasks.

2.7.4 Discussion

In this work we have argued that a defining quality of classification images is simplicity, in that they may be expressed as a sparse sum of smooth basis functions. We have derived a method for imposing this condition through a sparse prior on smooth basis coefficients in a GLM framework. We showed in simulations that classification images estimated with sparse priors are often more accurate for a given number of trials than those estimated through other methods.

A key advantage of sparse priors over Gaussian priors is enhanced interpretability, as coefficients which do not contribute significantly to an observer's decision process are discarded from the model. We showed that a sparse prior allowed efficient estimation of the internal template of a real observer on a blob detection task. This increase in efficiency allowed for the accurate estimation of partial classification images for the signal-present and signal-absent conditions, revealing important differences thought to be due to spatial uncertainty. This increase in efficiency also lead to an appreciable decrease in the number of trials necessary to reach a certain inference power.

Sparse priors are especially useful in high-dimensional tasks, such as tasks involving both time and space. The number of classification image pixels to estimate in such models can be very high, and estimated classification images can be too noisy to be useful. Even in cases where the extrinsic dimensionality of the classification image is very high, the *intrinsic dimensionality* of the observer's template, that is, the number of basis coefficients needed to accurately represent it, may be quite low. Sparse priors in bases use this fact to effectively estimate the parameters of seemingly complex models.

2.7.5 Prior assumptions

A common argument against using priors is that they might bias the estimated internal templates and therefore lead to erroneous interpretations of observers' strategies in performing a task. A more general statement would be that priors introduce a bias-variance tradeoff: By restricting the space of possible models to some subset of all models, bias is introduced, but variance (noise) in the estimated models is reduced. In this sense, classic psychophysics, in which a handful of parameters are allowed to change, can be seen as having high bias and low variance, while unprocessed classification images have low bias but high variance. Our approach can be seen as a balance between these two extremes, with modest bias and variance.

Classic denoising methods such as smoothing and thresholding also implement a bias-variance tradeoff. The main difference between using a prior and classic denoising methods is that assumptions are made explicit in the prior approach. Hence, the question is not so much whether it is appropriate to use assumptions, unless one wants to reject classic denoising methods as well, but whether the assumptions *that one uses* are appropriate. If the assumptions are indeed warranted, then a reduction in noise during model estimation can lead to more powerful inference. The Real observer section gives an example of this, where the model fitted with a sparse prior allows us to forcefully infer that the observer is using a strategy that is both nonideal and nonlinear, in contrast to the conclusion obtained via models fitted with a weight decay and a smoothness prior.

The validity of one's choice of prior, and more generally one's choice of model, can be measured by the AIC and the cross-validated deviance, provided that the choice of prior or model was made before the examination of data. When a new model form is suggested by fitting a baseline model to data, complexity-corrected deviance measures will be overly optimistic about the new model; this process is known, derisively, as data dredging or “double dipping” (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). There are straightforward ways to address this issue. The first is to collect a leave-aside validation set, which is not used to suggest model form, and base judgments of model validity exclusively on the

probability of the validation data under the different models. By Bayes' theorem, this approach provides an estimate of the probabilities of the models themselves. This can be regarded as a probabilistic version of the validation model suggested by (Neri & Levi, 2006). While this most certainly works, as shown in the Real observer section, binary responses have high variance, and much validation data must be collected for the results to be reliable. A more efficient use of the data is to pool the validation and fit sets into one large data set.

A more practical solution is to determine model form using preliminary data. It is customary, in classification image experiments, to collect a limited quantity of pilot data based on one or two conditions with a single subject to judge whether the complete experiment with multiple conditions and subjects is likely to yield meaningful results. Model form, including the prior up to a limited number of hyperparameters, is settled at this point, and the pilot data are discarded from subsequent analysis. AIC and cross-validated deviance scores in the full experiment based on the settled models are then reliable measures of their validity.

Assuming one decides to use a sparse prior in a particular basis, the choice of basis is determined at the preliminary stage. The choice of good bases is heavily constrained, as explained in the results section, so that the experimenter's choice reduces to a handful of bases related to a pyramid decomposition, save for a few special cases. The experimenter chooses whichever of these few options works best on the preliminary data and continues to use this basis to analyze the results of the full experiment. Note that these remarks are also applicable to standard methods of classification image estimation; the choice of smoothing kernel and threshold should not be determined with the same data that is used to judge the result of the experiment unless corrections for multiple comparisons are used (Chauvin et al., 2005; Worsley, Marrett, Neelin, & Evans, 1996). Hence, prior choice need not be subjective.

2.7.6 Relationship to previous work

2.7.6.1 Sparse GLM models

In this article, we have suggested using a sparse prior on smooth basis coefficients of GLM model parameters as a method of accurately estimating classification images. Similar techniques have been used to induce sparseness of GLM parameters in the context of neurophysiology.

Sahani and Linden (2003) suggest using Gaussian priors in a linear model with one hyperparameter per parameter, optimizing over the marginal likelihood of the model, a technique called automatic relevancy determination (ARD). In practice, ARD sets many coefficients in the chosen basis to zero, leading to

sparse solutions, as with the sparse prior. The authors also suggest an alternative method which combines automatic smoothness determination and ARD. This gives results similar to imposing a sparse prior in a basis composed of Gaussian blobs which all have the same size, this size being automatically determined. In principle, ARD could be used with an overcomplete basis such as the Laplacian pyramid. However, ARD involves a nonconvex optimization problem, and so there can be multiple local minima in the induced error function, and the algorithm may not converge. This is not a major issue when the chosen basis is orthonormal, but with an overcomplete basis, multiple local minima are likely to form, as there are many ways of expressing the same vector. In addition, it can be unwieldy to extend ARD to more complex GLMs than the linear regression model (Wipf & Nagarajan, 2008; Bishop, 2006), in particular the logistic and probit regression models relevant to classification image experiments.

David, Mesgarani, and Shamma (2007) suggest using boosting as a method of estimating sparse spectro-temporal receptive fields in the context of auditory coding. In this method, the model is fit in steps, starting with an empty model. At each iteration, the residual between the fitted model and the data is computed. The weight whose associated regression function is most correlated with the residual is then modified. Initial iterations fit to prominent effects while later iterations tend to fit to noise. Regularization is done by early stopping, which halts the procedure at an optimal number of iterations estimated through cross-validation. The stepwise fitting procedure means that several weights never enter into the model, resulting in a sparse model. Indeed, boosting can be shown to implicitly and approximately find the MAP estimate for a GLM with a sparse prior (Rosset, Zhu, Hastie, & Schapire, 2004).

As explained in Appendix A, the cost per boosting iteration is the same as the cost per fixed-point continuation iteration, and the number of iterations required by boosting with overcomplete bases can be substantially larger than with our proposed method. Our proposed method thus offers a way of estimating a sparse GLM model that is, under appropriate circumstances, computationally more efficient than boosting.

Seeger, Gerwinn, and Bethge (2007) have recently proposed using GLM models with sparse priors to estimate receptive fields. In contrast to our approach, they attempt to estimate the entire posterior rather than just its mode. This is done by finding a Gaussian density which minimizes the Kullback–Leibler divergence between the real and approximated posterior, a technique called expectation propagation (EP).

Although EP and MAP methods use the same underlying model and prior, they have rather different properties. Since the EP technique obtains an estimate of the full posterior distribution rather than just a point estimate, it can be used for selecting the stimulus that is most likely to help constrain model parameters at a given point in an experiment, a technique called active design. While the MAP method retrieves a point estimate in which several coefficients are exactly zero, the EP method retrieves the mean of the approximated posterior for which all coefficients are nonzero. This is motivated by the fact that it is impossible, with any finite amount of data, to conclude that a coefficient is exactly zero, and that hard subset selection is incompatible with active design (Seeger et al., 2007). However, we suggest that the subset selection effect of the MAP method is highly desirable for interpretation purposes, automatically removing from the model coefficients that are likely to be fitting to noise, and thus letting the experimenter focus on the more prominent effects visible in a classification image. Finally, the EP optimization problem is more involved, numerically unstable, and computationally more expensive than the MAP estimation problem. Hence, although EP can be used to estimate receptive fields and classification images, in general EP and MAP methods are complementary techniques with different target applications.

A key point is that none of the proposed sparsity-inducing methods (including ours) have so far shown a decisive edge in the *quality* of estimated models compared to other sparsity-inducing methods. In contrast, sparsity-inducing methods do have a decisive edge over using Gaussian priors on some problems (David et al., 2007; Seeger et al., 2007; Rosset et al., 2004). Thus, it is advisable to use a sparsity-inducing method in general, while choosing the particular method which is most mathematically convenient and computationally efficient for a given application. Our proposed estimation method is competitive in terms of implementation speed, as explained in Appendix A. For example, the model with the largest design matrix considered in this article (10,000 × 256) took roughly 40 seconds to fit on a recent desktop computer, including 5-fold cross-validation.

2.7.7 Basis projections in classification images

Basis projections have been used implicitly as an intermediate in estimating classification images. For example, radial averaging can be viewed as a projection of a classification image onto a basis of concentric rings (Abbey & Eckstein, 2002). Projection onto a cubic spline basis of lower dimensionality than the stimulus is possible in the generalized additive model (GAM) framework of (Knoblauch & Maloney, 2008b). A radial discrete Fourier transform (DFT) basis has been used in Abbey, Eckstein, Shimozaki et al. (2002). In all these cases, the basis is undercomplete; that is, it does not span the full

vector of possible templates. The purpose of the projection was thus *hard* dimensionality reduction. In contrast, a sparse prior in a basis decides which basis functions to keep, yielding *soft* dimensionality reduction.

Others have used one basis in both presentation and analysis, for example a Fourier basis (Ahumada & Lovell, 1971; Levi & Klein, 2002). A particularly interesting example of such a technique was used in Mangini and Biederman (2004) in the context of face classification. Noise fields were generated by taking weighted sums of truncated sinusoids of varying orientation and spatial frequency. The authors then estimated classification images by applying the weighted sums formula in this truncated sinusoid basis. They applied a threshold corresponding to a given *p*-value in the truncated sinusoid basis and visualized the results by projecting the coefficients back onto the pixel basis. Only a fraction of the coefficients (less than 5%) were kept in the process. The resulting estimated templates were similar to the nonthresholded templates, but with less high-frequency noise, and improved interpretability. The quality of the truncated reconstruction is likely due to the fact that faces can be represented sparsely in the truncated sinusoid basis, which, like wavelet and pyramid bases, is passband and local.

Importantly, the truncated sinusoid basis was used in both the construction of the noise fields and the estimation of classification images. In contrast, our method does not require presenting a special type of noise to the observer; white noise, colored noise, or natural textures may be used. The basis used in the analysis may be chosen after the classification image experiment has been performed. Finally, we do not advocate using a truncated sinusoid basis as ringing artifacts are apparent in reconstruction. The steerable pyramid basis would be appropriate for the task used in Mangini and Biederman (2004).

2.8 Directions for future work and conclusion

2.8.1 Directions for future work

One might wonder whether a more powerful prior could be used to obtain even more accurate estimates of classification images in a lower number of trials. While this is a possibility, a potentially more fruitful approach is to consider better experimental designs. For example, it has been shown in the context of neurophysiological reverse correlation that different noise fields cannot be expected to yield the same amount of information about a visual neuron's receptive field (Paninski, 2005).

In the context of classification images, rather than randomly generating noise fields on each trial, the field which maximizes the expected information can be shown. Without necessarily using a fully

adaptive design, optimal design theory could potentially indicate which of several classes of stimuli, such as white noise, colored noise, or natural textures should be used for a given task. Optimal design strategies depend on the prior used; an optimal design method for a linear model with a sparse prior is presented in Seeger, 2008.

The main challenge is to extend the classification image approach to capture nonlinear behavior. Two methods have been commonly used so far for that purpose: using partial internal templates for signal-present and signal-absent conditions and computing second-order classification images (Knoblauch & Maloney, 2008b). In the reverse correlation literature, a successful approach has been to consider *expanded input spaces*: the stimuli, instead of being described simply as intensity over time or over time and space, are redescribed in terms of redundant features. For example, an auditory stimulus $x(t)$ can be redescribed using a windowed Fourier transform as a spectrogram $x(f, t)$, which augments the representation with the evolution of the different component frequencies f over time. A neuron's response can be modelled as a linear weighting of $x(s, t)$, and the resulting two-dimensional weight pattern is then known as a spectro-temporal receptive field (STRF). The STRF is able to capture a range of nonlinear behaviors, such as time-varying frequency sensitivity.

Extensions of this technique are described in Ahrens, Paninski, and Sahani (2008) and applied to auditory neurons in Ahrens, Linden, and Sahani (2008). Our framework is directly applicable to these and related problems, by a simple preprocessing of the input. One could also leverage existing physiological models of low-level cortical neurons to describe linear observers that operate on the output of simulated neurons. An observer in a detection task could be modelled as taking a weighted sum of V1 complex cells, and again our framework can be used here without substantial modification. However, a large number of nonlinear behaviors, such as spatial uncertainty, cannot be described by input nonlinearities (Tjan & Nandy, 2006), so future research will have to establish whether this approach is indeed fruitful.

A recent article by Ross and Cohen (2009) suggests an alternative path towards modelling nonlinear classification image observers, based on the idea that observers individually match features of the classification image and nonlinearly combine them to form a decision. It is assumed that the observer performs a classification or detection task by matching the input with several linear templates. Matches are nonlinearly transformed by a logistic function, and these transformed matches are linearly combined and fed through a final logistic function, which drives the response of the observer. This model can be viewed as a robust Bayesian reinterpretation of an artificial neural network (ANN) with a single hidden

layer and is reminiscent of linear-nonlinear cascade models commonly used in neurophysiology (Rust, Mante, Simoncelli, & Movshon, 2006). One of the main challenges of the approach is the proliferation of parameters, which is roughly the product of the number of hidden units and the number of pixels in the stimulus. A Markov random field prior is used to impose a type of piecewise smoothness which encourages template values to spatially cluster around values of ±1. We believe that a sparseness-inducing prior on template parameters, together with an appropriate basis, as advocated in our approach, could prove quite potent in the context of such data-hungry multi-feature classification images.

2.8.2 Conclusion

Classification images are of great interest to visual psychophysicists, both because they can often be compared in a straightforward manner to neuronal receptive fields, and because they provide a powerful means of measuring observers' strategies in visual tasks. However, their use has been limited by the quality and quantity of the data that can be collected. In this work, we have described an analytical framework that allows the experimenter to formalize important assumptions that are necessary to the interpretation of classification images. Our method relies on the reasonable assumption that the internal templates used by psychophysical observers are sparse and locally smooth. These constraints can be represented by appropriate choices of a basis set to represent the space of possible images and a prior that constrains the number of parameters that contribute to the recovered image. We have shown through simulation and through experiments with a real observer that estimating classification images with our method is more efficient and accurate than previous methods.

2.9 Figures

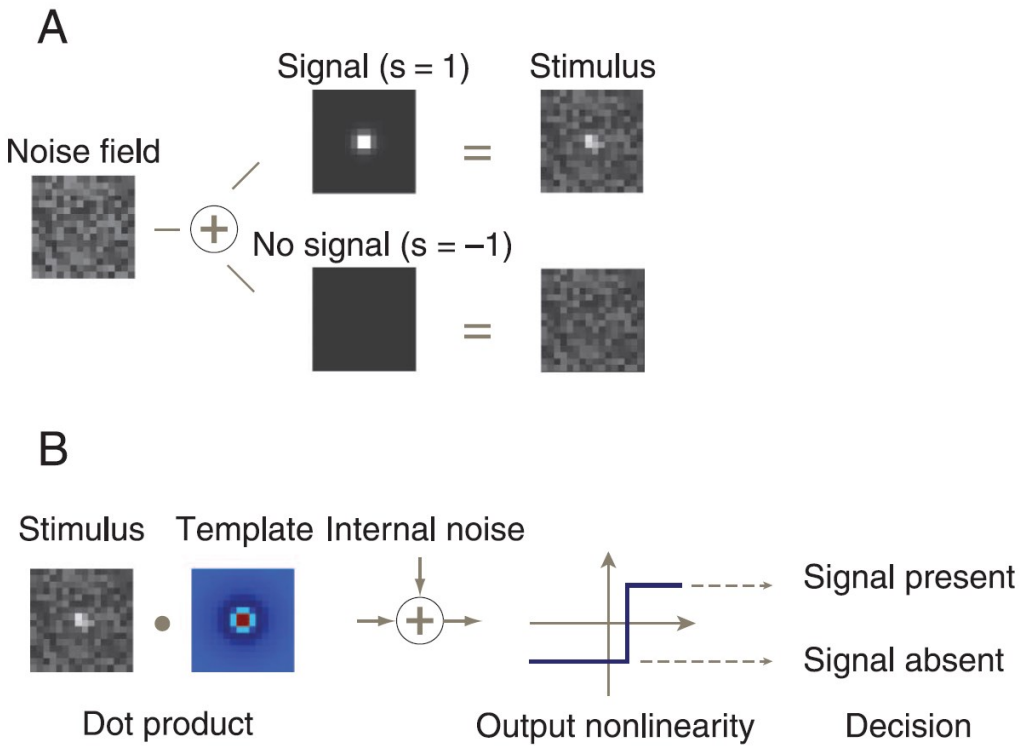


Figure 2-1: Outline of the classification image paradigm and the Linear Observer Model.

A: The classification image paradigm. On every trial, the observer is presented with a stimulus which is either the sum of a randomly generated noise field and the signal or just a noise field. The observer's task is to indicate whether the stimulus was present or absent on a trial. B: The Linear Observer model. The observer performs the task in A by correlating the stimulus with an internal template, giving a number which is corrupted by additive internal noise. The observer responds "signal present" when the number exceeds a criterion, and "signal absent" otherwise.

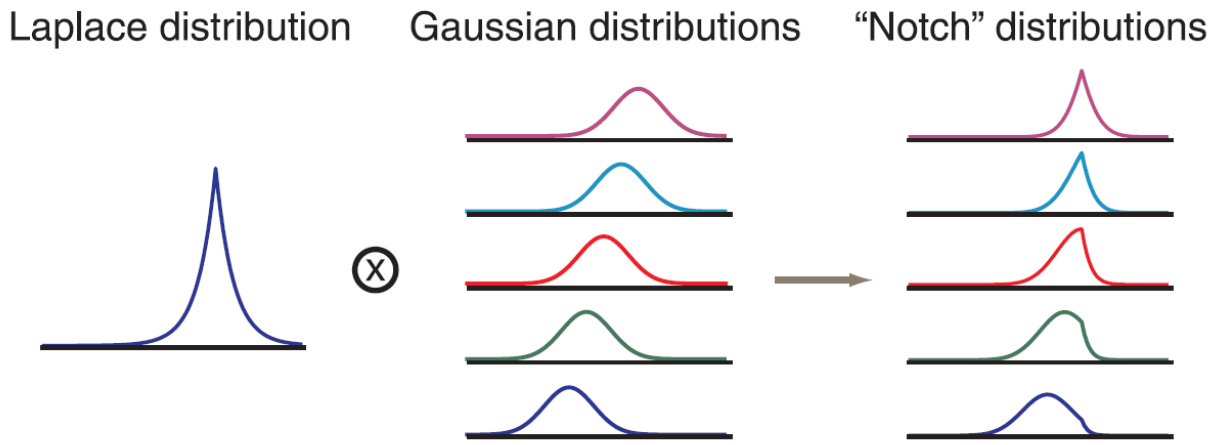


Figure 2-2: Effect of sparse prior on weights.

When a Laplace distribution centered at 0 is multiplied by Gaussian likelihoods with varying centers, the corresponding posteriors have discontinuities in their first derivatives at 0 (“notches”). For a wide range of Gaussian centers, the MAP estimate is exactly 0, as seen in the top three posterior distributions.

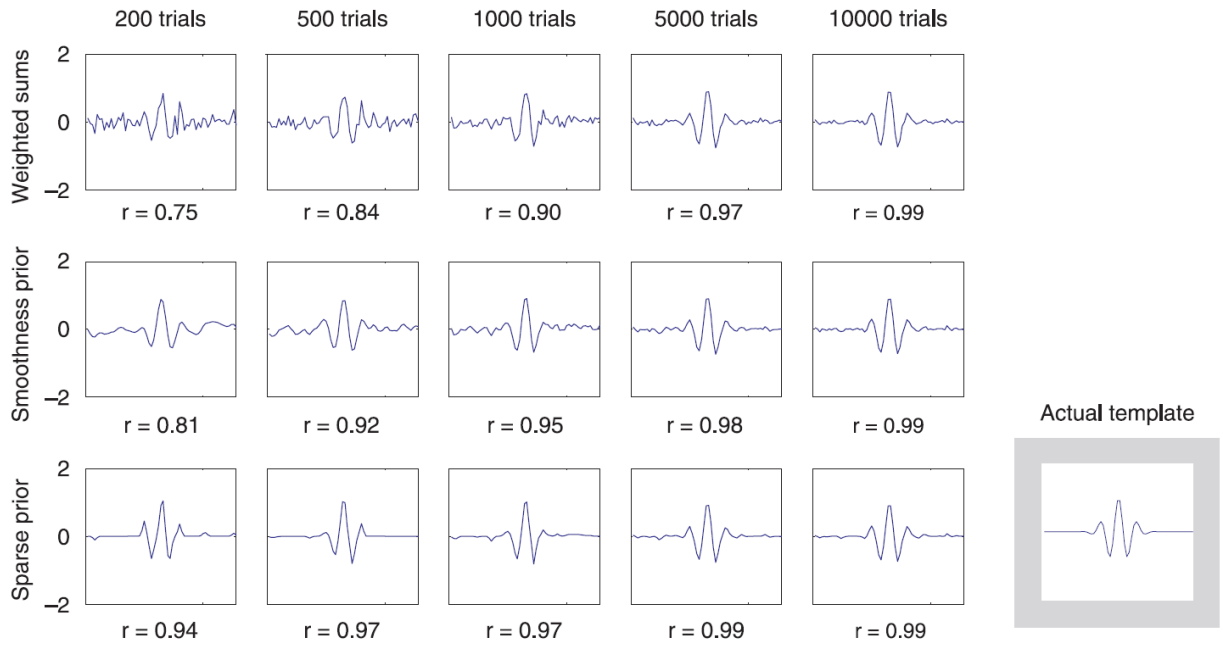


Figure 2-3: Estimated templates for a simulated linear observer.

The correlation between the estimate and the actual template is given below each estimated template. Inset: the actual simulated internal template. With a sparse prior, 200 trials suffice to obtain an accurate estimate of the actual template.

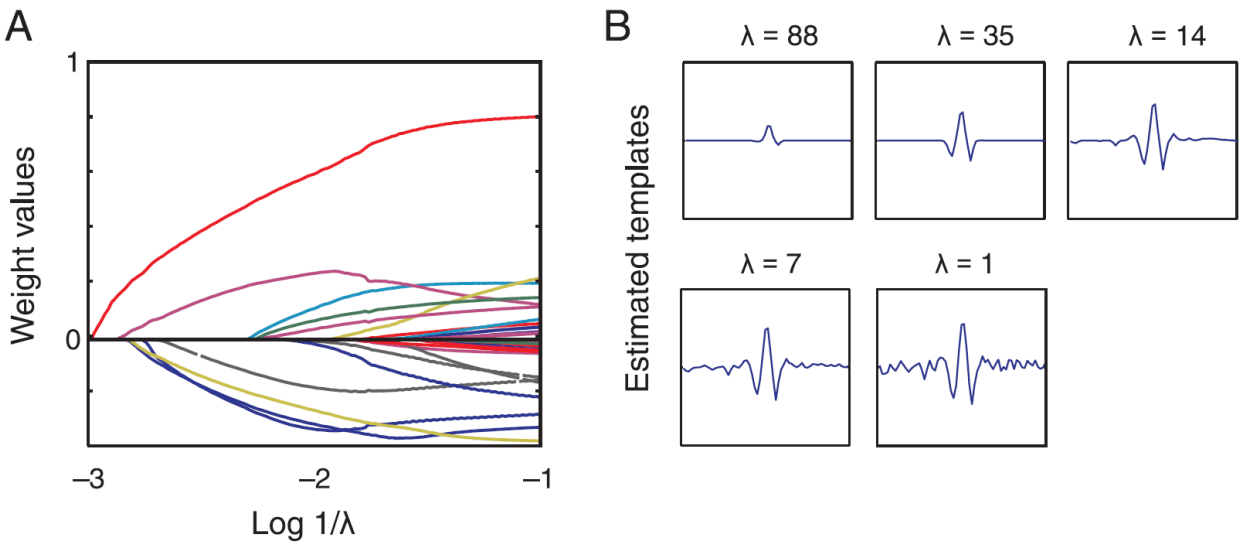


Figure 2-4: Weight paths with sparse priors.

(A) Estimated weights as a function of inverse regularization strength. Each line represents one of the weights of the fitted GLM model as a function of log 1/ λ. As λ decreases, more weights are added to the model. Once a weight is active, it tends to stay active. (B) Estimated templates for different values of λ. At large values of λ, only the most important areas of the template are recovered. The reconstruction becomes more complex and accurate for smaller λ. Beyond a certain λ, the model starts fitting to noise.

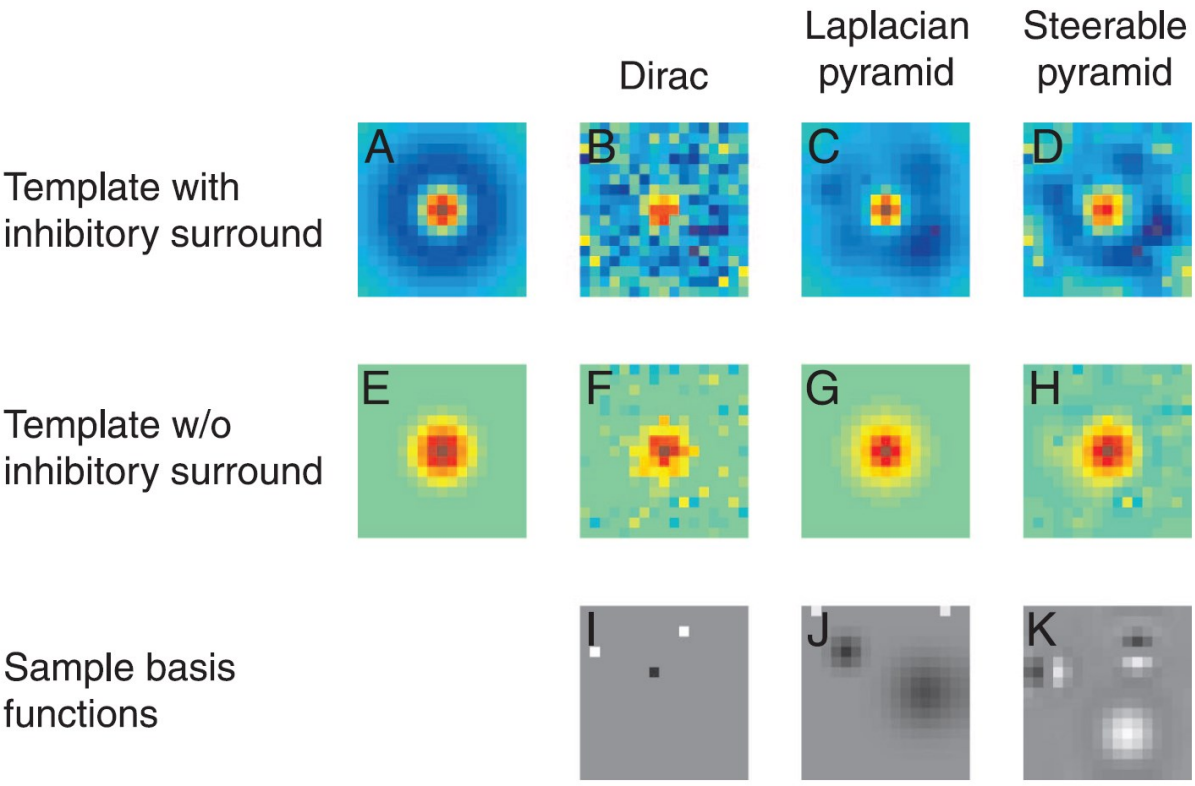


Figure 2-5: Estimated internal templates for simulated linear observer.

The Dirac basis cannot sparsely represent the template with inhibitory surround. The Steerable pyramid has basis functions which include positive and negative areas, and hence it is unsuitable for judging the existence of an inhibitory surround. The Laplacian pyramid is well suited for this task.

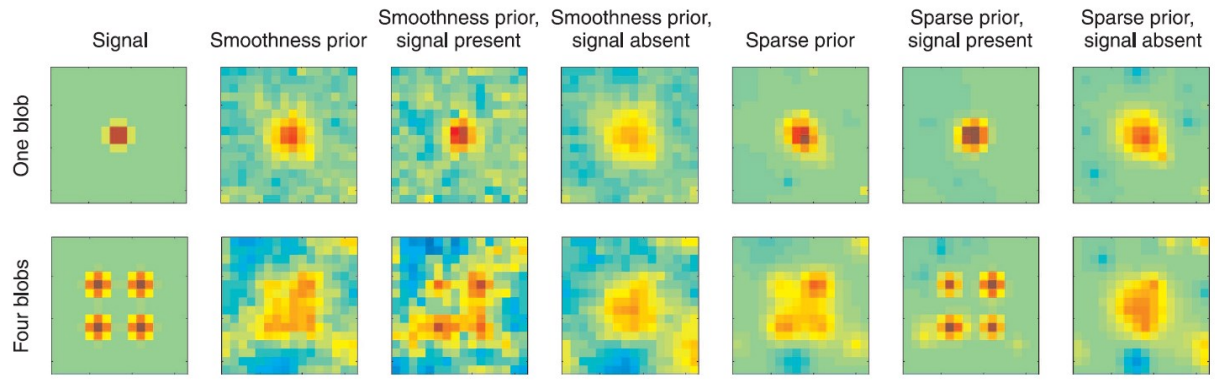


Figure 2-6: Estimated internal templates of real observer on one-blob and four-blob detection tasks.

Models estimated with sparse priors are less noisy than those estimated with smoothness priors. This increase in efficiency allows the accurate estimation of partial classification images for the signal present/signal absent conditions.

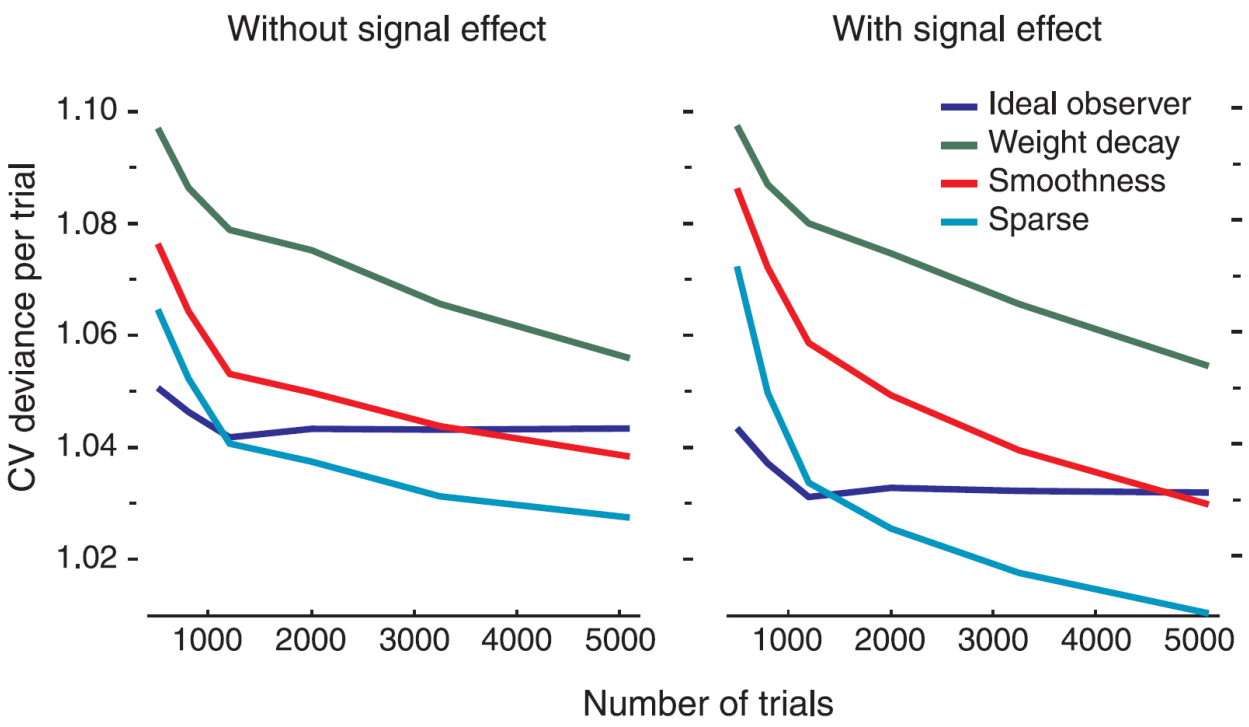


Figure 2-7: CV deviance per trial in the four-blob task estimated with different number of trials.

Left: without a signal effect. Right: with a signal effect. The model estimated with a sparse prior reaches significance against a null model appreciably faster, requiring less trials for the same quality of fit.

2.10 Tables

Gaussian prior name	**Regularizer *R* = −log *p*(w)**
Weight decay (ridge)	$\frac{\lambda}{2}\mathbf{w}^T\mathbf{I}\mathbf{w}$
Smoothness	$\frac{\lambda}{2}\mathbf{w}^T(\mathbf{D}^T\mathbf{D})\,\mathbf{w}$
Spline smoothness (Knoblauch & Maloney, 2008b)	$\frac{\lambda}{2}\mathbf{w}^T\mathbf{S}\mathbf{w}$
Arbitrary	$\frac{\lambda}{2}\mathbf{w}^T(\mathbf{A}^T\mathbf{A})\mathbf{w}$

Table 2-1: Gaussian priors in the context of classification images and corresponding regularizers.

Model type	***D***	***df***	**AIC**	D^{CV}	D^{val}
Ideal observer	2016	3	2022	2023	488.7
Weight decay prior	1748	166	2080	2109	503.2
Smoothness prior	1818	102	2022	2025	480.9
Sparse prior	1888	38	1964	1965	474.9
Pseudo-ideal observer	1985	4	1993	1994	474.0
Weight decay with signal effect	1722	203	2129	2134	492.3
Smoothness with signal effect	1680	173	2026	2028	461.9
Sparse prior with signal effect	1863	32	1927	1917	456.8

Table 2-2: Summary of fit results for several models in the 1-blob task.

D, deviance of the estimated model; *df,* degrees of freedom; AIC, Akaike Information Criterion; D^{CV}, cross-validated deviance; D^{val}, deviance of predictions on validation set.

Model type	***D***	***df***	**AIC**	D^{CV}	D^{val}
Ideal observer	2157	3	2163	2168	545.7
Weight decay prior	1989	112	2213	2218	571.4
Smoothness prior	2051	61	2174	2179	556.1
Sparse prior	2110	36	2182	2161	545.3
Pseudo-ideal observer	2127	4	2134	2140	542.9
Weight decay with signal effect	1980	114	2207	2202	569.2
Smoothness with signal effect	1956	106	2169	2161	560.6
Sparse prior with signal effect	2041	45	2131	2130	541.9

Table 2-3: Summary of fit results for several models in the four-blob task.

Legend as in Table 2-2.

Building on the parametric modelling methodology developed in Chapter 2, I extend and apply this framework to the analysis of visual representations at the level of neuronal ensembles in area V4 in Chapter 3. The results show that it is possible to infer, from multi-unit activity and local field potential (LFP) signals, the representation of visual space at a fine-grained scale over several millimeters of cortex. Analysis of the estimated visual representations reveals that LFPs reflect both local sources of input and global biases in visual representation. These results resolve a persistent puzzle in the literature regarding the spatial reach of the local field potential. This chapter was originally published in Frontiers in Computational Neuroscience as Mineault et al. (2013). Appendix B presents information relevant to the estimation of local field potentials from wideband signals contaminated with action potential traces; this appendix, which I authored, was originally published as the supplementary information in Zanos, Mineault and Pack (2011).

3. Local field potentials reflect multiple spatial scales in V4

Local field potentials (LFP) reflect the properties of neuronal circuits or columns recorded in a volume around a microelectrode (Buzsáki et al., 2012). The extent of this integration volume has been a subject of some debate, with estimates ranging from a few hundred microns (Katzner et al., 2009; Xing et al., 2009) to several millimeters (Kreiman et al., 2006). We estimated receptive fields (RFs) of multi-unit activity (MUA) and LFPs at an intermediate level of visual processing, in area V4 of two macaques. The spatial structure of LFP receptive fields varied greatly as a function of time lag following stimulus onset, with the retinotopy of LFPs matching that of MUAs at a restricted set of time lags.

A model-based analysis of the LFPs allowed us to recover two distinct stimulus-triggered components: an MUA-like retinotopic component that originated in a small volume around the microelectrodes (~350 µm), and a second component that was shared across the entire V4 region; this second component had tuning properties unrelated to those of the MUAs. Our results suggest that the LFP reflects neural activity across multiple spatial scales, which both complicates its interpretation and offers new opportunities for investigating the large-scale structure of network processing.

3.1 Introduction

Local field potentials (LFP) are low-frequency oscillations in the extracellular electric potential detectable through microelectrode recordings. Although they reflect a variety of electrical phenomena, the total synaptic current in a volume around the microelectrode is the major contributor to LFPs (Buzsáki et al., 2012). They thus offer a complementary signal to spikes, reflecting subthreshold network activity at larger spatial and longer temporal scales than are accessible through single unit recording.

The nature of the relationship between spikes and LFPs is a controversial issue that has attracted much interest (Bauer et al., 1995; Belitski et al., 2008; Eggermont et al., 2011; Gieselmann and Thiele, 2008; Henrie and Shapley, 2005; Hwang and Andersen, 2011; Jia et al., 2011; Katzner et al., 2009; Khawaja et al., 2009; Kreiman et al., 2006; Lashgari et al., 2012, 2012; Liebe et al., 2011; Lindén et al., 2011; Liu and Newsome, 2006; Mukamel et al., 2005; Nir et al., 2007; Rasch et al., 2008; Tsui and Pack, 2011; Xing et al., 2009; Zanos et al., 2011a). In a typical experimental scenario, a well-understood property of multi-unit activity (MUA) is compared and contrasted with that of the LFP. For instance, it has been established, by comparing the orientation tuning of V1 MUA and LFP activity, that the LFP activity in V1 could result from the spiking activity in a small volume around the electrode, on the order of 250 µm (Katzner et al., 2009). Xing et al. (2009) came to a similar estimate by comparing the size of MUA and LFP

receptive fields (RFs) in V1. On the other hand, Kreiman et al. (2006) found that selectivity of LFPs for objects is best explained by hypothesizing an integration radius of a few millimeters.

The difference in the estimated integration radii across experiments may reflect the selection of different components of the LFP for analysis. For instance, the power in the high-frequency gamma band tends to be correlated with spiking activity (Ray and Maunsell, 2011) while the amplitude of the signal at lower frequencies has distinct tuning (Belitski et al., 2008); distinct components may have distinct integration properties (Berens et al., 2008). In addition, the physical size of the area under study, the regularity of its organization, and the correlation structure of its input (Lindén et al., 2011) can influence the properties of the LFP, and this may explain some of the discrepancies in the literature.

Many studies of the LFP have focused on V1 in particular, which is an unusual cortical area in that it is quite large, and it is known to have extremely precise columnar organization based on orientation selectivity (Ohki et al., 2005, 2006) and retinal position (Blasdel and Fitzpatrick, 1984; Hubel and Wiesel, 1977). Thus, LFPs in V1 may be unrepresentative of visual or sensory cortex as a whole. As a step toward understanding the LFP in cortex at large, we have recorded LFPs and MUAs in cortical area V4, a region that occupies an intermediate position in the visual hierarchy.

V4 recordings were carried out in two macaque monkeys, both of whom were implanted chronically with 96-electrode Utah arrays, which allowed us to relate the properties of the receptive fields to their physical location on the cortical surface. Using a sparse-noise stimulation procedure, we found that LFPs in V4 exhibit well-defined receptive fields whose positions change smoothly as a function of position on the cortical surface. However, a detailed analysis of the temporal properties of these signals revealed striking changes in RF position and size as a function of time following stimulus onset, such that the retinotopy of MUAs matched that of LFPs only at a restricted set of time lags.

These results could be explained by a model in which the LFP reflects multiple sources of input: a local, retinotopic input and a distant, shared input that had a similar effect across all of V4. By fitting such a model to our data, we found that the local input exhibited consistent retinotopy that approximated that of the simultaneously recorded MUAs. The shared input arrived at latencies that differed from those of the retinotopic input and that differed substantially between the two animals. These results suggest that the LFP reflects neural activity across multiple spatial scales, which both complicates its interpretation and offers new opportunities for investigating the large-scale structure of network processing.

3.2 Results

3.2.1 Preliminary analysis

We used a sparse noise presentation paradigm to estimate the receptive fields of LFPs and multi-units (MUA). Sparse bar stimuli were flashed on a screen while the animal was rewarded for fixating a static red target (Figure 3-1A). The stimuli were located on a log-polar grid and scaled in length and width proportionally to eccentricity to account for the scaling of neuronal RFs with eccentricity (Motter, 2009).

We first removed the remnants of individual spikes on each electrode, using a Bayesian spike removal algorithm (Zanos et al., 2011b). We then determined whether the amplitude of the LFP or its power in different frequency bands were modulated by the stimulus. We filtered the signal in narrow frequency bands and applied the Hilbert transform to get an estimate of the instantaneous power of the signal as a function of time (Freeman, 2007). In this preliminary analysis, we estimated the receptive fields (RFs) of LFPs using reverse correlation (De Boer and Kuyper, 1968; Marmarelis and Marmarelis, 1978). We used these RFs to predict the signal in a separate validation dataset (see Methods for details). We obtained poor predictions (mean $r = 0.03$ for array 1, $r = 0$ for array 2) for all examined frequency bands (Figure 3-1B and 3-1C, red lines). This is likely due to the short duration of each stimulus; generally, power modulations tend to be visible after sustained stimulation (Khawaja et al., 2009).

We repeated the analysis using the *amplitude* of the LFP in different frequency bands. This yielded considerably better predictions (Figure 3-1B and 3-1C, blue lines) across the 5 lowest frequency bands examined, encompassing the range from 0.5 to 30 Hz (mean $r = 0.24$ for array 1, $r = 0.14$ for array 2). Note that shaded error bars represent ±2 s.d.; the lack of overlap in the lowest 5 frequency bands indicates that the worst fits to the amplitude of the LFPs are nevertheless better than the best fits to the power of the LFP. Thus for the following analyses, we focused on the amplitude of the LFP filtered between 0.5 and 40 Hz.

3.2.2 Receptive field profiles

Figure 3-1D shows a slice of a typical LFP RF, capturing the selectivity for space and orientation at a lag of 70 ms following stimulus onset. Casual inspection reveals little modulation of the spatial structure of the RF with orientation, aside from a scale factor. As this pattern of results was typical of both our LFP and MUA recordings, we assumed for the rest of the analyses that orientation tuning could be treated separately from space-time selectivity. As shown in Figure 3-1E, the consequent reduction in free parameters led to more reliable receptive field estimates (peak z-values shown next to color bars).

LFP spatial RFs were, however, strongly modulated as a function of time following stimulus onset. Figure 3- 2A (top) illustrates the spatial RF of an LFP measured on one array at three different time lags. The LFP was responsive to a large area of the visual field at early time lags (80-100 ms) and a smaller area at later time lags (140 ms). This pattern of changes was typical of LFPs measured on this array, as shown below in more detail. By contrast, the MUA RF measured on the same electrode (Figure 3-2A, bottom) showed little evidence of such a change in size.

In addition to changes in size, LFP RFs frequently appeared to shift their preferred positions as a function of time. Figure 3- 2C illustrates the receptive field of an LFP typical of the second array. Between 70 and 90 ms time lags, the RF shifted its preference towards high eccentricities. At 120 ms, the LFP responded to stimuli at a foveal location, far from the initial RF peak (peak z-value: 9.2; $|z| = 4.5$ corresponds to $p = .001$, corrected for multiple comparisons). Note that the polarity of the response reversed with time; while eccentric stimuli caused a low-latency negative deflection in the LFP signal, foveal stimuli caused a positive deflection at longer time lags. Again, the MUA RF (Figure 3-2C, bottom) showed no evidence of such a change.

To quantify these effects, we fit each RF time slice with a Gaussian, which captured the selectivity of the RF with 4 parameters: preferred eccentricity, preferred angle, radial size and angular size. We estimated the uncertainty in these parameters through bootstrapping. Figures 3-2B and 3-2D show the changes in RF position (top) and size (bottom) for the example LFPs (blue) and MUAs (red). LFP RF parameter changes were highly significant across time lag (blue lines; shaded error bars correspond to 95% confidence intervals). By contrast, parameters for MUAs were comparatively stable across time lag (red lines), partly as a consequence of their shorter duration.

Hence, while the LFPs in the examples were well tuned for space, their spatial RFs were not static over time. As a result, the position and size of LFP receptive fields diverged substantially from those of corresponding MUAs at some time lags.

3.2.3 Array analysis

Given the stability of MUA receptive fields across time lags (Figure 3-2), we refit the data on the assumption that their RFs were separable in time and space (see Methods). We then plotted the estimated preferred eccentricity and polar angle of each MUA according to their location on the array. Figure 3-3A shows the result for a single MUA; here the RF was found at roughly 15^{o} eccentricity along the vertical meridian in the lower visual field (bottom plot). These values were color-coded separately

and displayed at the location of the recording electrode within the array (cross-shaped outlines). For eccentricity (left), blue colors corresponded to small values (foveal locations), while red corresponded to high values (eccentric locations). The same color similarly mapped the range of polar angle (right) from 270^{o} (lower visual field) to 360^{o} (right visual field), which spanned the ensemble of RFs we recovered for this array.

Repeating this process for every electrode revealed the retinotopy of MUAs across 4 x 4 mms of cortex (Figure 3-3B). Here eccentricity and angle for each electrode are coded as described above, with electrodes that did not yield significant RFs represented in white. As expected, preferred eccentricity and polar angle changed smoothly as a function of position on the array, and the direction of the eccentricity gradient, illustrated by a green arrow, was roughly orthogonal to that of polar angle.

LFP RFs exhibited a similar retinotopic organization, as illustrated in Figure 3-3C for the same array. As expected from the examples shown in the previous section, LFP retinotopy changed as a function of time lag. These changes were coherent across the array: at later time lags (140 ms), a greater proportion of the array responded to foveal locations (as shown by the increased representation of dark blue colors, left plot) and angles near 270 degrees (again shown in dark blue, right plot). Figure 3-3D shows that the preferred angle and eccentricity of LFPs best matched those of MUAs at a time lag of 140 ms, as measured by the root-mean-squared (RMS) discrepancy.

As suggested by the example shown in Figure 3-2A, the mean LFP receptive field size changed dramatically as a function of time lag (Figure 3-3E). LFP RFs appeared larger at earlier lags, shrinking in size to a value close to that of the mean MUA receptive field size at longer time lags (dashed line).

Corresponding results for the second array are shown in Figure 3-4. In this case LFP RFs formed a retinotopic map at early time lags (Figure 3-4B, top), consistent with that of MUAs, with the best match occurring at 80 ms (Figure 3-4C). Strikingly, however, a foveal component appeared at later time lags, overtaking the retinotopy of the LFPs completely by 170 ms (Figure 3-4B, bottom). Thus, the majority of LFPs measured in the second array showed biphasic receptive fields similar to that illustrated in Figure 3-2C. The changes in retinotopy were accompanied by modest changes in mean RF size (Figure 3-4D).

These results show that the LFP retinotopy changes with time lag in a concerted fashion across the cortical surface. While at some time lags the LFP retinotopy matched that of the MUAs, at others it considerably diverged. Thus, LFP RFs reflect more than the underlying retinotopy of MUAs.

Furthermore, the relationship between LFPs and MUAs was qualitatively different between the two arrays. For the first array, the retinotopy of the LFPs best matched those of MUAs at late time lags (140 ms); in the second array, LFP RFs were aligned with MUAs at early time lags (80 ms). In addition, our second array showed an array-wide foveal component unseen in the first array.

3.2.4 Robustness of retinotopy

The striking differences between MUA and LFP receptive fields within an array and in the results between arrays could conceivably be a signature of a transient electrical artifact; that is, a source of noise that contaminated recordings on a particular recording day, such as line noise, reward artifact, cross-talk between electrodes, etc. To examine this, we repeated the analyses for data recorded on another day in each array. LFP RFs estimated on other recording days exhibited the same qualitative pattern of shift in retinotopy across time lag characteristic of each array. The day 1 LFP-day 2 LFP RMS discrepancy, averaged across time lags, was 0.8 for both arrays. By contrast, at the optimal temporal lag, the within-day LFP-MUA RMS discrepancy was ~2 (Figures 3D, 4C). The tuning of the LFP is thus consistent across days.

It is also possible that signal processing exaggerated the differences between the two signals. The results presented in Figures 3B and 4A reflect multi-unit activity obtained by full-wave *rectifying* a band-pass filtered (750-3500Hz) signal; we refer to this as the rMUA (cf. Xing et al., 2009). The MUA is also commonly defined by the density of *threshold* crossings in band-pass filtered voltage traces (Katzner et al., 2009); we term this the tMUA. These different definitions could potentially isolate different components of the signal (Supèr and Roelfsema, 2005). We thus repeated our analyses for the tMUA (see Methods for details).

We found similar retinotopies with both measures of multiunit activity, with rMUA-tMUA RMS discrepancies of 0.4 and 0.9 for arrays 1 and 2, respectively. We found fewer significantly tuned electrodes with the threshold method, however, especially in the second array (tMUA: N = 57 in array 1, N = 23 in array 2; rMUA: N = 65 in array 1, N = 69 in array 2). We found that tMUA RFs were slightly (7%), but significantly smaller than rMUA receptive fields ($p < 0.05$ for each array, two-sided Wilcoxon rank sum test). These results are consistent with the rMUA having similar properties to the tMUA, while integrating over a slightly larger cortical area.

Thus, the changing retinotopy across time lags and the qualitatively dissimilar properties of the MUA and LFPs are unlikely due to transient noise sources or to the choice of data preprocessing for the MUA (above) or the LFP (Figures 1B, 1C).

3.2.5 Temporal mixture model

The previous section showed that at some time lags, LFP RFs are organized in a retinotopic fashion similar to MUAs. Yet, this close correspondence between LFP and MUA retinotopy is broken at other time points. In the second array, in particular, the appearance of a foveal component at late time points (Figure 3-4B) hints at the interplay between an MUA-like retinotopic component, which changes from electrode to electrode, and a component tuned for foveal locations, shared by all electrodes.

Figure 3-5 illustrates how the interplay between these two mechanisms might account for the data. LFP RF time slices measured on two electrodes on the same array are plotted in Figure 3-5A. While at early time lags (left and center), the RFs are markedly different, they have similar shapes at later time points (right). These results are explained in Figure 3-5B by the interplay of electrode-specific components (green lines) and a second component (black lines) shared by all electrodes on the same array. Here the changing receptive field positions result from electrode-specific response components that are stronger at early time lags and a shared component that is stronger at later time lags.

More formally, we assumed that the RFs $f^e_{\tau,r,\theta,o}$ as a function of electrode number *e*, time lag τ, eccentricity *r*, angle θ and orientation *o* were given by:

$$f^e_{\tau,r,\theta,o} = a^e_\tau p_{r,\theta,o} + b^e_\tau q^e_{r,\theta,o} \quad (3\text{-}1)$$

Here $p_{r,\theta,o}$ is a component shared by all electrodes on a given array and $q^e_{r,\theta,o}$ is specific to each electrode; they are weighted differentially depending on electrode number and time lag by factors a^e_τ and b^e_τ. The shared component $p_{r,\theta,o}$ could take on any spatial configuration. The electrode-specific component $q^e_{r,\theta,o}$ was constrained to have a Gaussian spatial RF profile and a separable orientation tuning curve.

We fit the temporal mixture model for each array by minimizing the squared error between the model and the data (see Methods for details). The resulting model yielded a highly significant improvement

over a baseline model without the constant component (0.57 vs. 0.45 R^2, p << .001 for array 1; 0.37 vs. 0.30 $R^{2,}$ p << .001 for array 2; F-test).

Based on the model fits, we reconstructed composite LFP RFs that were then fit with Gaussians at every time point to reconstruct retinotopies. These are illustrated in Figure 3-6 for array 1. The simulations replicated the pattern of increased representation of low eccentricity and 270 deg. locations at longer time lags (Figure 3-6A) and the apparent decrease in receptive field size in time (Figure 3-6B).

The underlying mechanism for this switch is illustrated in Figure 3-7. The shared component was triggered by stimuli of any orientation across a fairly broad region of space (Figure 3-7A), with peak selectivity at central locations (~10 degrees eccentricity, 315 degrees polar angle). In absolute terms, both the shared and retinotopic components were strongest at early time lags (Figure 3-7B). However, because the retinotopic component decayed more slowly, it was relatively stronger at late time points (Figure 3-7B, bottom). It follows that at early lags, the observed RFs were both more broadly spatially tuned and biased towards representing central locations than at later lags.

Similar results are shown for the second array in Figure 3-8. The model captured the gradual overtaking of the array by a constant component at later time lags (Figure 3-8A), along with the decrease in receptive field size with increasing lag (Figure 3-8B). Figure 3-9A shows that this shared component was strongly tuned for a foveal portion of the visual field. As with the first array, the retinotopic component peaked at early time lags (Figure 3-9B, middle); unlike the other array, however, the shared component manifested itself mostly at later time lags (Figure 3-9B, top).

Together, these results explain the observed changes in retinotopic organization in terms of a gradual switch in the importance of two distinctly tuned components. The first, retinotopic component was strongest in both arrays at early time lags, while the second, shared component differed qualitatively between the two arrays. Thus, shared components may represent idiosyncratic, large-scale biases in visual representation (Jia et al., 2011), a matter we explore in more detail in the discussion.

3.2.6 Retinotopic component

The extracted retinotopic components are illustrated in Figures 10A and 10B; they were retinotopically arranged in a manner similar to MUAs (Figures 3B and 4A). This link is shown in more detail in Figures 10C and 10D, which compares the positions (top row) and sizes (bottom row) of LFP and MUA RFs measured on the same electrodes. The eccentricity and polar angle of the extracted RFs were similar to those of MUAs measured at the same location, save for a cluster of observations at the bottom right of

the second array (Figure 3-10B). These electrodes had a retinotopic component that was foveal and thus overlapped with the shared component; this made precise estimation of their location and size problematic.

Consistent with previous literature, retinotopic LFP RFs were larger than corresponding MUA receptive fields (Figures 10C and 10D); mean and median sizes are documented in Table 1. We estimated the integration radius of the LFP using the method introduced in (Xing et al., 2009). This involves first estimating the integration radius in visual coordinates, then translating this into cortical coordinates σ_{cLFP} using the estimated cortical magnification factor (*m*), according to the formula:

$$\sigma_{cLFP} = [m^2(\sigma_{vLFP}^2 - \sigma_{vMUA}^2) + \sigma_{cMUA}^2]^{1/2} \quad (3\text{-}2)$$

Here σ_{vLFP} and σ_{vMUA} correspond to the mean size of LFP and MUA RFs in visual coordinates, respectively, and σ_{cMUA} is the integration radius of the MUA in cortical coordinates. Because the retinotopy of V4 is less regular than in V1, magnification can change depending on the position on the cortical surface. We therefore estimated *m* by averaging the magnitude of the gradients of eccentricity and polar angle across the array (see Methods for details). This yielded an integration radius of 300 microns (95% CI: [100,500]) for the retinotopic component of the LFP in array 1 and 400 microns (95% CI: [150,650]) in array 2. Thus, the retinotopic component of the V4 LFP arises from the integration of activity proximal to the electrode, consistent with previous results in V1 (Katzner et al., 2009; Xing et al., 2009).

3.2.7 Orientation and temporal tuning

Additional information about the relationship between the MUA and the LFP may be gained by comparing the orientation and temporal tuning of the two signals. We found that the LFPs were essentially untuned for orientation, with a mean circular variance (CV; Ringach et al., 2002) of 0.95 for the retinotopic component (minimum CV: 0.88) and 0.94 for the shared component across both arrays. On the other hand, some MUAs were tuned for orientation (min CV: 0.65), with a mean CV of 0.87 for significantly tuned MUAs across both arrays. This data is consistent with the idea that the LFP integrates over a larger area than the MUA, although the poor tuning prevents further analysis of the integration radius in the manner of Katzner et al. (2009).

More interesting is the temporal tuning of both signals. Figure 3-11A illustrates the temporal filters of significantly tuned MUAs (blue lines; 50 ms to 240 ms) as a function of their position on the first array.

Temporal filters are stereotyped through the array, with a rapid rise followed by a slower decay. There is some indication of suppression at late time lags (segments below the gray line). The majority of filters have a peak latency of 70 ms (Figure 3-11C), with a minority having a peak around 120 ms. These results are mirrored in array 2 (Figure 3-11B), where the filters are also highly stereotyped, although here the decay appears faster. The peak latency is also centered around 70 ms (mean: 73ms; Figure 3-11D).

These results contrast strongly with the time filters of the retinotopic component of the LFP in array 1 (Figure 3-11E), in which the temporal filters differed dramatically across the array. Moreover, the filters are of considerably longer duration than the corresponding MUA filters, reflecting the fact that the signal is modulated at low frequencies (Figure 3-1B). While the dominant polarity is negative, some filters have roughly equal positive and negative polarity phases (bottom center) or have mostly positive polarity (right middle). The peak latency occurs late compared to the MUA, and varies widely (mean: 118ms; Figure 3-11G). Similar trends are visible for array 2 (Figure 3-11F), with filters showing large variation in shape and polarity. Peak latency is also more broadly distributed than the corresponding MUA data, and longer, with a mean of 90ms.

Thus, while MUAs and the retinotopic component of LFPs have similar retinotopy (Figure 3-10), the two signals diverge strongly in terms of temporal selectivity. LFPs have more sluggish dynamics, longer duration, and longer peak latencies; they are also more variable in shape and polarity than MUAs. One must therefore be careful in interpreting the LFP as a spatially smoothed version of the MUA, since the LFP does not reflect the MUA per say, but rather subthreshold activity (Buzsáki et al., 2012). While in some instances, like retinotopy, the relationship between subthreshold activity and the MUA is sufficiently well understood to make the relationship between LFP and MUA transparent (Carandini and Ferster, 2000), in the case of temporal tuning the relationship is complex, and the LFP and MUA reveal themselves as highly distinct.

3.3 Discussion

3.3.1 General discussion

The local field potential is a complex signal which offers a window into cortical processing at larger spatial and longer temporal scales than those associated with single units. Components of this signal have been shown to correlate with attention (Fries et al., 2001, 2008; Gregoriou et al., 2009), cortical inhibition (Atallah and Scanziani, 2009; Henrie and Shapley, 2005), arousal (Andretic et al., 2005; Van

Swinderen et al., 2004), synchronicity (Gray and Singer, 1989; Mukamel et al., 2005; Nir et al., 2007), and other network phenomena.

While the LFP has proven a highly interesting signal, its interpretation has been marred by our lack of understanding of its biophysical sources and its relationship to spikes. Action potential generation and passive propagation have been well understood for several years (Hodgkin and Huxley, 1952; Koch, 1999). By contrast, LFPs have only recently been modeled in a biophysically detailed fashion (Bedard et al., 2006; Buzsáki et al., 2012; Lindén et al., 2011; Milstein et al., 2009). Modeling studies have unequivocally concluded that the LFP is an intrinsically more complex signal than spikes, reflecting a variety of distinct electrical phenomena (Buzsáki et al., 2012). Lindén et al. (2011) show that the integration radius of the LFP depends both on cortical layer and the correlation structure of the input.

It follows that the structure of the LFP and its relationship to spikes may well vary from area to area in idiosyncratic and unpredictable ways. We thus set out to estimate the receptive fields of LFPs in area V4 of two macaques and compared their properties to those of MUAs. Our results show that, in the context of a sparse noise presentation paradigm (reverse correlation), where the LFP signal is dominated by transient as opposed to sustained activity, the LFP reflects multiple sources of inputs.

In both our subjects, one component of the LFP reflected the underlying retinotopic organization of the cortical sheet (Figure 3-10). This component was strongest at 80-90 milliseconds following stimulus onset (Figures 3-7B and 3-9B), and decayed slowly to baseline at 200-250 ms. It arose from the integration of activity within a patch of cortex of ~350 um. Retinotopic signals were mixed with components shared by all electrodes on a given array. In the case of our first array, the shared component was broadly tuned for space and peaked at early time lags (Figure 3-7). For the second array, the shared component was tuned for foveal locations and peaked at late time lags (Figure 3-9).

What is the source of the shared component? We used state-of-the-art signal processing to eliminate potential signal distortion by analog filters and spike remnants (Nelson et al., 2008; Zanos et al., 2011a; see Methods for details). While it remains possible that the shared component is artifactual, its tuning properties are inconsistent with distortion caused by faulty grounding, for example. In the second array, in particular, we see that the shared component has temporal tuning properties which are very different from MUAs or early LFPs recorded on the array. Careful inspection of the shared component measured at late time lags (Figure 3-4B, 170 ms) shows that there is a small but visible gradient in angular

selectivity from the left to the right of the array; this gradient is not captured by the temporal mixture model (Figure 3-8A).

Hence, the shared components actually change across the array, albeit more modestly than the retinotopic components. We thus hypothesize that the shared components reflect large-scale biases in the input to area V4. Jia et al. (2011) found that one component of low-gamma LFPs in V1 have similar orientation tuning across 4 mm of cortex, independent of the preference of local MUAs. It has been hypothesized that this reflects large scale biases in orientation representation in striate cortex, where orientations aligned with the preferred polar angle of neurons are slightly overrepresented (Freeman et al., 2011). Such large-scale biases in representation, which have no functional role per say, could vary idiosyncratically from animal to animal. We conjecture that this could explain the sharp difference between the arrays in the tuning of the shared common component.

Another potential source of discrepancy in the tuning of the shared component lies in the the sampling of cortical layers. Different layers could be targeted as a function of position due to the curvature of cortex. The changing polarity of the temporal filters in Figures 11E, 11F is consistent with this idea. The second animal was much smaller (5-6 kg) than the first animal (9-10 kg). Consequently, the distance between the lunate and superior temporal sulci at the level of the implant was smaller in the second animal (4-5 mm vs. 6-7 mm) and the cortical surface was more curved. Thus, it is unlikely that the sample of layers is exactly the same in both animals, although this could not be verified histologically. These factors highlight that as an epiphenomenal signal (although see Anastassiou et al., 2011), the LFP is complex and noisy, and care must be taken in its interpretation.

3.3.2 The integration radius of the LFP

These results may bear on the continuing debate regarding the integration radius of the LFP. We and others have demonstrated that different components of the LFP have tuning properties which are consistent with either local (~300 um) or global (several mms) integration of MUA activity (Jia et al., 2011; Katzner et al., 2009; Kreiman et al., 2006; Liu and Newsome, 2006; Xing et al., 2009). We thus conclude that it is not meaningful to speak of *the* integration radius of the LFP, for the simple reason that the LFP reflects *multiple* sources of inputs with different integration scales.

We conjecture that the discrepancies in previous studies are due to signal processing and analysis choices which enhance the relative strength of one component of the LFP over another. An important distinction is that some studies (Katzner et al., 2009; Xing et al., 2009) have examined the amplitude of

the LFP, similar to what is done here, while others (Jia et al., 2011; Kreiman et al., 2006; Liu and Newsome, 2006) have examined power at a selected frequency. The amplitude of the LFP is sensitive to transients in the local field potential, while power is sensitive to sustained oscillatory activity. Lashgari et al. (2012) have found that transient and sustained LFPs in V1 have markedly different tuning properties; this translates into changing relationships with MUA activity. It will be interesting to compare the retinotopy of amplitude and power components of the LFP directly in a paradigm which triggers both stereotyped deflections and oscillatory activity; natural movies may be able to evoke both phase and power modulations (cf. Figure 6, Rasch et al., 2008).

Another potential source of variability in the properties of the LFP may result from the treatment of the dimension of time. Katzner et al. (2009), for instance, analyzed the first component of the Singular Value Decomposition of the temporal-orientation tuning curve. Xing et al. (2009) instead analyzed responses at a latency corresponding to the peak deviation of the LFP. Such choices would not permit analysis of the multi-component temporal responses of the kind we have reported here (Figure 3-2).

It may well be the case that in V1 the major contribution to the LFP is separable with respect to time lag and highly local. In this respect, V1 may be a special case, as its retinotopy is remarkably precise (Blasdel and Fitzpatrick, 1984; Hubel and Wiesel, 1977; Ohki et al., 2005, 2006); by contrast, higher-level areas have less precise retinotopy. Tonotopy in primary auditory cortex is significantly less precise than retinotopy in V1, at least in rodents (Castro and Kandler, 2010). Interestingly, LFP spectrotremporal receptive fields are much more broadly tuned than those of MUAs in both cat (Eggermont et al., 2011) and monkey (Kajikawa and Schroeder, 2011).

Given that the integration radius of the LFP varies with the correlation structure of the input (Lindén et al., 2011), part of the disagreement may reflect genuine inter-areal differences in the LFP. While this complicates the interpretation of the LFP, it may afford an opportunity to study how input correlation structures are forwarded and modified in a hierarchy, with wide implications for our understanding of encoding and decoding neural activity (Averbeck et al., 2006).

3.4 Methods

3.4.1 Task

The recording methods have been described in detail previously (Zanos et al., 2011b). Briefly, we implanted chronic microelectrode *Utah* arrays in area V4 of two macaques (macaca mulatta). Area V4

was identified based on stereotactic coordinates and anatomical landmarks (Ghose and Ts'O, 1997). After recovery, the monkey was seated comfortably in a primate chair (Crist Instruments) and trained to fixate for liquid reward. Eye position was monitored at 200 Hz with an infrared camera (SR Research). All aspects of the experiments were approved by the Animal Care Committee of the Montreal Neurological Institute and were conducted in compliance with regulations established by the Canadian Council of Animal Care.

3.4.2 Signal acquisition and processing

We recorded wideband signals at 10 kHz (bandpass filtered in hardware between 0.07-2500 Hz) over the 96 channels of each Utah array. We monitored the power spectrum of recorded wideband signals on a daily basis to minimize line noise and other artifacts. Recordings were referenced against a ground located on the skull 2-3 cm away from the array. The same recording equipment was kept in place for both animals.

The wideband signal was band-pass filtered between 750 and 3500 Hz, rectified, band-passed between 2 and 40 Hz, and downsampled to 100Hz to form the MUA signal (Xing et al., 2009). For comparison with previous literature, we also computed an alternative MUA based on applying a low threshold (3σ) to the wideband signal bandpassed between 750 and 3500Hz (Katzner et al., 2009). We used a detection deadtime of 1ms, a two-sided threshold, and binned the events at 100Hz (10 ms time bins); this gave similar results to the rectification-based method (see Results, Robustness of retinotopy). Action potential artifacts were removed from the wideband signal using a Bayesian method (Zanos et al., 2011a); the despiked wideband signal was then downsampled and band-pass filtered to produce the LFP (see Preliminary Analysis and Receptive Field Estimation section for filters specific to each analysis).

3.4.3 Stimulus

Each animal was trained to fixate a red spot (2 degree fixation window) while sparse dark bar stimuli were flashed on a uniform gray screen . The monitor was refreshed at a rate of 75Hz; stimuli changed every odd frame (37.5Hz); and each bar stimulus lasted for 6 monitor frames (12.5Hz; 80ms). Stimuli were presented in a single continuous trial, which lasted 25 minutes for the data presented for the first array and 30 minutes for the data presented for the second array. We repeated the experiment on other recording days for each array, with similar results (data not shown). The bar stimuli were placed along a 12x12 polar grid (Figure 3-1A), such that stimuli at the periphery were longer and wider than those near the fovea; bars scaled linearly with eccentricity. The length of the bars was chosen so that no bars touched when presented simultaneously; bar width was set to 0.25 times the eccentricity. 4 different

orientations were used. The grid was confined to the lower right corner of the screen. It spanned 120 degrees of polar angle and 5-40 degrees of eccentricity for array 1 and 3-50 degrees of eccentricity for array 2. On average, 7 bar stimuli were on the screen at any given time.

3.4.4 Preliminary analysis

In a preliminary analysis (Figure 3-1B and 1C), we first downsampled the despiked wideband signal to 200Hz, then filtered it in 7 bands (Freeman, 2007): delta (0.5-4Hz), theta (4-8Hz), alpha (8-12Hz), low beta (12-20Hz), high beta (20-30Hz), low gamma (30-50Hz) and high gamma (50-80Hz). We also took the absolute value of the Hilbert transform of each band-passed filtered LFP to obtain estimates of the instantaneous power in each frequency band.

We split the data into a fit dataset and a validation dataset: for each 5 second block of the data, the first 4 seconds were assigned to the fit dataset and the last second to the validation dataset. We estimated orientation-spatial-temporal receptive fields for both band-pass filtered LFPs and their power using standard reverse correlation on the fit dataset (De Boer and Kuyper, 1968; Marmarelis and Marmarelis, 1978; Figure 1B-E). We applied a Gaussian spatial kernel (σ=0.8) to the estimated receptive fields and predicted the signal in the validation dataset, assuming a linear model with the estimated filter.

3.4.5 Receptive field estimation

The results of the preliminary analysis (previous section, Figures 1B and 1C) showed that the stimulus modulated the amplitude of the LFP but not its power, and that the modulation was concentrated at low frequencies. For the remaining analyses, we thus band-pass filtered the despiked wideband signal in the range (0.5-40 Hz) and downsampled to 100Hz to produce the LFP signal. Inspection of the reverse correlation filters led us to a low-dimensional parameterization for each temporal slice of the LFP receptive fields: the selectivity of the RF is given by the product of an orientation filter and a Gaussian spatial envelope. Specifically, we assumed that the contribution of the stimulus presented τ epochs ago to the internal response in the kth time bin was given by:

$$\eta_k = \sum_{r,\theta,o} s_{k-\tau,r,\theta,o} v_o G_r(r_0,\sigma_r) G_\theta(\theta_0,\sigma_\theta) + d \qquad (3\text{-}3)$$

$s_{k-\tau,r,\theta,o}$ is the stimulus presented τ epochs ago. *d* is a bias. $G_a(b,c)$ is a Gaussian curve evaluated at *a*, centered around *b*, with width (standard deviation) *c*. v_o is the orientation filter. We fit the model for each time slice τ = 1 to 24 by least-squares. We initialized the parameters by fitting a Gaussian to the spatial envelope of the reverse correlation estimate of the filter through least-squares; this envelope

was determined by taking the first singular vector of the SVD of the reverse correlation estimate (Ahrens et al., 2008a).

The model for MUAs was similar, but this time we assumed that the receptive fields did not change in shape across time slices, but were simply scaled by a time-dependent gain:

$$\eta_k = \sum_{\tau,r,\theta,o} s_{k-\tau,r,\theta,o} u_\tau v_o G_r(r_0,\sigma_r) G_\theta(\theta_0,\sigma_\theta) + d \quad (3\text{-}4)$$

Here *u* corresponds to the weights of a separable time filter. We fit the model through least-squares.

An MUA RF was deemed significantly tuned if the fit was significant at the $p < 0.0001$ level according to a χ^2 test (Wood, 2006). We found this criterion too lenient for LFP RFs, presumably because the correlation structure of the LFP did not follow the assumptions of the test. Instead, an LFP RF fit was deemed significant if its R^2 value was greater than the observed R^2 values on any electrode on the same array for time slices from 10 ms to 40 ms. Importantly, this simulation-based method preserves the correlation structure of both data and input, while eliminating the relationship between the two signals (Goldfine et al., 2013), and the resulting threshold corresponds to an effective $p \sim 0.01$.

3.4.6 Temporal mixture fit

The temporal mixture model illustrated in Figure 3-5 was as follows. The LFP RF $g^e_{\tau,r,\theta,o}$ measured on electrode *e* was assumed to be a noisy version of the underlying RF $f^e_{\tau,r,\theta,o}$. The underlying RF was given by a mixture of a shared component $p_{r,\theta,o}$ and a component specific to the electrode $q^e_{r,\theta,o}$:

$$f^e_{\tau,r,\theta,o} = a^e_\tau p_{r,\theta,o} + b^e_\tau q^e_{r,\theta,o} \quad (3\text{-}5)$$

The shared component was unconstrained while the specific component was a Gaussian in space modulated by orientation:

$$q^e_{r,\theta,o} = v^e_o G^e_r(r_0,\sigma_r) G^e_\theta(\theta_0,\sigma_\theta) \quad (3\text{-}6)$$

The temporal mixture model was fit using an iterative least-squares algorithm to minimize the mismatch between $f^e_{\tau,r,\theta,o}$ and $g^e_{\tau,r,\theta,o}$, which was estimated by reverse correlation. The shared component was initialized to the mean of all RFs at t = 90 ms for array 1 and t = 200 ms for array 2. a^e_τ was then set by

least-squares on the assumption that b_τ^e = 0. Then the following steps were alternatively repeated until convergence:

1. The residual $g_{\tau,r,\theta,o}^e - a_\tau^e p_{r,\theta,o}$ was reshaped into a matrix with {e,τ} in one dimension and {r,θ,o} along the second dimension. The first singular values of this matrix were used to determine b_τ^e and $q_{r,\theta,o}^e$. $q_{r,\theta,o}^e$ was then fit to eq.**Error! Reference source not found.**.

2. The first singular values of $g_{\tau,r,\theta,o}^e - b_\tau^e q_{r,\theta,o}^e$ were used to determine a_τ^e and $p_{r,\theta,o}$.

Once the temporal mixture model was fit, we extracted $f_{\tau,r,\theta,o}^e$ for each electrode and time lag and fit the reconstructed RF as an orientation filter multiplied by a Gaussian spatial envelope. The parameters determined through this process are plotted in Figures 6 and 8.

3.4.7 Estimation of the integration radius of the LFP

We applied the method of Xing et al. (2009) to estimate the integration radius of the LFP on the cortical surface σ_{cLFP}:

$$\sigma_{cLFP} = [m^2(\sigma_{vLFP}^2 - \sigma_{vMUA}^2) + \sigma_{cMUA}^2]^2 \quad (3\text{-}7)$$

σ_{vLFP} and σ_{vMUA} correspond to the size of LFP and MUA RFs in visual coordinates, respectively, and σ_{cMUA} is the integration radius of the MUA in cortical coordinates. We used σ_{cMUA} = 100 µm (Xing et al., 2009) and estimated $(\sigma_{vLFP}^2 - \sigma_{vMUA}^2)$ by taking the 20% trimmed mean of this quantity for electrodes where we could measure both LFP and MUA receptive fields.

The cortical magnification factor *m*, measured in millimeters per unit of visual space, captures the change in visual coordinates that corresponds to a unit change in position on the cortical surface. Thus, the local cortical magnification factor corresponds to the inverse of the magnitude of the gradient of the visual quantity measured (log eccentricity or polar angle in our case).

Unfortunately, estimating the magnitude of the gradient of the retinotopies of the MUAs directly is infeasible due to missing measurements. The measured retinotopies were too irregular to be fit reliably with a simple surface such as a plane. Instead, we obtained a smoothed estimate of log eccentricity and

polar angle using Gaussian Process Regression (Rasmussen and Williams, 2006). We used the parameters suggested by the GPML for Matlab toolbox manual (Rasmussen & Williams 2006; Gaussian likelihood, isometric squared exponential covariance, linear + constant mean function, marginal likelihood optimization for hyperparameters, exact inference).

We then computed the magnitude of the gradients of the smoothed surfaces, took their average across the surfaces, and inverted them to obtain the cortical magnification factor for log eccentricity and polar angle for each array.

3.4.8 Orientation and temporal selectivity

We evaluated the orientation selectivity of the retinotopic component of the LFP and the MUA by computing the circular variance of the orientation selectivity coefficients (v_o in eqs. **Error! Reference source not found.** and **Error! Reference source not found.**) as follows (Ringach et al., 2002):

$$CV = 1 - \frac{\left| \sum_{\theta} v_\theta \exp(2i\theta) \right|}{\sum_{\theta} |v_\theta|} \quad \text{(3-8)}$$

For the non-retinotopic component of the LFP, the same formula was used, with v_o being estimated from the first singular vector of the SVD of the spatial-orientation filter.

We used the time mixture parameters b_τ (eq. **Error! Reference source not found.**) as an estimate of the temporal selectivity of the retinotopic component of the LFP (Fig. 11E and F). For the MUA, we opted to take the first singular vector of (spatial-orientation)-temporal filters estimated by reverse correlation as an estimate of the temporal filters. These are more directly comparable to b_τ than u_τ in equation **Error! Reference source not found.**, since u_τ corrects for the slight auto-correlation in the stimulus while b_τ, being ultimately based on a reverse correlation estimate, does not.

3.5 Figures

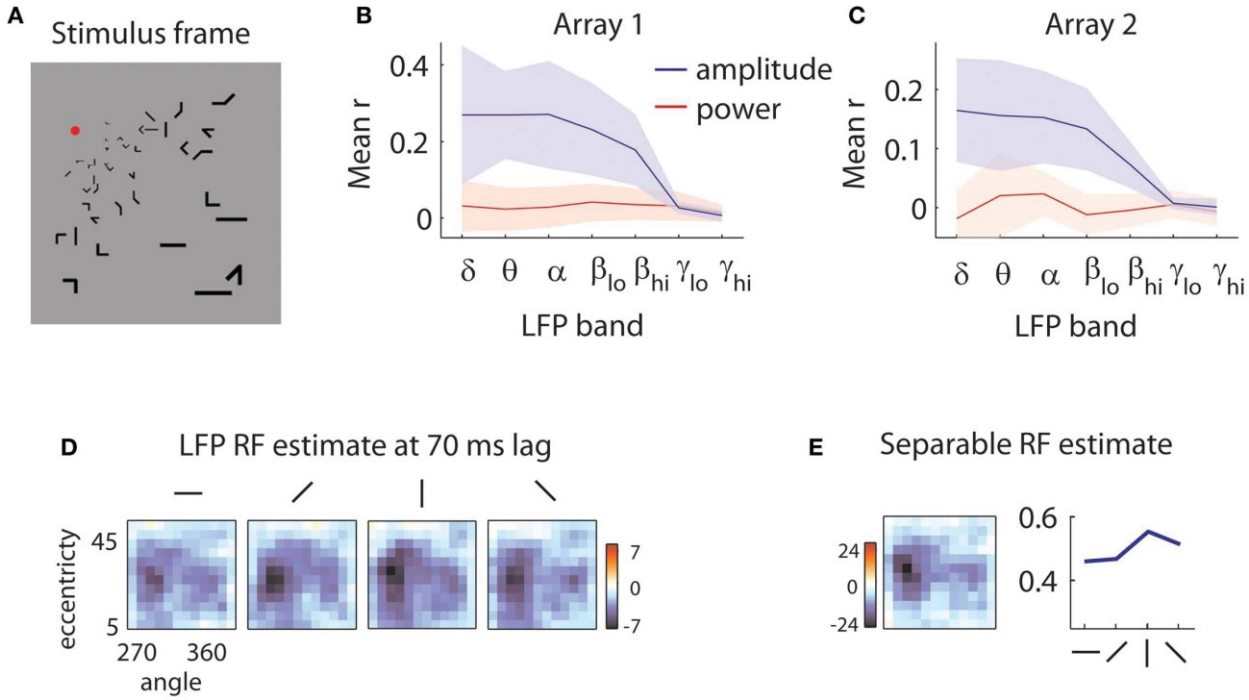


Figure 3-1 - Sample stimulus and receptive field

A) Sample stimulus frame. The stimulus was composed of sparse bars arranged on a log-polar grid to account for the magnification of the visual field with eccentricity. B) Quality of reverse correlation fits to the amplitude of the LFPs (blue lines) and power (red lines) in different frequency bands for the first array. delta: 0.5-4Hz, theta: 4-8Hz, alpha: 8-12Hz, low beta: 12-20Hz, high beta: 20-30Hz, low gamma: 30-50Hz, high gamma: 50-80Hz. Quality of fit was evaluated by the correlation (r) of the predicted and measured responses in a validation dataset. Shaded error bars represent ±1 s.d. Power is not modulated by the stimulus; rather, the stimulus influences the low-frequency components of the amplitude of the LFPs. C) Same as in B, for the second array. D) Sample LFP RF estimate. The RF of the LFP is measured at a 70 ms time lag relative to stimulus onset. Each square represents the spatial RF for a given orientation. The shape of the spatial RF varies little with orientation. The color bar indicates peak z-values estimated through bootstrapping. Here, as in all subsequent RF illustrations, a Gaussian smoothing kernel with σ = 0.7 is applied. E) Separable RF estimate. The RF shown in D) is approximated as separable in space (left) and orientation (right). Little information is lost in the process, and z-values for the spatial envelope of the RF are markedly increased (see legend of color bar).

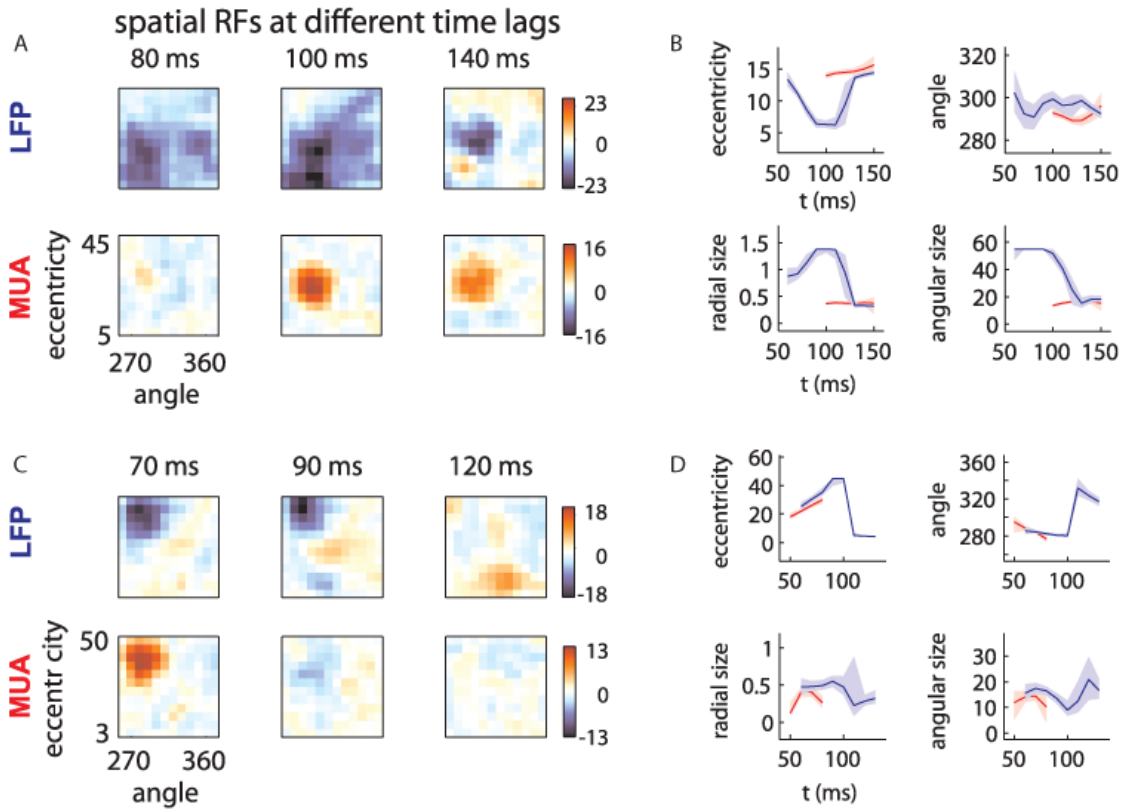


Figure 3-2 - LFP receptive fields change with time lag

A) Spatial envelope of an LFP RF (top) and MUA RF measured on the same electrode (bottom). The measured spatial envelope of the LFP RF (top) becomes markedly smaller at longer time lags, while the spatial envelope of the MUA RF (bottom) is stable. This pattern was typical for electrodes on the first array. B) RF position and size as a function of time. LFP RF parameters (blue lines) corresponding to preferred eccentricity, polar angle, radial size and angular size change markedly as a function of time. MUA RF parameters, illustrated in red, are comparatively stable. Parameters were estimated by fitting a Gaussian to the spatial RFs at different time lags and shaded error bars represent 95% confidence intervals for the parameters estimated through bootstrapping. C) and D): as in A) and B) for an example electrode on the second array. Here, the LFP RF shows a late, foveal excitatory region (120 ms) far from the initial, peripheral preference (70 ms).

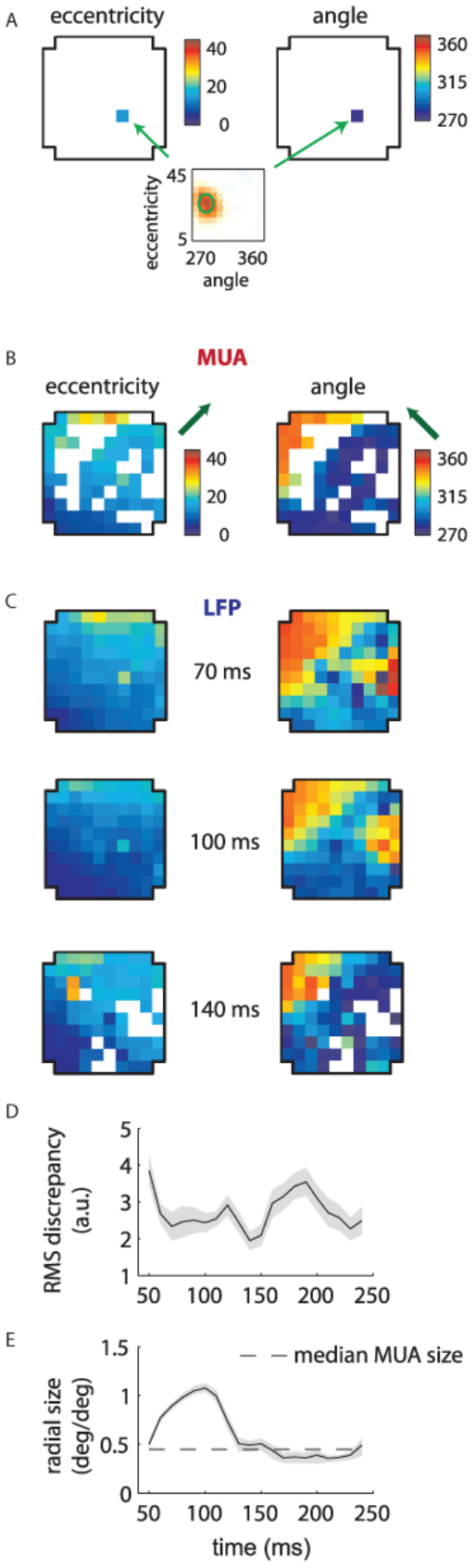

A
eccentricity
angle
40
20
0
360
315
270
45
5
eccentricity
270
360
angle
B
MUA
eccentricity
angle
40
20
0
360
315
270
C
LFP
70 ms
100 ms
140 ms
D
RMS discrepancy (a.u.)
5
4
3
2
1
50
100
150
200
250
E
median MUA size
radial size (deg/deg)
1.5
1
0.5
0
50
100
150
200
250
time (ms)

Figure 3-3 - MUA and LFP retinotopy - Array 1

A) Construction of retinotopies based on measured RFs. The preferred angle and eccentricity of a RF (bottom) is measured by fitting a Gaussian. These measurements are shown using a color code (top) at the location of the electrode on the array (cross-shaped outline). By repeating this process for all electrodes, the underlying retinotopy of the cortical sheet is revealed. B) Measured retinotopy of MUAs. Electrodes yielding non-significant fits are left in white. The preferred eccentricity and polar angle change smoothly across the cortical sheet in a linear gradient (green arrows). C) Measured retinotopy of LFPs as a function of time lag. The retinotopy evolves in a concerted fashion across time lag; at 140 ms, the array appears to represent more foveal locations and more locations around 270 degrees polar angle. D) Root mean square (RMS) discrepancy between MUA and LFP retinotopies as a function of time lag. MUA and LFP retinotopies are best matched at 140 ms time lag. Shaded error bars represent ±2 s.d. E) Mean receptive field size as a function of time lag. The dashed line represent the mean MUA size measured on this array. The angular size of the RFs (not shown) showed a similar effect. Receptive fields are 3 to 4 times larger in linear dimensions at 100 ms compared to late time lags. Shaded error bars represent 95% confidence intervals for the mean.

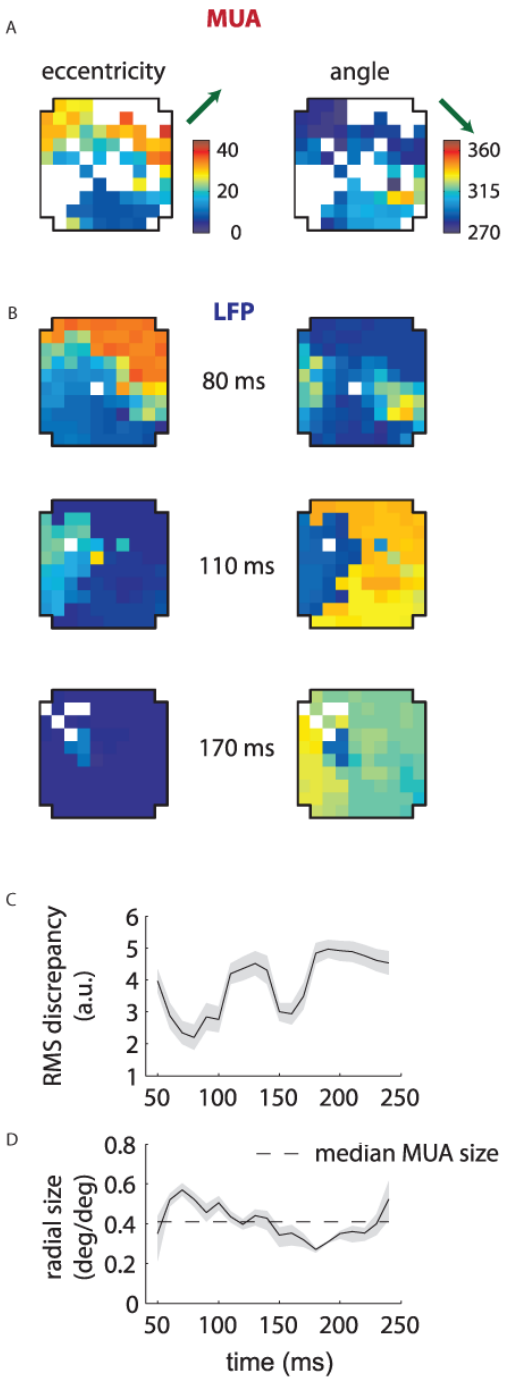


Figure 3-4 - MUA and LFP retinotopy - Array 2

A), B), C) and D) as in Figure 3-3 B), C), D) and E), now for the second array. The second array shows qualitatively different changes in retinotopy as a function of time. In particular, while at early time lags (80 ms) the array forms a smooth retinotopy, at later time lags (170 ms) the entire array represents a foveal location.

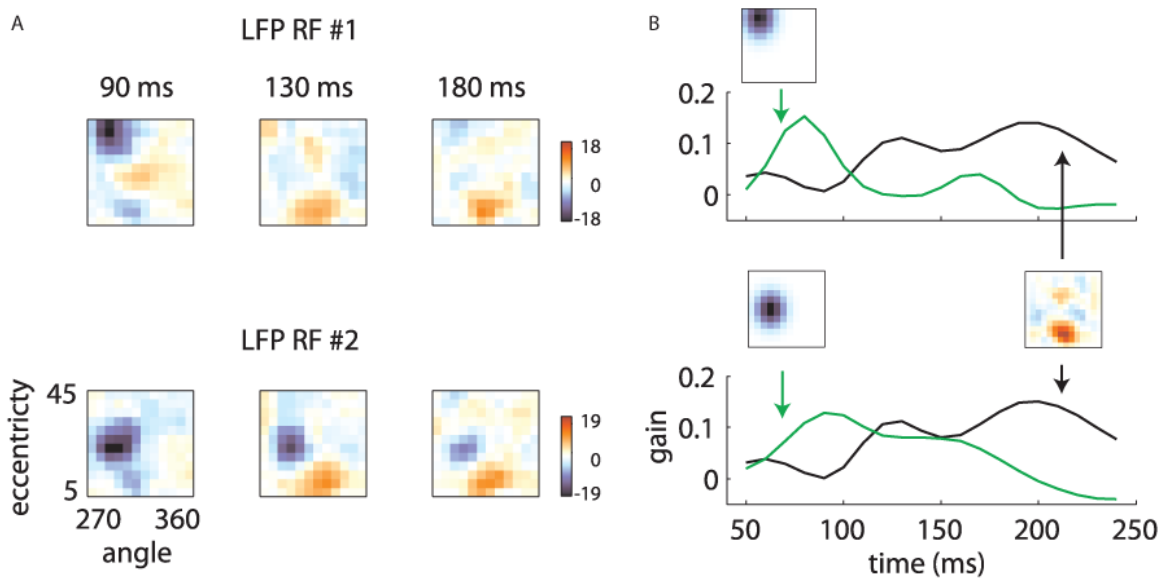


Figure 3-5 - Temporal mixture model

A) Two LFP RFs measured on the same array are illustrated. At early time lags, the RFs are quite different, with the first RF representing high eccentricity locations and the second representing intermediate eccentricities. At late time lags, however, both RFs show a secondary excitatory lobe in a foveal location. B) These results are explained by positing that each RF is a mixture of two components: a retinotopic component, specific to each electrode, constrained to take the shape of a Gaussian, and a shared component, which can take an arbitrary shape but is shared across electrodes. Each RF time slice is obtained by a weighted sum of the electrode-specific retinotopic component (green lines) and the shared component (black line). The relative strength of the two components as a function of time determines the shape of the RFs.

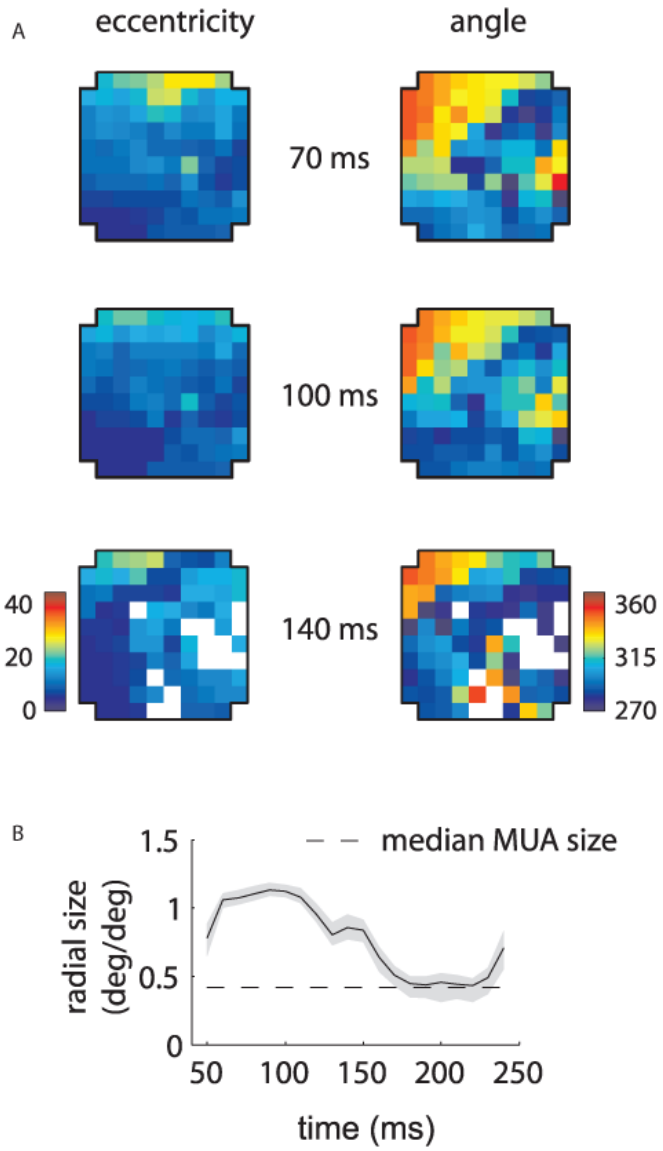


Figure 3-6 - Reconstructed retinotopies based on temporal mixture model - Array 1

A) Reconstructed retinotopy of the array as a function of time. These retinotopic maps were obtained by fitting the temporal mixture model to the array 1 data, creating simulated RFs based on the measured parameters, and fitting the simulated RFs with Gaussians. The mixture model captures the greater representation of low eccentricity and 270 degree locations at late time lags (140 ms). B) Reconstructed mean RF size as a function of time. The model captures the dramatic change in measured RF size as a function of time. Shaded error bars represent 95% confidence intervals for the mean.

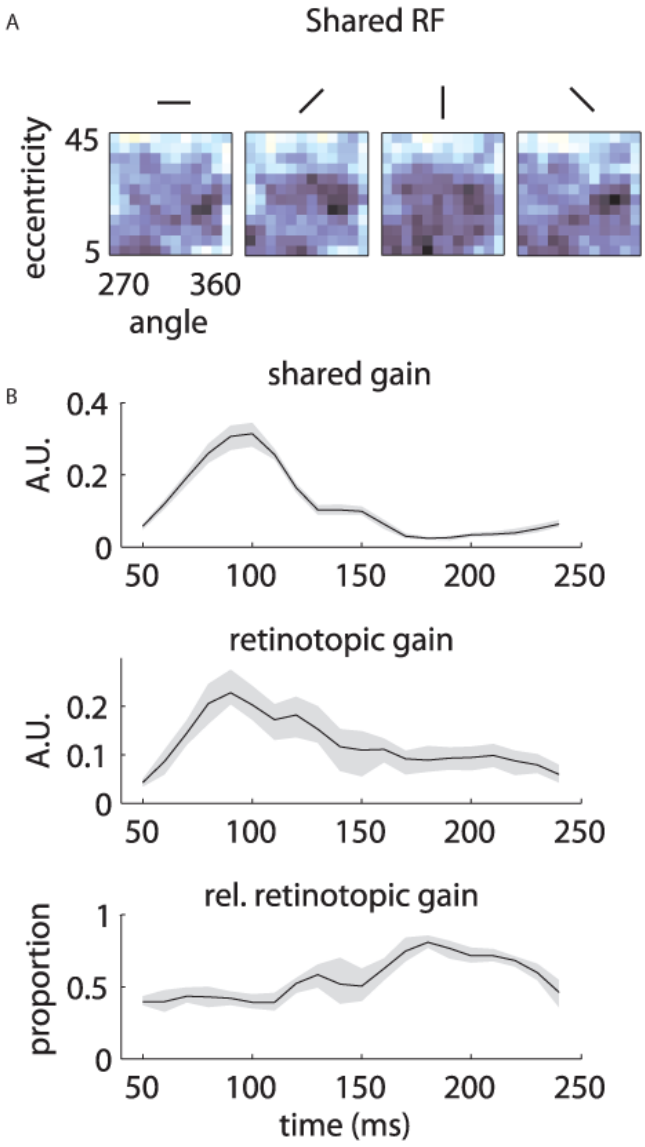


Figure 3-7 - Temporal mixture model parameters - Array 1

A) Shared receptive field estimated from the data. The RF is broadly tuned for space and orientation. B) Gains of each component as a function of time. Both the median shared gain (top) and the retinotopic gain (middle) peak at early time lags. However, the retinotopic gain decays more slowly as a function of time. Therefore, the retinotopic gain is relatively larger at late time lags (bottom). This creates a shift in the representation from broadly tuned (shared component) to more tightly tuned (retinotopic component). Shaded error bars represent 95% confidence intervals for the median.

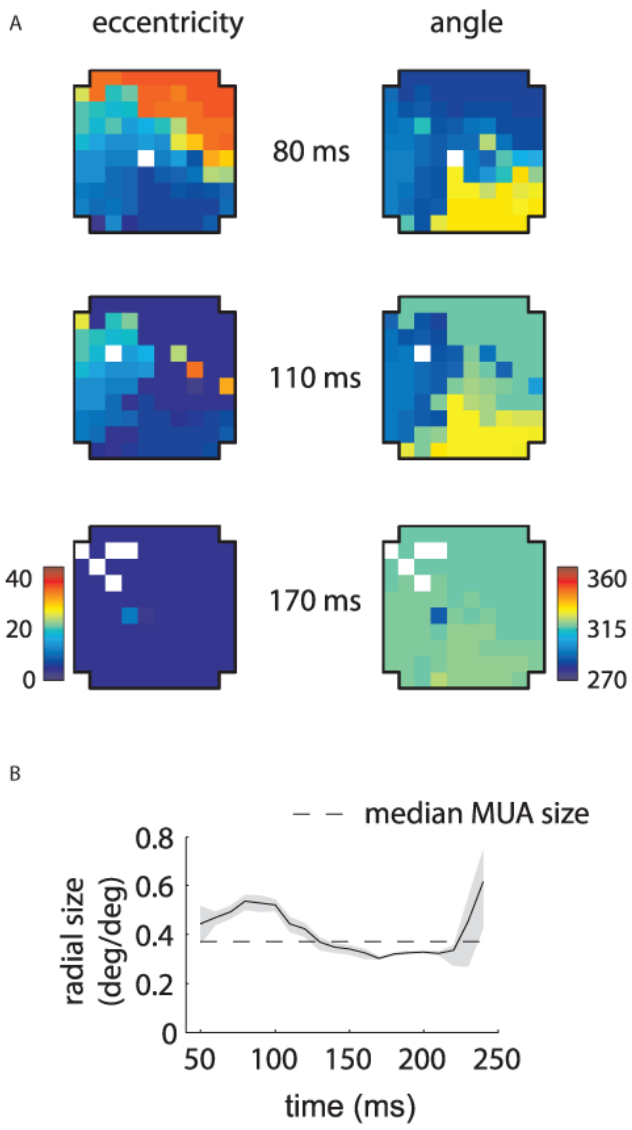


Figure 3-8 - Reconstructed retinotopies based on temporal mixture model - Array 2

A) and B) as in Figure 3-6. The model captures the change in the representation from retinotopic at early time lags to exclusively representing foveal locations at late time lags.

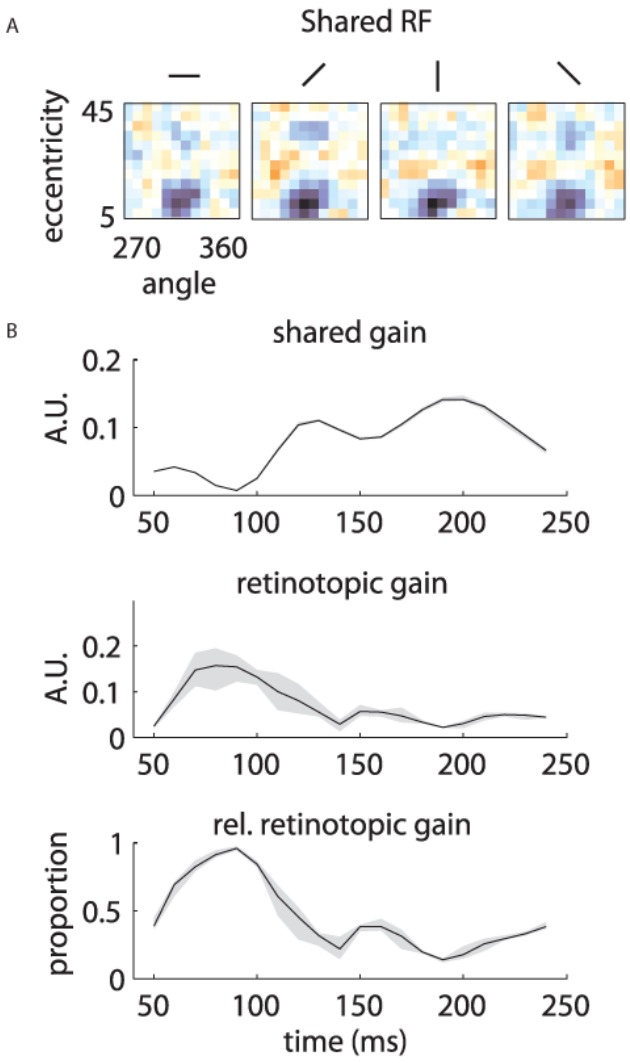


Figure 3-9 - Temporal mixture model parameters - Array 2

A) and B) as in Figure 3-7. The shared RF is tightly tuned for foveal locations. The gain of the shared RF grows larger at late time lags. Therefore, the RFs switch from an early retinotopic to a purely foveal late representation.

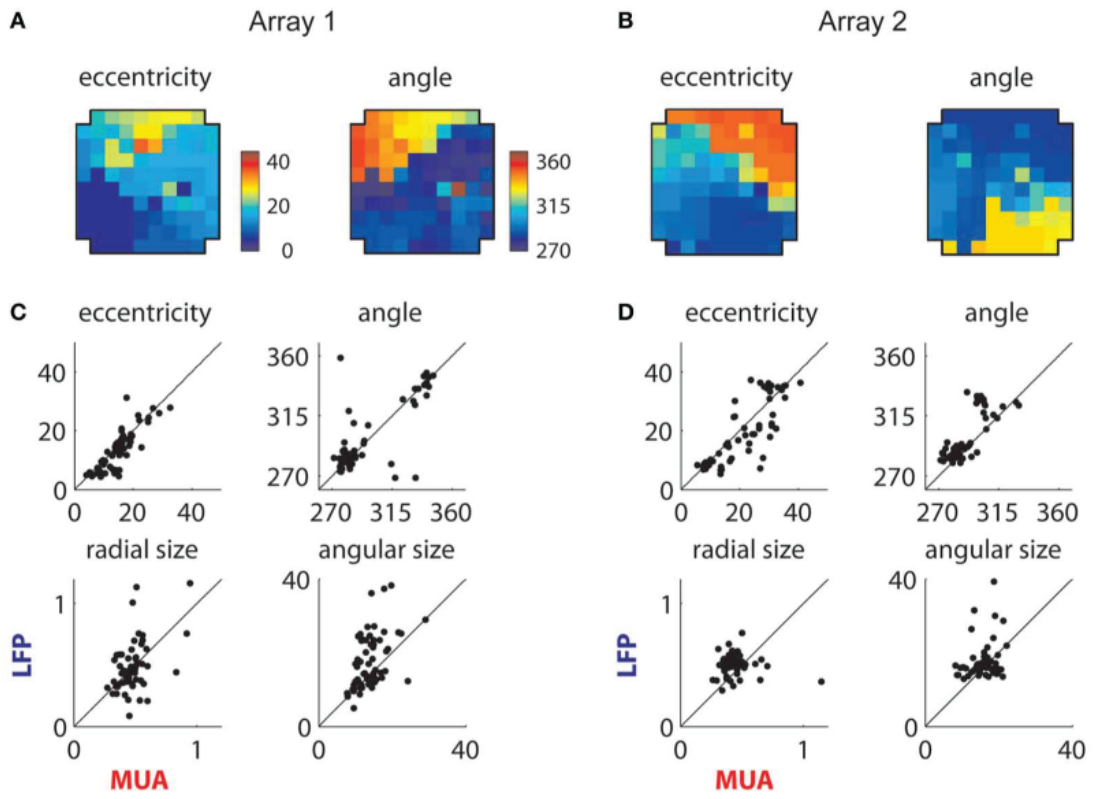


Figure 3-10 - Retinotopic components

A) Measured retinotopy of retinotopic component for Array 1 and B) for Array 2. These can be compared to the corresponding MUA-based estimates in Figures 3B and 4A. C) Receptive field parameters of LFPs and MUAs measured on the same electrode compared for Array 1 and D) Array 2. Eccentricity and angle match between MUAs and LFPs, while LFPs display larger receptive fields on average.

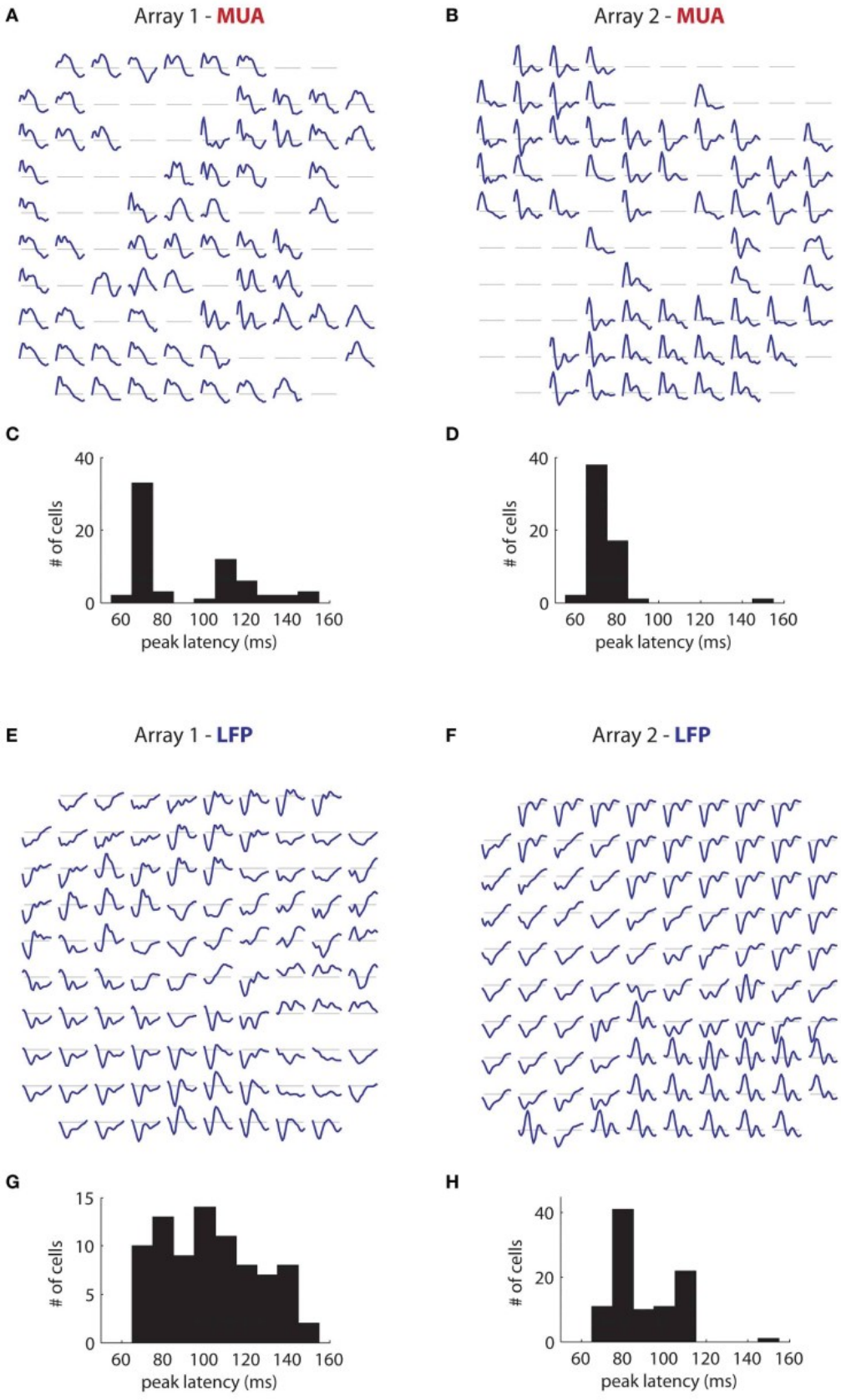

A
Array 1 - MUA
B
Array 2 - MUA
C
# of cells
peak latency (ms)
D
# of cells
peak latency (ms)
E
Array 1 - LFP
F
Array 2 - LFP
G
# of cells
peak latency (ms)
H
# of cells
peak latency (ms)

Figure 3-11 - Temporal filters

A) Estimated temporal filters for the MUAs measured on the first array and B) on the second array. Filters span 50 to 240 ms time lags. Gray lines correspond to a coefficient of 0. Filters are highly stereotyped across each array. C) distribution of peak latency for the MUAs for the first array and D) the second array. Peak latency is centered around 70ms. E) and F), same as in A and B, but for the retinotopic component of the LFPs. LFP temporal filters vary much more than corresponding MUA temporal filters, Their duration is generally longer, and they are smoother. G) and H) LFP peak latencies. These are more widely distributed than corresponding MUA peak latencies, and generally longer.

3.6 Tables

	Radial size - median	Radial size - mean	Angular size - median	Angular size - mean
LFP	0.50 (0.50,0.51)	0.54 (0.51,0.57)	16.6 (16.2,16.8)	18.0 (17.2,19.0)
MUA	0.44 (0.42,0.45)	0.46 (0.44,0.48)	14.0 (14.4,15.7)	15.0 (14.4,15.7)

Table 3-1 - Summary statistics of measured RF sizes

(in parentheses: 95% confidence intervals estimated through bootstrapping)

Building on the parametric modeling framework developed in Chapters 2 and 3, I now turn to the analysis of signals at still smaller scales: single-neuron responses in area MST of the dorsal visual stream. Results reveal that MST responses can be explained by the integration of their afferent input from area MT, provided that this integration is nonlinear. Estimated models reveal long suspected, but previously unconfirmed receptive field organization in MST neurons that allow them to respond to complex optic flow patterns. This receptive field organization and nonlinear integration allows more accurate estimation of the velocity of approaching objects from the population of MST neurons, thus revealing their possible functional role in vergence control and object motion estimation. This manuscript was originally published in Proceedings of the National Academy of Sciences as Mineault et al. (2012). Appendix C contains information on the technical aspects of the receptive field estimation method; it was originally published as the supplementary information in Mineault et al. (2012).

4. Hierarchical processing of complex motion along the primate dorsal visual pathway

Neurons in area MST of the primate visual cortex respond selectively to complex motion patterns defined by expansion, rotation, and deformation. Consequently they are often hypothesized to be involved in important behavioral functions, such as encoding the velocities of moving objects and surfaces relative to the observer. However, the computations underlying such selectivity are unknown. In this work we have developed a novel, naturalistic motion stimulus and used it to probe the complex selectivity of MST neurons. The resulting data were then used to estimate the properties of the feedforward inputs to each neuron. This analysis yielded models that successfully accounted for much of the observed stimulus selectivity, provided that the inputs were combined via a nonlinear integration mechanism that approximates a multiplicative interaction among MST inputs. In simulations we found that this type of integration has the functional role of improving estimates of the three-dimensional velocity of moving objects. As this computation is of general utility for detecting complex stimulus features, we suggest that it may represent a fundamental aspect of hierarchical sensory processing.

4.1 Introduction

In the early stages of the primate visual system the receptive fields of neurons can be readily estimated from the responses to simple stimuli such as spots, bars, and gratings, or even by hand-mapping (DeAngelis et al., 1993; Hubel and Wiesel, 1962; Movshon et al., 1978). However for neurons farther along the visual pathways, the relationship between stimulus input and neuronal output is often far from obvious, particularly in areas that respond to complex stimuli such as faces, objects, or optic flow patterns (Brincat and Connor, 2004; Freiwald et al., 2009; Pasupathy and Connor, 2001; Yu et al., 2010). Uncovering this relationship is crucial for understanding the computations that underlie important behavioral functions such as object recognition and navigation.

One well-known example of complex cortical processing is the range of selectivities found in the medial superior temporal (MST) area of the primate visual cortex. Previous work has shown that MST neurons are highly selective for visual stimuli comprised of combinations of motion patterns such as expansion, deformation, translation, and rotation (Duffy and Wurtz, 1991b; Graziano et al., 1994; Orban et al., 1992; Raiguel et al., 1997; Tanaka et al., 1986). Although this selectivity has been documented many times over the last 25 years, very little is known about the computations by which it is derived. One prevalent hypothesis is that the selectivity of MST neurons is determined by specific strategies used by

the brain to calculate one's direction of motion, or heading, through the world (Lappe, 2000; Perrone and Stone, 1994; Perrone and Stone, 1998). In these models, heading is computed by combining the output of detectors tuned to specific motion patterns, and these patterns are reflected in the internal structure of an MST neuron's receptive field.

While this hierarchical account of MST selectivity is appealingly simple, it has been difficult to confirm experimentally. Indeed previous studies have concluded that MST responses to complex stimuli often cannot be predicted, even qualitatively, from their responses to simple ones (Duffy and Wurtz, 1991b; Tanaka et al., 1989; Tanaka et al., 1986; Yu et al., 2010). For example a recent paper by Yu et al. (Yu et al., 2010) found that MST receptive field substructure failed to account for the response patterns of MST neurons to combinations of motions. This led the authors to speculate that highly complex interactions must occur among MST inputs, perhaps involving specific wiring of dendritic compartments. Such findings call into question the simple hierarchical scheme that has been at the heart of most previous models.

In this work we have examined the hierarchical nature of MST processing using a novel experimental stimulus and a rigorous computational framework. Specifically, we have developed a visual stimulus that efficiently and thoroughly explores the space of complex motion stimuli and used the resulting data to test MST models with different structures. We find that the most successful models take into account the specific properties of MST's most proximal source of afferent input, the middle temporal (MT) area (Albright, 1984; Maunsell and Van Essen, 1983; Rust et al., 2006). Furthermore, we find that such hierarchical models are capable of capturing all of the main features of MST stimulus selectivity, provided that a particular style of nonlinear integration is used to transform MT inputs into MST outputs. We show that this mechanism is consistent with the known properties of cortical neurons and that it can be expressed in a simple mathematical form. Finally we demonstrate in simulations that this type of integration is useful for extracting the three-dimensional velocity of objects relative to the observer, as it provides strong tuning for velocity with little dependence on other stimulus features. This work therefore provides quantitative validation of a number of existing notions about MST function, while supplying a crucial element (nonlinear integration) that has been previously missing.

4.2 Results

4.2.1 MST neurons are tuned to complex optic flow

We recorded from 61 neurons in area MST of two awake, fixating macaque monkeys. In most cases we first obtained an estimate of the neuron's selectivity for optic flow by measuring responses to the *tuning curve* stimuli depicted in Figure 4-1A. For a given position in space, 24 tuning curve stimuli were presented, with 8 stimuli corresponding to translation (motion in a single direction), 8 to spirals (including expansion, contraction, rotation and their intermediates), and 8 corresponding to deformation (expansion along one axis and contraction along the other). These tuning curve stimuli span the space of first-order optic flow patterns, and have proven useful in characterizing optic flow selectivity in the dorsal visual stream (Lagae et al., 1994; Orban et al., 1992). These 24 tuning curve stimuli were presented at 9 positions lying on a 3x3 rectangular grid that spanned most of the central 50^{o} of the visual field, allowing us to examine the positional invariance of the selectivity (Duffy and Wurtz, 1991a).

Figure 4-1A shows the responses of an example MST cell to the 24 optic flow stimuli when they were displayed in the lower-middle part of the 3x3 grid. Here the cell preferred downwards-translational motion (left panel), contracting counterclockwise spirals (center panel), and deformation with a horizontal divergent axis (right panel). These responses are replotted in Figure 4-1B as *tuning mosaics*, which are color-coded versions of the standard direction tuning curves. Each mosaic shows the response of a cell to 8 stimuli of a type at a given position in the receptive field, with reds representing responses above baseline firing rate and blue responses below baseline. The most saturated red corresponds to maximal firing rate across all stimuli, while white corresponds to the median firing rate; tuning mosaics are not otherwise normalized. The mosaics highlighted in green correspond to the tuning curves shown in Figure 4-1A.

The translation mosaics (Figure 4-1B, left) indicate that this cell shows a preference for downwards motion in the bottom and center portions of the screen. The spiral mosaics (Figure 4-1B, center) show that the cell's spiral tuning shifts from position to position, with the strongest preference being for expansion motion at the top and center of the visual field. A weaker response to contraction can be seen near the bottom of the visual field. The deformation mosaics (Figure 4-1B, right) show that tuning for deformation motion also varies from position to position. This cell therefore shows selectivity for a range of stimuli, and a strong dependence of stimulus preference on spatial position.

In contrast, Figure 4-1C shows a cell with tuning for expansion (center panel) that is nearly invariant with spatial position. This second cell's translation tuning (left panel) is similar to that of the cell in Figure 4-1B, indicating that there is no obvious relationship between the tuning for translation and that for spirals. Thus our results, in agreement with previous reports (Duffy and Wurtz, 1991b; Tanaka et al., 1986), suggest that MST neurons exhibit complex tuning in a high-dimensional stimulus space. In order to explore this tuning in quantitative detail, we developed a stimulus that sampled the space of optic flow far more thoroughly than the tuning curve stimulus described above. Specifically, we used a *continuous optic flow* stimulus that consisted of continuously evolving, random combinations of translation, spirals, and deformation stimuli, each of which elicited robust responses from most MST neurons (sample stimulus viewable at http://www.youtu.be.com/l0QNtKGL7vU). This approach typically allowed us to measure responses to several thousand optic flow stimuli.

Based on the responses to this rich repertoire of stimuli, we sought to develop a quantitative account of the neuronal computations that lead to the variety and complexity of neuronal responses exemplified in Figure 4-1. Our approach was to describe each neuron's responses using several mathematical models, all of which shared the same basic structure. In the first stage, the input stimulus is processed by a number of subunits, each of which is selective for motion in a part of the visual field. The output of these subunits is fed to the simulated MST neuron, which sums its inputs and translates the result into a predicted firing rate through an expansive static nonlinearity. Such linear-nonlinear cascade models have strong theoretical foundations that have been described elsewhere (Paninski, 2004; Wu et al., 2006).

For each MST neuron we optimized the choice of subunits so as to maximize the quality of the fit to the *continuous optic flow* data (see Methods for details). We controlled the complexity of the model by cross-validation, and evaluated its performance by predicting a neuron's response to the *tuning curve* stimuli, on which the model was not trained. As a check on the validity of our approach and its implementation, we verified that our methods converge to correct estimates of receptive fields in simulated data (Appendix C and Figure C-1). As described in detail below, our approach allowed us to examine particular hypotheses about neuronal computation in MST.

4.2.2 Hierarchical processing partially accounts for MST responses

The simplest model that could in principle account for the data shown in Figure 4-1 involves a computation in which MST neurons linearly compare the visual stimulus to an internal template, with the output reflecting the degree of match. This *linear model* is directly analogous to the linear

spatiotemporal receptive field models that have been used in the luminance domain to study early visual areas (Carandini et al., 2005; DeAngelis et al., 1993). Furthermore, it is mathematically tractable , and previous modeling work has shown promise in capturing the complex tuning properties seen in MST (Poggio et al., 1991; Zhang et al., 1993). We found, however, that while such a model can capture some preference to translation, it is unable to capture the more complex selectivities of MST neurons (see Appendix C and Figure C-2).

This may be expected, as MST neurons have no direct access to the visual stimulus, instead receiving the bulk of their input from MT neurons, which are tuned for both direction and speed (Boussaoud et al., 1990; Maunsell and Van Essen, 1983). Thus a more promising model involves a computation in which MST neurons linearly sum the output of appropriately tuned MT subunits. Indeed this idea is implicit in many existing MST models (Duffy and Wurtz, 1991b; Grossberg et al., 1999; Lappe, 2000; Perrone and Stone, 1994; Tanaka et al., 1989). We thus developed a *hierarchical model* in which the input stimulus is first transformed into the outputs of a population of MT-like subunits tuned for stimulus direction and speed (Figure 4-2A). The mathematical form of these subunits was chosen to provide an accurate and parsimonious account of the responses of real MT cells. Specifically, MT subunits had receptive fields that were smaller than those found in MST and responses that were tuned for direction and speed, with bandwidths matching those found in real MT cells (see Methods for details).

Figure 4-2B shows the predicted tuning curves under this hierarchical model for the example cell shown in Figure 4-1B. In this case the model captures the tuning, including the general preference for downward translation (left) and the variety of selectivities for spiral and deformation motion (middle and right). The quality of the prediction can be assessed using $\bar{R}^2$, the proportion of explainable variance accounted for (Sahani and Linden, 2003; see Methods for details). For this example cell, $\bar{R}^2 = 0.55$, which compares favorably with results reported previously in other areas (Cadieu et al., 2007; David and Gallant, 2005; Mante et al., 2008; Willmore et al., 2010). Across the MST population, however, the model fared considerably worse, with median $\bar{R}^2 = 0.31$. Indeed we found some cells with tuning characteristics that could not be explained even qualitatively with this model structure, and the neuron originally shown in Figure 4-1C is an example of this category. Figure 4-2C shows that, while the hierarchical model successfully captures this cell's tuning for translation (Figure 4-2C, left panel), it consistently underestimates the responses to spiral stimuli (Figure 4-2C, center panel). This pattern of errors in the hierarchical model was common across our population of cells, being present in 58% of the cells (21/36, stimulus class comparisons, $p<0.001$; see Methods for details). Thus we conclude that,

while a hierarchical model can account for some MST tuning properties, there is strong evidence that such a model responds too strongly to translation and too weakly to complex optic flow.

4.2.3 Nonlinear integration is necessary to explain MST stimulus selectivity

Stated in more general terms, the stimulus selectivity of the hierarchical MST model is too similar to that of its inputs, and there appears to be no spatial arrangement of inputs that can bring this model into closer agreement with the data. This suggests that MST selectivity requires a nonlinear operation that transforms the output of one area prior to summation by the next (DeAngelis et al., 1993; Rust et al., 2006; Tsui et al., 2010); indeed such a mechanism been proposed in other contexts throughout the primate visual system (Anzai et al., 2007; Brincat and Connor, 2004; Cadieu et al., 2007; Movshon et al., 1978). We therefore examined the consequences of adding a nonlinearity (Figure 4-3A) that shaped the output of each MT subunit. In particular, we added a flexible, static nonlinearity, represented by a single free parameter β, that could be either compressive (β < 1) or expansive (β > 1; see Methods for details). For each MST cell the nonlinearity was constrained to be identical across subunits.

Remarkably, this minimal change to the hierarchical model structure yielded far better fits to the data (Figure 4-3B) for the expansion-selective cell originally presented in Figure 4-1C. In particular, the *nonlinear integration model* showed enhanced responses to optic flow stimuli such as expansion and rotation, while maintaining strong tuning for translation, with an overall increase in the goodness-of-fit from a $\bar{R}^2$ of 0.41 to 0.70. This improved fit to the data was not a trivial consequence of the additional free parameter, as the model was evaluated with a validation procedure (defined in Methods) that was robust to the overall model complexity. Figure 4-3C shows that predictions improved for the majority (75%) of MST cells from which we recorded, with the median goodness-of-fit improving from 0.31 to 0.50. These improvements are also reflected in the cross-validated goodness-of-fit measured with the continuous optic flow stimulus, shown in Figure C-3B. Similar results were obtained if we allowed each subunit to have its own nonlinearity (Table 4-1, unrestricted nonlinear model; see also Appendix C), suggesting that the shared nonlinearity is sufficient.

In principle there are two ways in which the introduction of nonlinear integration could improve the fit of the model to the data. The first would be to increase the overall *level* of responses to spiral and deformation stimuli relative to translation stimuli, while preserving the shape of tuning curves within stimulus categories. This would compensate for the above-mentioned tendency of the hierarchical model to underestimate firing rates for spiral and deformation stimuli. The second would be to improve

the ability of the model to match the *shape*s of the tuning curves, apart from overall response levels for individual stimulus classes. To untangle these two factors, we performed an additional analysis after first normalizing the responses within each stimulus class (translation, spirals, and deformation). Figure C-3C shows that the nonlinear integration model still improves the quality of predictions in 78% (28/36) of the cells (stimulus class comparisons; see Methods for details). This indicates that the nonlinear integration model captures aspects of the MST responses that cannot be related simply to stimulus-specific level modulation. Rather the nonlinear integration mechanism is necessary for producing the stimulus selectivity seen in MST responses to optic flow.

We also verified that the success of the model was not influenced by errors in the centering of the stimuli, stimulus position profoundly affects MST stimulus selectivity (Lagae et al., 1994). We estimated receptive field centers from the tuning curve stimuli and compared the quality of model fits for recordings in which the stimuli were well-centered (within 7 deg. of the centers) and those in which the centering was worse (12.4 deg. on average). The addition of the nonlinearity improved the model fits for both groups of neurons (15/19 in the first group, 12/17 for the second group), indicating that our conclusions about nonlinear integration are robust to stimulus centering. Indeed the results were noticeably better when the stimulus was well-centered (median $\overline{R}^2$ = 0.56 for well-centered cases, 0.36 when the centering was worse), which indicates that the model captures the bulk of the selectivity in the center of the receptive field.

4.2.4 Substructure of MST receptive fields

The success of the nonlinear modeling approach allowed us to examine the types of subunit arrangements that were recovered for each neuron. Figure 4-4A shows the subunits that contribute most critically to the highly nonlinear neuron shown in Figure 4-3B (see Appendix C for details). Each circle in the figure corresponds to the position and size of a single MT subunit's receptive field; the direction of each arrow indicates the preferred direction of the subunit; the opacity of the color indicates the weighting; and the color denotes the sign of the contribution, with red being excitatory and blue being inhibitory. The results of this analysis show that this MST neuron's response is largely explained by the selectivity of subunits tuned to downward-left motion in the bottom left portion of the visual field and downward-right motion in the bottom right. This is consistent with this cell's tuning for both expansion and downwards motion.

For some MST cells the subunit nonlinearity was less critical, and an example of this type of receptive field is illustrated in Figure 4-4B (same cell as in Figure 4-1B). Here the cell's receptive field is summarized by a single downwards-tuned, centrally located subunit. This cell's nonlinearity had an exponent of 0.6, closer to unity than most neurons in the MST sample (see below for details); the quality of the prediction went from 0.55 to 0.62 with the additional nonlinearity, a comparatively small change. Thus this MST cell's response properties were similar to those found in MT.

The receptive fields of three more MST neurons are shown in Figure 4-4C-E. Like the cell originally shown in Figure 4-1C, these three cells are selective for expansion at multiple positions in the visual field. However, despite the similarity in the tuning, the most critical subunits of these neurons revealed a variety of receptive field substructures. In particular, the position and relative motion directions of the subunits varied substantially from cell to cell, suggesting that these MST cells are not detectors of expansion *per se*. Rather, the selectivity of these cells appears to be captured by nonlinear combinations of a small number of excitatory and inhibitory inputs,. Estimated time filters and additional examples of receptive fields are shown in Figure C-6.

As can be seen in Figure 4-4, another prominent feature of MST receptive fields is the spatial overlap of the subunits. Although differences in direction and speed preference tended to increase with spatial distance between subunits (Figure C-8D,E), there was also substantial variation on spatial scales smaller than a single subunit (e.g., Figure 4-4C). This variation may be important for estimating optic flow quantities such as motion parallax, in which multiple motion vectors occur at nearby spatial locations. More generally, the complex selectivity observed here is likely to be useful in natural contexts, in which motion patterns are determined in part by the structure of the surrounding environment and hence are not constrained to resemble the canonical flow fields typically used experimentally. Overall these results parallel the finding that selectivity for analogous stimuli (e.g., non-Cartesian gratings) in the ventral stream of the visual cortex is related to selectivity for combinations of orientations or other features (Cadieu et al., 2007; Gallant et al., 1993; Pasupathy and Connor, 2001).

As suggested by Figure 4-4, the number of subunits recovered by the model differed from cell to cell. This variability is summarized in Figure 4-4F, which shows that the number of subunits contributing significantly to individual MST neurons ranged from 2 to 45, with a median value of 9 (see Appendix C for details). Most of these subunits were excitatory, with a median proportion of excitatory subunits of 81% across our population of cells. The remaining inhibitory subunits can be interpreted either as removal of excitation from tonically active MT cells or as indirect MT influences via MST interneurons, as

inter-areal projections are almost exclusively excitatory. These conclusions are of course contingent upon the assumptions underlying our modeling approach. However, for the most part these assumptions are quite conservative, and, as we show in the next section, relaxing them does not change the main results.

4.2.5 Importance of compressive nonlinearities across the MST population

While Figure 4-4 shows that the receptive field substructure varied substantially from cell to cell, we found that the shape of the nonlinearity recovered by the model was highly consistent across neurons. This is illustrated in Figure 4-5A, which plots the distribution of the parameter β for all the cells in our MST population. The distribution is heavily skewed towards values less than 1, as shown earlier in individual examples, suggesting that a *compressive* input nonlinearity is an important property of MST neurons.

4.2.6 Influence of surround suppression

Given the importance of the compressive nonlinearity in accounting for the MST data, we next sought to relate it to potential physiological mechanisms. One important candidate mechanism is *surround suppression* at the level of MT (Allman et al., 1985; Raiguel et al., 1995; Xiao et al., 1995; Xiao et al., 1997). Surround suppression attenuates the responses of MT neurons to pure translation, and so it might account for the abovementioned observation that the compressive nonlinearity decreases the relative influence of translation on MST responses (Figure 4-3B). We therefore extended the model output for each MT subunit to include divisive modulation (Heeger, 1992) by a suppressive field that could vary in terms of its spatial extent, its tuning to motion, and its strength. We defined these quantities as free parameters, and allowed the model to specify which characteristics best fit the data (Figure 4-5B; see Appendix C for details).

The results of these simulations indicate that in most cases the optimal surround was well tuned for motion direction and, surprisingly, that it covered a spatial extent similar to that of each subunit's excitatory receptive field (Figure C-4A). In other words the suppressive influence recovered by the model was typically identical to the excitatory influence, so that stimuli that activated a subunit also limited its output. This type of suppressive mechanism is mathematically indistinguishable from a pure compressive nonlinearity. Indeed the full center-surround model yielded little or no improvement in the quality of the fits relative to the simple nonlinear integration model (Figure C-4B, Table 1). Similar results were obtained if we used spatially asymmetric surrounds (Xiao et al., 1997), symmetric surrounds that

interacted with the centers via subtraction (Raiguel et al., 1995; Tsui et al., 2010) rather than division, and surrounds that had their own output nonlinearities (see Appendix C). Although these models generally performed better than the linear integration model, none consistently outperformed the one-parameter nonlinear integration model. These results are summarized in Table 4-1.

Of course these results do not contradict the important role for MT surrounds in motion processing (Gautama and Van Hulle, 2001; Raiguel et al., 1995), but they do suggest that the contribution of these surrounds to MST optic flow selectivity might be fairly subtle; we return to this issue in the Discussion.

4.2.7 Computational properties of nonlinear motion integration

Intuitively the compressive nonlinearity has a straightforward interpretation: as the input to an individual subunit increases, the output saturates quickly, and as a consequence the MST cell responds best to stimuli that drive many different subunits, even if each subunit is activated weakly. This mechanism thus favors stimuli, such as complex motion, that activate many subunits.

This operation is similar to multiplicative subunit interactions described in other contexts (Gabbiani et al., 2002; Hatsopoulos et al., 1995; Peirce, 2007; Pena and Konishi, 2001). That is, the compressive nonlinearity is similar to a logarithm (see Figure C-3A), and thus the combination of compressive input nonlinearities and expansive output nonlinearity approximates multiplication through the identity $a \cdot b = \exp(\log a + \log b)$. Indeed, we verified in additional simulations that explicit multiplicative interactions between subunits outperformed models of similar complexity in 79% of the MST cells (see Appendix C and Figure C-5).

To quantitatively examine the functional utility of this mechanism we used optimal linear decoding to measure the ability of area MST to represent stimulus information, with and without the nonlinear integration mechanism in place. Specifically, we used our model MST cells to estimate the responses to various stimuli and then trained a simple decoding algorithm to extract various quantities from the population response (Figure 4-6A). This method provides insight into the type of information that would be available to a brain region that had access to the output of the MST population (Ben Hamed et al., 2003; DiCarlo and Cox, 2007).

In our simulations the model MST population responded to a series of discrete objects moving in various directions and speeds, in various positions in the visual field (Figure 4-6B). The goal of the decoder was to recover the different components of each object's velocity, independently of its position in visual

space. Although we have not explored more complex situations involving different visual environments and observer motion, the position-invariant readout of 3D object velocity is necessary for common behavioral situations, such as vergence eye movement control (Takemura et al., 2001; Takemura et al., 2007) and estimation of time-to-contact (Sun and Frost, 1998).

The results of this simulation (Figure 4-6C) show that the model with nonlinear integration of MT inputs (black bars) outperforms the linear hierarchical model (gray bars) in reconstructing velocity in all three dimensions. The difference is especially large (a 60% drop in reconstruction error) in the case of the z-component of the velocity, which is defined by expansion optic flow. As mentioned above, the nonlinear integration approximates a multiplicative operation that renders the model less sensitive to the individual components of expansion stimuli, which are ambiguous with respect to the speed of motion in depth. This suggests that the nonlinear aspects of MST motion encoding are useful for functions which rely heavily on measurement of motion in depth and for which retinal position is relatively unimportant (see Discussion).

4.3 Discussion

4.3.1 Hierarchical encoding of visual stimuli

In this work we have found that neurons in area MST can be effectively characterized by a hierarchical model that takes into account the properties of neurons in MT. An important result from this work is that cells with similar stimulus selectivity, as assessed by relatively low-dimensional tuning curve stimuli, can have subunit structures that differ significantly (Figure 4-4). While we cannot say that the subunits recovered by our model correspond exactly to the anatomical inputs received by each MST neuron, they do represent an optimal estimate under a conservative set of assumptions about MT responses. The formidable challenges associated with a direct characterization of the feedforward inputs to the extrastriate cortex (Movshon and Newsome, 1996) suggest that a model-based approach is particularly valuable.

In addition to a plausible subunit representation, the model requires a nonlinear integration mechanism, which for most neurons is compressive (Figure 4-5). Functionally, the compressive nonlinearity appears to be useful primarily for implementing a multiplicative operation similar to that seen in other visual cortical areas (Brincat and Connor, 2004) and in sensory processing in other species (Gabbiani et al., 2002; Hatsopoulos et al., 1995; Pena and Konishi, 2001). A similar approach has recently been proposed to account for the pattern and speed selectivity of MT neurons (Perrone and Krauzlis, 2008) and for

shape selectivity in V4 (Cadieu et al., 2007). Indeed a similar idea was suggested as a qualitative account of optic flow tuning in MST (Duffy and Wurtz, 1991b). To the extent that the tuning properties found in different brain regions share the same nonlinear integration mechanism, one might expect to find that they share similar temporal dynamics (Brincat and Connor, 2006; Pack and Born, 2001) and contrast dependencies (Pack et al., 2005); these predictions will be tested in future work.

In a complementary analysis, we tested the hypothesis that the compressive effect could be a result of center-surround interactions (Allman et al., 1985; Raiguel et al., 1995; Xiao et al., 1995; Xiao et al., 1997). We tested a wide variety of interaction types (Table 1), with the result that no mechanism consistently outperformed the simple nonlinear model. Moreover, the surrounds recovered by the model were typically the same size as the centers, suggesting that a spatially extended surround is not necessary to account for MST optic flow selectivity. A likely functional rationale for these surrounds is in performing motion segmentation and shape from motion (Gautama and Van Hulle, 2001).

Regardless of its precise functional interpretation, the compressive nonlinear operation could plausibly be implemented through inhibitory interactions among MT neurons with similar receptive field positions and stimulus selectivities; a similar "self-normalization" operation at the level of V1 has been posited to be of primary importance in explaining selectivity in MT cells (Nishimoto and Gallant, 2011; Rust et al., 2006; Tsui et al., 2010). An alternate explanation is synaptic depression at the level of the MT-MST synapse (Abbott et al., 1997). Both mechanisms are equivalent to a compressive static nonlinearity for slowly-varying inputs (Chance et al., 1998). However, self-normalization would have visible effects on the tuning of MT cells, including bandwidth broadening. Given the current knowledge of MT, synaptic depression appears somewhat more plausible, and would reconcile our use of a compressive nonlinearity with previous work showing that expansive output nonlinearities are sufficient for modeling the MT output (Rust et al., 2006). On the other hand, our results are unlikely to arise from contrast normalization or untuned surround suppression at the level of MT (Figure 4-4A).

An alternative explanation for the compressive effect is a form of normalization among MST neurons. A number of different nonlinear tuning operations can be performed through the interplay of feedforward excitation and divisive normalization (Kouh and Poggio, 2008), including multiplicative input interactions. While it is reasonable to assume that normalization shapes MST responses given its important role in areas V1 and MT (Britten and Heuer, 1999; Heeger, 1992; Rust et al., 2006; Tsui et al., 2010), the nature of the normalization pool in MST is unexplored, and as a result it would be difficult to incorporate into our model.

Previous MST models include those that are linear in the velocity domain (Poggio et al., 1991; Zhang et al., 1993) and those that derive their selectivity primarily from the spatial arrangement of MT-like inputs (Grossberg et al., 1999; Lappe, 2000; Orban et al., 1992; Perrone and Stone, 1994), as well as other more informal proposals (Duffy and Wurtz, 1991b; Tanaka et al., 1989; Yu et al., 2010). Each of these models is capable of reproducing certain qualitative aspects of the MST data, but to date there has been no statistical comparison of different model classes. Most recently, Yu et al. (Yu et al., 2010) attempted to estimate MST receptive field substructure by stimulating each cell with a small set of 52 canonical optic flow patterns. These authors concluded that the failure of the resulting receptive field models to account for tuning to complex optic flow stimuli implied that MST stimulus selectivity might result from an unknown mechanism that is sensitive to specific pairwise interactions within MST receptive fields.

While this idea is of course possible, there are two main methodological shortcomings in the Yu et al. (Yu et al., 2010) work. First, the use of a small stimulus set permitted very limited inference power; our results suggest that thousands of different stimuli are necessary to estimate MST receptive field substructure. Second, the model-fitting approach implemented by the authors involved a comparable number of data points and free parameters, and hence would be unlikely to generalize to novel stimuli even with a sufficiently rich training data set. We therefore suggest that the previously reported lack of correspondence between receptive substructure and stimulus selectivity is not due to any intrinsic feature of MST, but rather to the stimulus and modeling methods used in that study.

4.3.2 Decoding of MST population activity

Functionally, MST neurons are likely to be involved in navigation (Britten and van Wezel, 1998; Gu et al., 2010; Perrone and Stone, 1994). Indeed, many previous MST models have assumed that MST receptive fields are arranged to compute heading angle during self-motion (Lappe, 2000; Perrone and Stone, 1994). However, our nonlinear integration model suggests that the properties of MST neurons reflect a more general mechanism that allows MST to participate both in heading and three-dimensional velocity estimation. Indeed, in naturalistic scenes, heading and object velocity often cannot be estimated separately (Zemel and Sejnowski, 1998).

In addition to heading, MST is likely involved in controlling tracking eye movements that maintain fixation on moving objects (Takemura et al., 2007). Such eye movements require accurate estimates of motion direction, and our simulation results (Figure 4-6C) suggest that the estimation of three-dimensional object velocity relies critically on the computational properties we have identified in MST. Specifically, while frontoparallel motion can be recovered with reasonable accuracy by the MT

population, accurate calculation of the velocity of motion in depth requires the nonlinear integration mechanism of the kind used by MST neurons. Consistent with this idea, previous work has shown that MST is important for estimating object velocity (Takemura et al., 2001), and lesions of MST impair vergence movements (Takemura et al., 2007).

Our simulations (Figure 4-6C) show that position-independent estimate of three-dimensional velocity can be readily extracted from the output of the MST population, and that nonlinear integration improves such estimates substantially. Thus, our findings indicate that nonlinear integration allows MST to form a distributed representation of three-dimensional object that supports a wide range of behaviors through a simple decoding mechanism (Ben Hamed et al., 2003; DiCarlo and Cox, 2007).

4.4 Methods

4.4.1 Electrophysiological recordings

Two rhesus macaque monkeys took part in the experiments. Both underwent a sterile surgical procedure to implant a titanium headpost and a plastic recording cylinder. Following recovery the monkeys were seated in a custom primate chair (Crist Instruments) and trained to fixate a small red spot on a computer monitor in return for a liquid reward. Eye position was monitored at 200 Hz with an infrared camera (SR Research) and required to be within 2° of the fixation point in order for the reward to be dispensed. All aspects of the experiments were approved by the Animal Care Committee of the Montreal Neurological Institute and were conducted in compliance with regulations established by the Canadian Council on Animal Care.

We recorded from well-isolated single neurons in the medial superior temporal (MST) area. Single waveforms were sorted on-line and then resorted off-line, using spike-sorting software (Plexon). MST was identified based on anatomical magnetic resonance imaging (MRI) scans and its position relative to MT (just past MT during a posterior approach to the superior temporal sulcus). Most of the neurons from which we recorded had large receptive fields that extended into the ipsilateral visual field, and that responded to expansion and rotation stimuli in addition to translation. This suggests that most of our recordings were from the dorsal, rather than the ventral, portion of MST, but this has not been verified histologically.

4.4.2 Procedure and visual stimuli

Upon encountering a well-isolated MST neuron, we performed a preliminary receptive field mapping with flashed bars and dot fields. For any neuron that was visually responsive, we characterized its responses in terms of *tuning curves* for three optic flow types: translation, expansion/rotation (spirals) and deformation (8 measurements per optic flow type; see Appendix C for equations). Random-dot stimuli were presented in a 24- or 30- degree aperture at 9 different spatial positions on a 3x3 grid with adjacent center positions 12 or 15 degrees apart. The grid was placed over the approximate center of the receptive field as determined by preliminary hand-mapping.

To explore the space of optic flow stimuli more thoroughly, we also developed a novel *continuous optic flow* stimulus consisting of dots moving according to a continuously evolving velocity field generated by random combinations of six optic flow dimensions (see Appendix C for equations). Dots moving according to this velocity field were presented in a circular aperture 24 or 30 degrees wide, which moved slowly around the screen. The stimulus was presented for 6 – 10 minutes.

In all cases, dots were 0.1 degrees in diameter at a contrast of 100% against a dark background. The screen subtended 104x65 degrees of visual angle at a distance of 32 cm. The stimuli were presented at a resolution of 1920x1200, and refreshed at frame rates of 60 or 75 Hz. During continuous stimulus presentation, the animal was rewarded after maintaining fixation for 1 sec.

4.4.3 Models

In order to understand the computations underlying MST optic flow selectivity, we fit the *continuous optic flow* data from each cell to models with various types of subunits. In all cases we first binned the spike trains at 50 ms resolution and excluded time periods during which more than half of the stimulus was off the screen or the animal's gaze deviated more than 1.5 degrees from the fixation point, as well as from 100 ms before loss of fixation to 250 ms following recovery of fixation. This yielded a series of firing rates, which we describe as a response vector $\mathbf{y}$. For the model, we assumed that this response was generated by a Poisson process with rate $\mathbf{r}$, computed deterministically from the stimulus. The log-likelihood of the model $L(\mathbf{y},\mathbf{r})$ is then given up to an additive constant by (Paninski, 2004):

$$L(\mathbf{y},\mathbf{r}) = \log p(\mathbf{y}|\mathbf{r}) = \sum_t y_t \log(r_t) - r_t \tag{5-1}$$

We assumed that the firing rate was given by the rectified output of the receptive field acting on the stimulus, $r_t = g(\eta_t)$. g must be nonnegative for r to be meaningful; additional constraints on the

derivatives of g are required to yield a model that is straightforward to optimize (Paninski, 2004; Wood, 2006). We thus chose $g \equiv \exp$.

The spatiotemporal receptive field acted on the stimulus to yield a response η_t:

$$\eta_t = c + \sum_{\tau} F\left(\rho(t,x,y), \theta(t,x,y)\right) w(t-\tau) \tag{5-2}$$

Here, $F(\rho,\theta)$ is a nonlinear spatial filter that acts on the optic flow stimulus, which is described by the local motion speed $\rho(t,x,y)$ and direction $\theta(t,x,y)$. c is a constant offset. We sampled the stimulus at a spatial resolution of 24 by 24 samples, generally covering from 48 to 60 degrees of visual angle. The temporal filter $w(\tau)$ was assumed to last 5 time steps, spanning from -50 ms to -250 ms. This formulation embodies an assumption of separable, linear temporal processing, which is supported by earlier studies of the temporal behavior of MST neurons (Khawaja et al., 2007; Paolini et al., 2000).

The nonlinear spatial filter $F(\rho,\theta)$ was assumed to be given by the sum of M nonlinear subunits $f(\rho,\theta,\mathbf{p}^m)$, where $\mathbf{p}^m$ denotes the parameters of the mth subunit:

$$F(\rho,\theta) = \sum_{m=1}^{M} f(\rho,\theta,\mathbf{p}^m) \tag{5-3}$$

We examined the compatibility of the data with several different models, each of which was defined by the structure of its subunits.

Hierarchical model. This model embodies the assumption that MST responses are approximately linear in terms of their feedforward input from area MT, which provides one of the strongest projections to MST (Boussaoud et al., 1990). The tuning of the modeled subunits is determined by 3 components. Subunits were assumed to have log-Gaussian speed tuning with preferred speed p_ρ:

$$R\left(\rho(x,y), p_\rho\right) = \exp\left(-\left(\log\left(\rho(x,y)+1\right) - p_\rho\right)^2 / 2\sigma_\rho^2\right) - \exp\left(-\left(\log(\rho(x,y)+1) + p_\rho\right)^2 / 2\sigma_\rho^2\right) \tag{5-4}$$

Note that a second log-Gaussian is subtracted from the first to constrain the response to be zero when there is no motion. Although MT cells tuned to low speeds have robust responses to static stimuli (Palanca and DeAngelis, 2003), we did not model such responses, as our stimulus poorly sampled slow

speeds. We set the speed tuning width to $\sigma_\rho = 1$, similar to the mode of the distribution of speed tuning widths reported in MT (Nover et al., 2005).

The direction tuning of the subunits was given by a Von Mises function with preferred direction p_θ:

$$D\left(\theta(x,y), p_\theta\right) = \exp\left(\sigma_\theta \cos\left(\theta(x,y) - p_\theta\right)\right) - 1 \tag{5-5}$$

The value 1 is subtracted from the result so that by convention a stimulus moving in a direction orthogonal to the preferred direction elicits no response, and a stimulus moving in the non-preferred direction elicits a negative response; a similar convention was used in previous models of MST (Perrone and Stone, 1994; Perrone and Stone, 1998). The bandwidth parameter was chosen to be $\sigma_\theta = 2.5$, corresponding to a full-width at half maximum bandwidth of 86 degrees, similar to the mean value of 83 degrees measured with moving random dots reported in (Albright, 1984). Finally, subunits had a Gaussian spatial profile:

$$G\left(x, y, p_x, p_y, p_\sigma\right) = \exp\left(-\left(\left(x - p_x\right)^2 + \left(y - p_y\right)^2\right) / 2p_\sigma^2\right) \tag{5-6}$$

The direction, speed and spatial response of the subunits were combined to form the response of the subunit:

$$f\left(\rho, \theta, \mathbf{p}\right) = \quad p_g h\left(\sum_{xy} R\left(\rho(x,y), p_\rho\right) \cdot D\left(\theta(x,y), p_\theta\right) \cdot G\left(x, y, p_x, p_y, p_\sigma\right)\right) \tag{5-7}$$

Here p_g denotes the gain of the subunit, and the function $h(x) = \max(x, 0)$ returns the positive part of the response (half-wave rectification).

Hierarchical model with nonlinear integration. This model provides each subunit with a nonlinearity that exhibits either compressive or expansive behavior depending on a free parameter (expansive when β > 1, compressive when β < 1). Subunits take the same form as equation (4-7), but with the nonlinearity replaced by $h(x) = \max\left(x, 0\right)^\beta$. This model reduces to the previous model when $\beta = 1$. Importantly, β is shared across all subunits for a given model fit. In practice, we fit the model for 7 different values of β ranging from 0.2 to 1.4 and selected the optimal β for a cell based on the cross-validated likelihood.

4.4.4 Model fitting

Estimating the models described above is challenging, as they contain many free parameters and must be fit with rather noisy data. To constrain the parameters and to obtain fits which extrapolate well to novel data, the fitting procedure must limit the dimensionality of the model. This is typically done by including explicit assumptions about the parameters (Wu et al., 2006). A particularly powerful assumption is that a model is sparse, meaning that most of its parameters are zero (Friedman et al., 2000). In a neurophysiological context, this corresponds to the assumption that only a modest number of subunits are driving a given cell, which is consistent with anatomical and correlation studies of early sensory areas (Anderson et al., 1998; Usrey et al., 2001). Models fit with assumptions of sparseness have proven increasingly useful in estimating the receptive field properties of high-level neurons (David and Gallant, 2005; Paninski, 2004; Willmore et al., 2010). We thus used gradient boosting, a stepwise fitting procedure that introduces an assumption of sparseness (Friedman et al., 2000). The number of free parameters was limited through 5 fold cross-validation (Wu et al., 2006).

4.4.5 Validation and accuracy metrics

For those cells for which the *continuous optic flow* stimulus spanned the spatial range of the *tuning curve* stimulus (36/61), we predicted the responses to the *tuning curve* stimuli based on the *continuous optic flow* fit. Note that the continuous optic flow stimulus samples a large, six-dimensional space of optic flow, of which the tuning curve stimuli comprised a small number of points. Thus this approach is a rigorous test of the model's ability to extrapolate to novel stimuli.

For these simulations we ignored the temporal component of the responses, instead predicting the total spike count in response to a stimulus. We allowed the gain and baseline firing rate to be estimated from the data using standard techniques (Wood, 2006) rather than predicted from the continuous optic flow stimulus. Given a predicted response $\mathbf{r}$ and an observed response $\mathbf{y}$, the quality of the prediction may be assessed using the standard R^2 metric of variance accounted for):

$$R^2 = \frac{\mathrm{Var}(\mathbf{y}) - \mathrm{Var}(\mathbf{y} - \mathbf{r})}{\mathrm{Var}(\mathbf{y})} \tag{5-8}$$

In practice the value $R^2 = 1$ cannot be attained, as $\mathrm{Var}(\mathbf{y} - \mathbf{r})$ for a perfect prediction is the variance of the noise, which is non-negligible in physiological measurements. To recover a natural scale we thus used a corrected R^2 metric, also known as predictive power (Sahani and Linden, 2003):

$$\bar{R}^2 = \frac{\mathrm{Var}(\mathbf{y}) - \mathrm{Var}(\mathbf{y} - \mathbf{r})}{\mathrm{Var}(\hat{\mathbf{y}})} \tag{5-9}$$

Here $\mathrm{Var}(\hat{\mathbf{y}})$ is the variance of the unobserved noiseless signal $\hat{\mathbf{y}}$. The explainable signal variance $\mathrm{Var}(\hat{\mathbf{y}})$ is estimated from the pattern of disagreement between responses in different presentations of the same stimulus (equation 1 in Sahani and Linden, 2003).

To determine whether the relative level of responses to different classes of optic flow (translation, spirals, deformation) was correctly accounted for by the different models, we also computed a *stimulus class* $\bar{R}^2$ which introduced a free gain per optic flow type. In the case of the hierarchical model, we found that the relative level of responses across stimulus types was misestimated for 70% of cells (25/36, $p<0.001$, likelihood ratio test), and in a majority of these cases (84%, 21/25) predicted responses were too weak for spiral stimuli relative to translation stimuli. We emphasize that the stimulus class metric is not an accurate reflection of the quality of the model predictions, but rather is an artifice that allowed us to isolate one mechanism underlying quality of fit.

4.4.6 Decoding simulations

We compared the capacity of an optimal linear estimator to extract information relevant to behavior. From the 61 fits (one per cell) under the hierarchical models, we generated 61x4 = 244 virtual cells through reflections across the x and y axes to compensate for inhomogeneous sampling of visual space. Since the cells were tested at different resolutions and at different screen positions, we scaled and repositioned the receptive fields to span the central 120x120 degrees of the visual field. Stimuli were cropped to the central 90x90 degrees of the visual field to avoid artifacts around receptive field edges.

An object 1/16th the size of the visual field was simulated as undergoing three-dimensional motion in one of 17 directions (left, up, down, right, towards the observer, and intermediate directions; see Figure 4-6B). The object could be located in one of 25 positions inside the receptive field lying. The speed of the object was chosen on a log scale from 2 to 16 Hz; the physical speed of the object may be reconstructed in m/s or deg/s if the distance to the object is known.

We reconstructed the physical parameters of the stimulus using an optimal linear estimator given the outputs of a population of MST cells (Ben Hamed et al., 2003). We picked 122 cells at random out of the pool of 244 to yield a decoding population of a size comparable to that previously used in the literature (Ben Hamed et al., 2003). The variables to reconstruct were the signed log velocities in each direction,

for example $\mathrm{sign}(v_x)\log(|v_x|+1)$ for the velocity in the x direction. To do so we computed the weights $\mathbf{w}$ that minimized the squared error between the reconstruction $\mathbf{Xw}$ and the variable to decode $\mathbf{y}$. Here $\mathbf{X}$ is a matrix with one row for each stimulus and 123 columns (one for each cell and an offset). The quality of the reconstruction was determined by the root mean square (RMS) error, and was expressed as a percentage of the range of log velocity in the x direction (5.67 log Hz). Each decoding simulation was repeated for 50 different random choices of decoding population to yield a mean value and standard deviation.

4.5 Figures

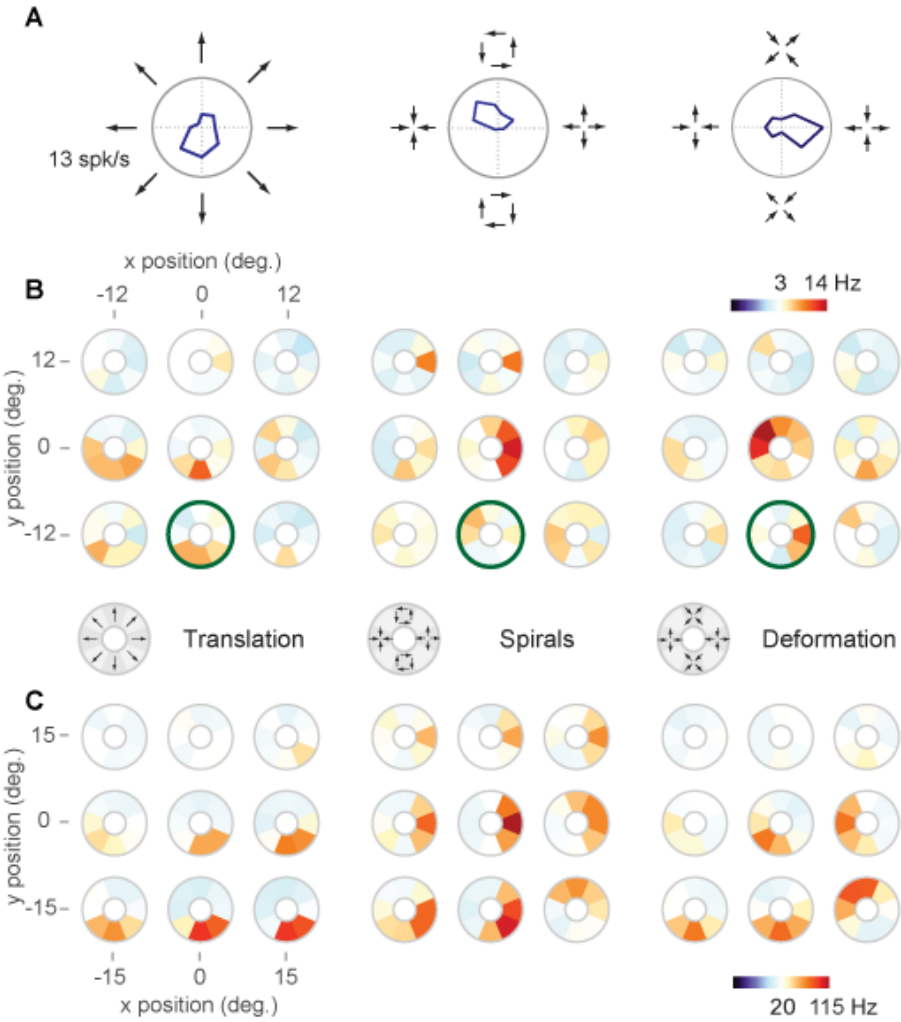


Figure 4-1. Tuning of MST neurons for complex optic flow

(A) Tuning curves for a single MST neuron to visual motion comprised of translation (left), spirals (center), and deformation (right). Stimuli were presented at one position on a 3x3 grid centered on the fovea. (B) Tuning mosaics, in which large responses are represented by red colors, small responses by blue and median responses by white. Each mosaic captures the tuning for one of the stimulus types shown in (A) at nine positions in the visual field. The mosaics highlighted in green correspond to the tuning curves shown in (A). This cell consistently preferred downwards translation (left); and tuning for spirals (center) and deformation (right) varied across positions. (C) Tuning mosaics for a second example cell. This cell consistently preferred downwards translation (left) and expansion (center) at most spatial positions.

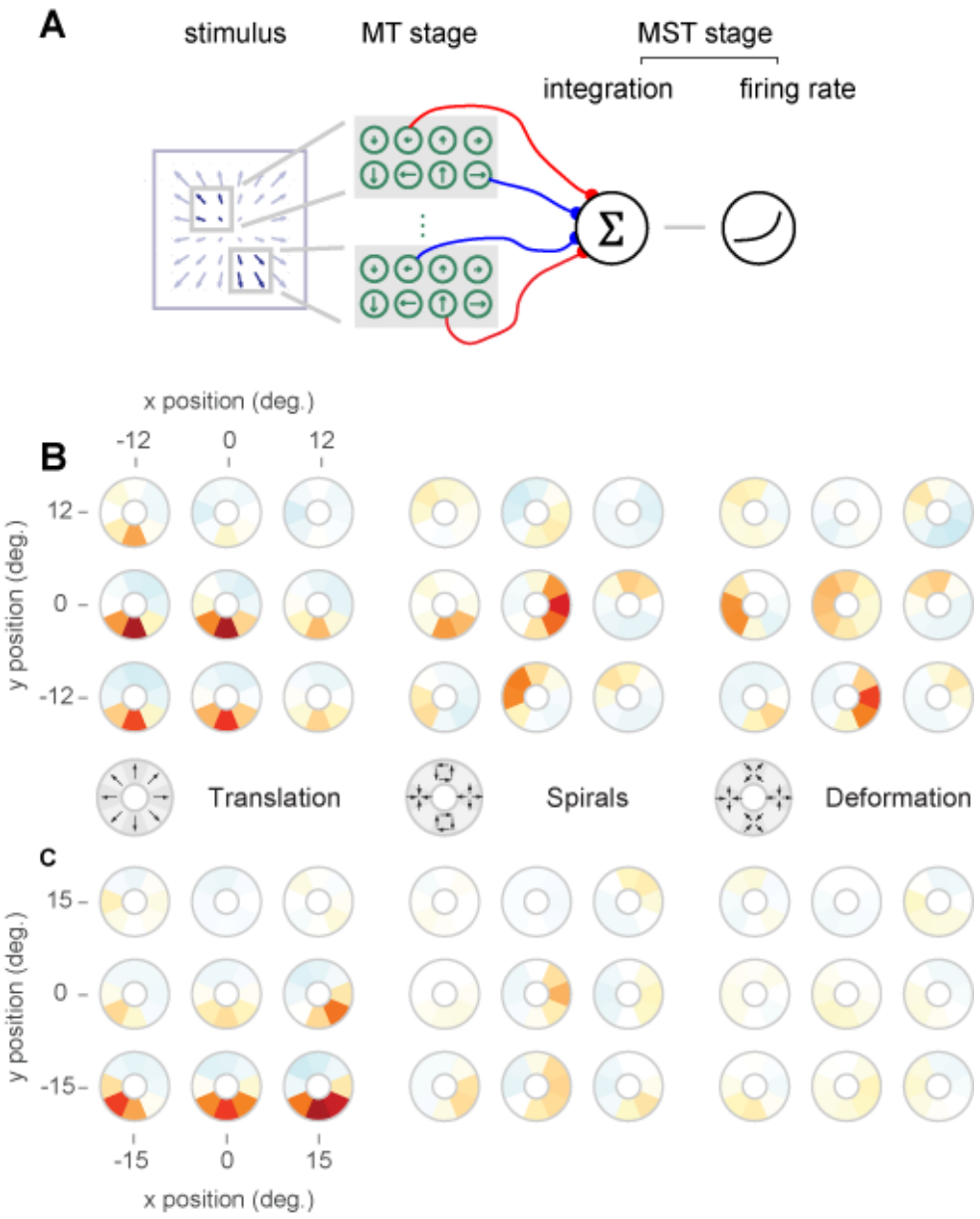


Figure 4-2. Performance of the linear hierarchical model.

(A) In the linear hierarchical model, the stimulus was processed by groups of MT-like filters (only 2 groups shown for clarity), which could vary in preferred direction, spatial position, and speed. The outputs of these filters were weighted, summed, and nonlinearly transduced to a firing rate. (B) Predicted tuning mosaics for the same cell as in Figure 1B under the hierarchical model. The hierarchical model correctly captures the optic flow tuning of this cell, including the preferences for spiral motion (center panel). (C) As in (B) but for the example cell shown in Figure 1C. The hierarchical model fails to capture this cell's tuning to complex optic flow (spirals and deformations).

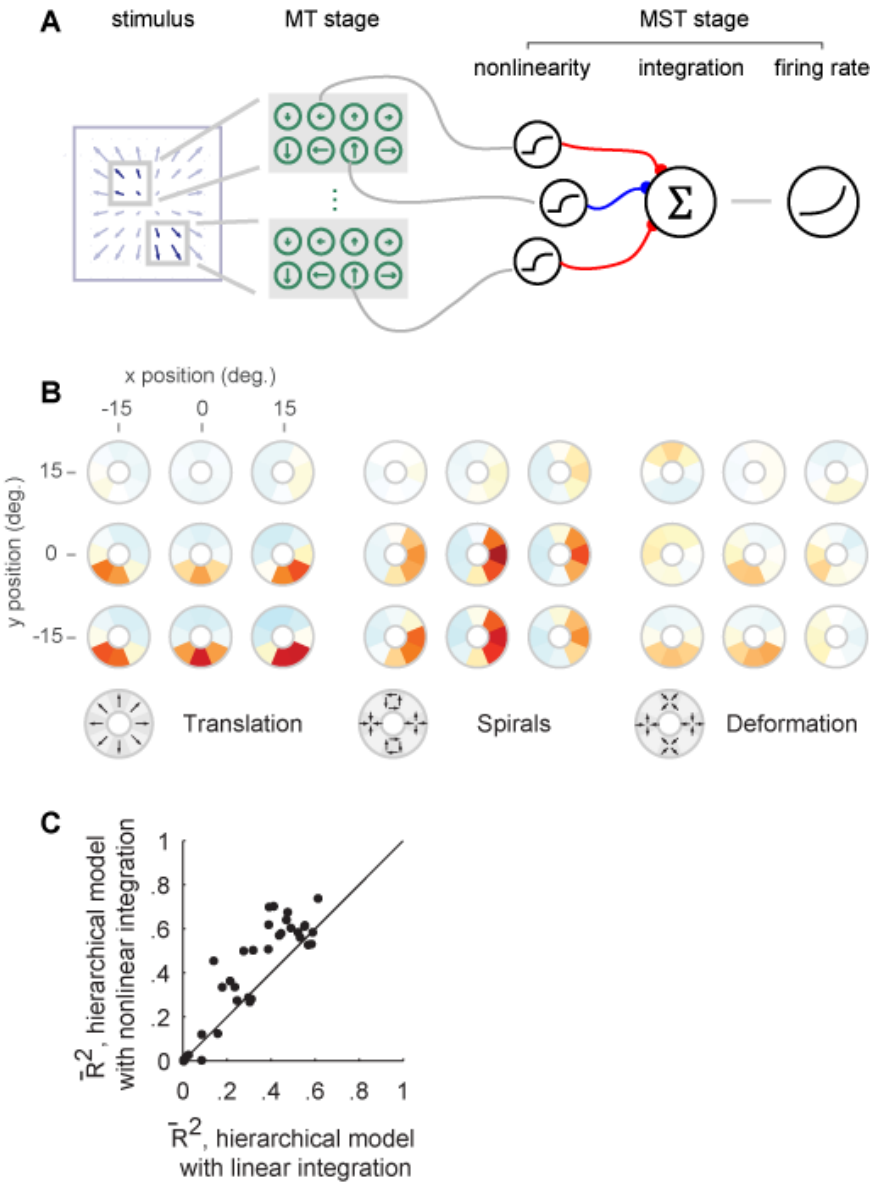


Figure 4-3. Performance of the hierarchical model with nonlinear integration.

(A) In the hierarchical model with nonlinear integration, the stimulus was processed by groups of MT-like filters. The output of these filters was passed through a nonlinearity, then weighted, summed, and transduced to a firing rate. For each MST cell, the nonlinearity could vary from compressive to expansive and was identical across all subunits. (B) Predicted tuning mosaics for the same cell as in Figure 1C under the nonlinear integration model. This model accurately captures the tuning and relative response levels of this cell to translation and spirals. (C) Quality of tuning curve predictions for the hierarchical model with and without nonlinear integration. The nonlinear integration model improves performance in 75% of the tested cells.

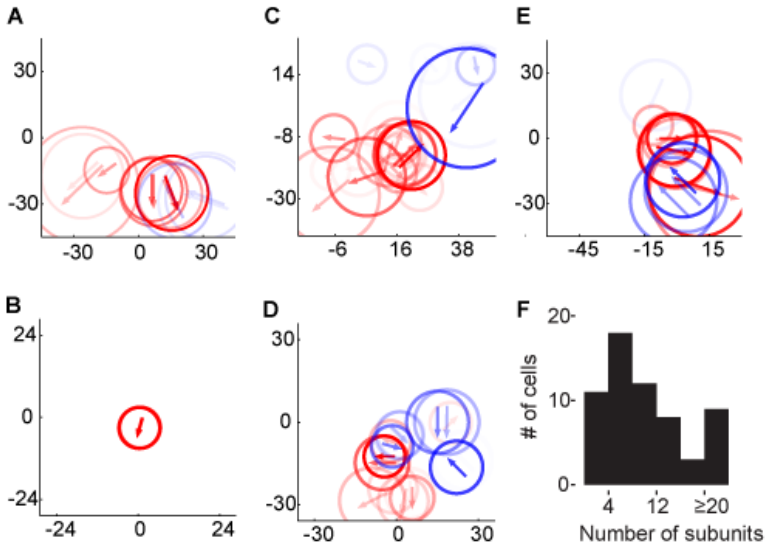


Figure 4-4. Diversity of receptive field substructures in MST.

(A) Receptive field substructure for the example cell shown in Figure 1C. This visualization was produced by constructing a compact representation of the subunits in the nonlinear integration model. Red represents excitatory input, blue inhibitory, opacity the magnitude of the weight of the subunit, and the direction of the arrow the preferred direction of the subunit. This cell's tuning for downwards motion and expansion is explained by downwards-left tuned subunits in the lower left portion of the visual field and downwards-right tuned subunits in the lower right. (B) Substructure of example cell shown in Figure 1B. This cell's receptive field was composed of a single, downwards-left tuned subunit. (C), (D) and (E) Most critical subunits for three expansion-tuned cells. While these cells and the one presented in Figure 5A can all be described as expansion-tuned, they show a diversity of receptive field arrangements. (F) Histogram of number of subunits found by the visualization procedure.

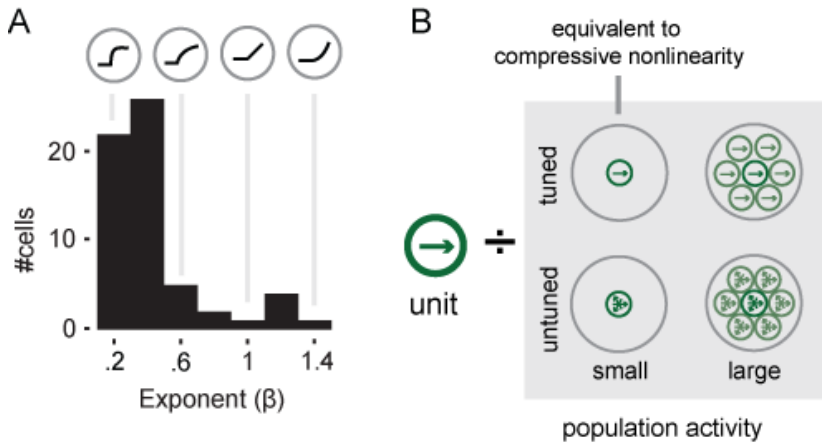


Figure 4-5. Analysis of optimal subunit nonlinearity across MST population.

(A) In the nonlinear integration model, subunit outputs were processed by a nonlinearity of the form $f(x) = \max(0, x)^{\beta}$ β values smaller than 1 correspond to a compressive nonlinearity, while values greater than 1 indicate an expansive nonlinearity. Most MST cells required a compressive nonlinearity at the level of each subunit. (B) In the divisive surround model, the output of the center of subunits is divided by the output of a pool of subunits differing in tuning bandwidth and spatial extent. A strongly tuned divisive surround with small spatial extent is equivalent to a static compressive nonlinearity.

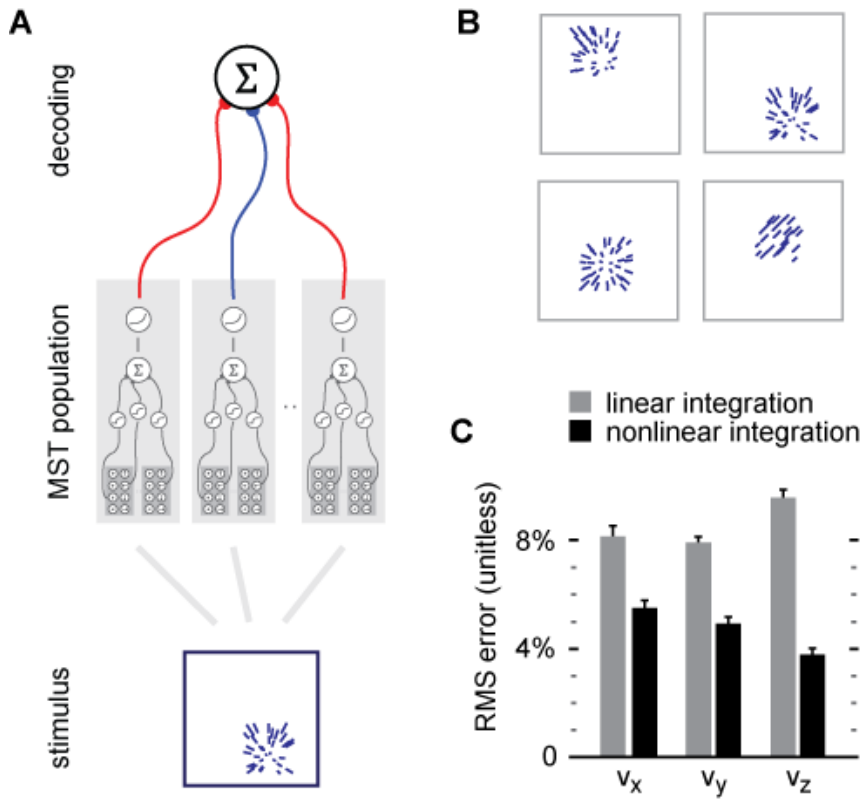


Figure 4-6. Role of nonlinear integration revealed by population decoding.

(A) In a decoding simulation, stimuli were processed by a population of MST model cells estimated from the recorded data. The goal of the linear decoder (top) was to deduce physical parameters of the stimulus based on the output of the MST population (B) Example stimuli used in the object decoding simulation corresponding to motion of an object in three dimensions. (C) Performance of the decoder based on input from the hierarchical model population with (black bars) and without (gray bars) nonlinear integration. Results are quantified as the mean error relative to the range tested; smaller values indicate better performance. Error bars indicate one standard deviation from the mean, determined through a resampling procedure (see Methods). The sensitivity of the nonlinear integration mechanisms to combinations of inputs facilitates the decoding of object velocity based on the output of the MST population.

4.6 Tables

Model	Median LL/s, continuous stimulus (median %difference relative to nonlinear MT model)	Median R2, tuning curve stimuli (median % difference relative to nonlinear MT model)
Linear	0.38 (−54)	0.19 (−48)
MT	1.11 (−9)	0.31 (−21)
Nonlinear MT	1.23	0.50
Nonlinear MT (unrestricted)	1.23 (2)	0.45 (−6)
Divisive surround	1.20 (4)	0.48 (−1)
Asymmetric surround	1.12 (−2)	0.34 (−15)
Nonlinear asymmetric surround	1.22 (5)	0.45 (0)
Subtractive surround	1.09 (−3)	0.36 (−15)
Nonlinear subtractive surround	1.22 (4)	0.47 (−1)

Table 4-1- Summary of quality of fits of all models considered.

Goodness-of-fit for continuous stimulus is defined as cross-validated log-likelihood accounted for per second of data. Quoted percentage values are the median ratio of goodness-of-fit for target model divided by goodness-of-fit for nonlinear MT model. Note that the ratio of medians is not necessarily equal to the median of individual ratios.

Chapter 5 summarizes the results from the previous chapters, discusses limitations of the presented results and outlines possible remedies for these issues. I then look at future directions for this work, and outline a research programme which could further help reveal the nature of hierarchical visual representation.

5. Discussion and conclusion

5.1 Summary of results

The aim of this thesis was to gain an understanding of the mechanisms by which the visual hierarchy transforms visual information into representations which can drive behaviour. To do so, I developed and applied parametric modeling methods to study the visual system at multiple scales – at the psychophysical level in Chapter 2, at the neuronal ensemble level in Chapter 3, and at the scale of single neurons in Chapter 4.

In Chapter 2 (Mineault et al. 2009), I introduced the general parametric modelling framework which was used as a building block for subsequent studies. Specifically, I showed how observer's decisions in the context of psychophysical systems identification – the classification image paradigm – could be captured with generalized linear models (GLMs).

GLMs furnish an appropriate framework with which to model many signals of interest in the context of sensory neuroscience, as I've argued in Chapter 2 and in the introduction: the linear component of GLMs naturally maps onto the concept of the linear receptive field of neurons and neural ensembles and onto the linear observer model of psychophysics. The nonlinear transduction and non-Gaussian noise processes of GLMs can account for non-continuous data generated in the context of sensory neuroscience, i.e. psychophysical decisions and spike trains. Using assumptions on model parameters, justified by biological considerations, it becomes possible to constrain the parameters of such a model with noisy neuronal or psychophysical data. Finally, with appropriate parameterization of the stimulus, it is possible to account for nonlinear processes, which is necessary when studying systems far removed from the retinal input.

I showed that the assumption that the internal decision mechanism of a psychophysical observer – the classification image – was sparse in a multiscale basis was particularly useful to estimate classification images. In simulations and in real data, I showed that this assumption allowed for the estimation of significant classification images in a small fraction of the number of trials required by previously proposed methods. This notable increase in efficiency made it possible to explore more complex models of the observer's decision process. In an example application, that in the context of a detection task, the resulting classification images showed clear signatures that the observer used a spatially invariant detection mechanism; likely driven by a population of invariant neurons, i.e. simple and complex cells.

These results show that parametric modeling in the context of psychophysics can reveal properties of the underlying neural representation.

In Chapter 3 (Mineault et al. 2013), I applied a similar analysis framework in the study of the visual representation of space at the level of neural ensembles in area V4. Using a multi-electrode array, I showed that it is possible to estimate highly significant local field potential (LFP) receptive fields on every electrode. Surprisingly, the resulting receptive fields were highly non-separable with respect to time-lag, which was not the case for multi-unit activity (MUA) receptive fields. As a consequence, the retinotopy of LFPs matched that of MUAs for a restricted set of time lags.

I explained these findings using a low-rank parameterization of the LFP receptive fields, in which the RFs are generated by a sum of two components: an electrode-specific component and a global, array-wide shared component. This model captured the apparent change in retinotopy as a function of time lag. The electrode-specific component's retinotopy matched that of MUAs, indicating that it is of local origin, while the shared component, being the same across a given multi-electrode array, is of a global origin.

These results reconcile a controversy in the literature concerning the integration radius of the LFP: the LFP reflects both local activity and global biases in visual representation, and signal processing can selectively enhance each component, causing apparently divergent estimates. These results show the local field potential can be a useful signal to study visual representations at the scale of neuronal ensembles, provided that the signal is carefully processed to isolate the component of interest.

Finally, in Chapter 4 (Mineault et al. 2012), I applied parametric modeling techniques to the study of single neuron computation in area MST of the primate dorsal visual stream. The selectivity of MST neurons to complex optic flow had been a source of much speculation and countless modeling studies over the past 25 years (Tanaka et al. 1986), but thus far, it had proven impossible to crack open the receptive fields of MST neurons.

By fitting MST neuronal responses on the assumption that they integrate from the output of simulated MT neurons with realistic properties, I showed that it was possible to account for some, but not all, of the optic flow selectivity properties of MST neurons. Rather, MST neurons appear to integrate MT neuronal output using a nonlinear integration mechanism which shapes the selectivity of these neurons to non-preferred stimuli. By considering several different models for the nonlinear integration, I showed that a biologically plausible mechanism for this effect is that a compressive nonlinearity at the level of

the input, possibly implemented by synaptic depression, interacts with the static thresholding nonlinearity of MST neurons to create AND-like selectivity.

The resulting estimated receptive fields at last revealed the internal selectivity of MST neurons, and confirmed, as had been long suspected, that MST receptive fields are tuned for complex combinations of motion which are frequently found in wide-field optic flow. Finally, in decoding simulations, I showed that the nonlinear integration mechanism was particularly useful for communicating, in a position-invariant fashion, the velocity of an approaching object. These results indicate that the uncovered selectivity of MST neurons could have a functional role in the estimation of object motion, in particular to drive vergence.

Put together, these studies show that parametric modeling methods, applied at multiple scales in the visual system, can reveal the mechanisms by which visual information is represented into a format that is useful for behaviour.

5.2 Limitations and extensions

5.2.1 Parametric modelling

5.2.1.1 Addressing lack of fit

It is often difficult, beyond the sensory periphery, to estimate models which are predictive of a system's output given complex, naturalistic input. Even in primary visual cortex, state-of-the-art models can only account for roughly 25% of stimulus-driven variance in spike rate in a standard benchmark (J. Gallant, F. Theunissen, http://neuralprediction.berkeley.edu).

I have presented a parametric modelling framework which addresses lack-of-fit issues inherent in linear models in three ways:

- Modeling a more realistic transduction and noise process within the context of Generalized Linear Models
- Considering nonlinear representation of the input in which high-level systems can be described linearly, in particular in Chapter 4
- Softly constraining model parameters to be smooth, sparse and low-rank.

As I have demonstrated in the previous chapters, this parametric modelling methodology is potentially more predictive of visual responses than classical systems identification. Nevertheless, there is

significant room for improvement, and I now criticize and address each of the components of the parametric modeling framework.

5.2.1.2 Transduction and noise process

Real systems rarely behave like idealized statistical models. This extends to the Generalized Linear Models used throughout this thesis to analyse different visual representations.

I modelled psychophysical decisions as a binary process. This, however, makes inefficient use of the psychophysical data. In particular, reaction times are frequently instructive about the difficulty of a psychophysical task, and consequently about an observer's decision process (Ratcliff and Rouder, 1998). Decision time conditional distributions can be modelled, in the GLM framework, using any number of positive, continuous distributions (MacCullagh and Nelder, 1989).

A particularly useful distribution here is the inverse Gaussian distribution, which models the distribution of waiting times for a system to reach threshold when driven by temporally-integrated Gaussian noise. This corresponds exactly to waiting times in the race model of decision-making (Ratcliff and Rouder, 1998). The inverse-Gaussian is an exponential family distribution, and consequently inverse Gaussian decision processes can be modeled within the GLM framework (MacCullagh and Nelder, 1989).

To the best of my knowledge, however, such a GLM has yet to be used to estimate classification images with reaction times. It will be interesting to see if simultaneous modelling of reaction times and decisions can reveal different aspects of an observer's strategy.

In the analysis of local field potentials, I assumed that the noise in local field potentials is temporally independent and normally distributed. Local field potentials, of course, have sluggish dynamics, and a more appropriate noise model may be that the observation noise is temporally correlated and normal. Such a formulation requires the specification of a model of the temporal correlations.

It is possible to estimate the correlation structure of the noise as well as the deterministic relationship between local field potential and stimulus in an Empirical Bayes framework, in a manner similar to the generative model of wideband signals presented in Appendix B. The resulting estimates of the LFP receptive fields should be less noisy; it will be interesting to see if this additional modelling complexity results in qualitatively significant or merely incremental improvements in receptive field estimates.

Finally, in the analysis of spike trains, I assumed that spikes are generated according to an inhomogeneous Poisson process whose rate is determined by the projection of the stimulus onto the

neuron's internal template. Such a linear-nonlinear-Poisson (LNP) model has a rich history in sensory neuroscience (Simoncelli et al. 2004). It can be considered a first-order improvement on linear models, providing a more realistic transduction and noise process; in particular, the LNP model necessarily predicts positive spike rates, and the variance in the observed number of spikes increases in proportion with the mean of the firing rate (Dayan and Abbott, 2001).

However, real neurons diverge from Poisson statistics in a number of ways. In the sensory periphery, spike times are highly precise (Butts et al., 2007), and spike count variability across trials is much smaller than expected from a Poisson process. A simple way to account for temporally precise responses is to assume that, following a spike, a neuron's rate changes according to a self-coupling filter. Such a model is straightforward to estimate within the GLM framework, and has been applied successfully to the study of neurons with high temporal precision in the retina and the LGN (Butts et al., 2011; Pillow et al., 2008).

Higher-level neurons, in contrast, are frequently more variable across trials than the Poisson process suggests (Dayan and Abbott 2001). Apparent overdispersion could be the result of falsely assuming that the underlying spike rate of the neuron is the same across trials. While it is possible to roughly equate the visual stimulus across trials in a systems identification experiment, a high-level neuron responds both to the visual stimulus and to internal variables which vary uncontrollably across trials. These state variables include the strength and location of the attentional spotlight, energy reserves, and local network activity (Fries et al. 2001).

Thus, apparent overdispersion can result from failing to account for internal states. This can be resolved, in a straightforward extension of the GLM framework, by assuming that the neuron responds to unobserved correlated noise sources (Paninski et al., 2010). An alternative is to use a surrogate signal – e.g. the local potential – to measure internal state, and to model the neuron as responding to both the visual stimulus and the measured internal state. Again, this may result in improved receptive field estimates, although whether the improvements are qualitatively significant remains to be shown.

5.2.1.3 Nonlinear representations

To analyze single neuron receptive fields in area MST in Chapter 4, I represented the visual stimulus in a basis that mimics MST's afferent input in area MT. This nonlinear stimulus representation is biologically motivated by the hierarchical nature of visual processing, and furnished key insights into the receptive

fields of MST neurons. This approach hinged on previous quantitative characterizations of neurons in area MT, one of the best studied areas of the visual cortex (Born and Bradley, 2005).

To extend this approach directly to other areas of cortex will require parameterizing the selectivity of the afferent input to a given set of neurons in a low-dimensional fashion, which is problematic when the afferent input is poorly characterized. For example, to estimate the selectivity of neurons in V4 with this approach would require good knowledge of visual representation in V2, an area that is currently poorly understood (Freeman et al., 2013; Willmore et al., 2010).

In this scenario, it will be necessary to characterize different visual areas in a sequential fashion to obtain effective models of hierarchical processing. I return to this point in the Future Directions section.

An alternative to inferring the full hierarchy is to find an effective representation of the visual stimulus in which the response of a given neuron is approximately linearized. This is the idea behind nonlinear cascade models, which describe the responses of sensory neurons by representing the stimulus using a flexible nonlinear transformation, followed by a linear-nonlinear-Poisson (LNP) process. For different choices of flexible nonlinear transformation, different model classes are obtained, including the nonlinear input model (NIM; McFarland et al., 2013), generalized quadratic model (Park et al., 2013), convolutional GLM (Vintch et al., 2012), and maximally informative dimension model (Sharpee, 2013).

While such models, being based on flexible, rather than fixed representations of the visual stimulus, are potentially more flexible than the hierarchical scheme that I applied in Chapter 4, this flexibility comes at a cost. Indeed, the number of parameters that must be estimated increases sharply compared to fixed representation models, which limits their applicability to cases where the relevant stimulus subspace is low-dimensional. I present an alternative framework which combines the advantages of both fixed and flexible representation schemes in the Future Directions section.

5.2.1.4 Constraints on model parameters – the bias-variance tradeoff

At the heart of the proposed systems identification methodology is the idea that assumptions on model parameters – typically embodied in Bayesian priors - can be judiciously used to strongly constrain model parameters in finite experimental time. An alternative position is to *let the data speak for itself* – to forgo assumptions on model parameters in the hopes of finding novel, unsuspected relationships in the data (Gelman et al., 2003).

Allowing sufficient flexibility for the discovery of novel relationships is an important design goal in systems identification and statistical analysis. However, using few constraints on model parameters limits one's ability to consider higher-dimensional – and potentially more powerful and insightful – model forms.

The psychophysical detection task shown in Chapter 2 is illustrative of this trade-off: with weak assumptions on model parameters, inferring that the observer is performing the task non-ideally takes more than 5000 trials (Figure 2-7). On the other hand, with a stronger prior – sparseness in a basis – reaching the same inference requires only 1,200 trials.

This allowed me to consider more flexible model forms – in this case, allowing that the observer uses different internal decision rules when the signal was present or not, which doubled the number of parameters in the model. This proved insightful into the observer's decision process – the analysis showed that the observer is using a representation with spatial uncertainty to perform the task.

Maximum a posteriori (MAP) estimators are biased (Gelman et al., 2003) – that is, for many realizations of the data, the mean estimate of the model parameters will not be equal to the true parameters. They trade off this bias with a decrease in variance – that is, for many realizations of the data, the estimates of the model parameters will be more similar to each other. This results in better constrained model parameters – i.e. less noise and smaller error bars.

This is an example of a bias-variance trade-off (Mackay, 2002). A key element to perform such a trade-off judiciously is to estimate the relative strength of the likelihood and the prior as part of model estimation; e.g. using cross-validation or Empirical Bayes estimation, as outlined in the introduction. In the limit where there is sufficient data to strongly constrain model parameters, the prior's relative weight vanishes, and the estimated model weights converge to the optimal estimates regardless of prior form, provided that the prior is non-zero everywhere (Mackay, 2002).

Another key element to use the bias-variance trade-off to our advantage is to compute baselines – e.g. ideal observer analysis in psychophysics, linear models in neuronal systems identification – and perform sensitivity analyses – i.e. ensuring that the results are qualitatively similar with different model forms - to determine the sensitivity of the results to model assumptions (Gelman et al., 2003). Within this framework – *careful* parametric modeling – I've shown that it is possible to estimate complex, biologically-relevant models within limited experimental time (Mineault et al. 2008, 2012, 2013).

5.2.2 Interpretation of parametric modelling results

I have used parametric modelling to study visual representations at multiple scales. As with any form of systems identification, the interpretation of these results is complicated by the gap between the necessarily simplified model forms and the actual, biophysical mechanisms which underlie the system's response.

In the context of psychophysical systems identification, it is clear that the estimated decision mechanisms are not detailed descriptions of the biological processes which underlie a decision. Considering that behaviour is a whole-brain process, it may appear futile to try to relate the effective decision process reflected in a classification image and the actual biophysical processes that underlie the decision. However, simulations of idealized neurons hypothesized to underlie a decision process are frequently instructive in interpreting classification images (Neri and Heeger, 2002). Furthermore, under appropriate circumstances, classification images can be meaningfully compared and contrasted with neuronal receptive fields (Neri and Levi, 2004). Thus, psychophysical systems identification can reveal relevant aspects of a neural representation, provided that appropriate care is taken to relate it to the underlying biology.

With signals derived from neural populations, i.e. local field potentials and fMRI, the primary difficulty is in relating the observed signal and the underlying neuronal activity. LFPs, in particular, are a reflection of the extracellularly-measured voltage, and although correlated synaptic activity is its primary source, Buzsáki et al. (2012) denote more than a dozen distinct types of electrical events which are reflected in the LFP. I have, through a signal processing algorithm (Appendix B), isolated and removed one of these components – action potential traces from foreground neurons – which is frequently a nuisance in interpreting LFPs.

Nevertheless, there remains much work to be done in understanding the many sources of local field potentials. Through careful parametric modelling, I have identified two such components in a systems identification paradigm in V4 (Mineault et al. 2013). One of these components, a retinotopic component, had properties consistent with a local source, and is likely a reflection of local correlated synaptic activity. The properties of the second, shared component appeared to reflect global biases in the visual representation.

For many analysis purposes, the first component, as a reflection of the local synaptic activity, would be of interest. Thus, through careful parametric modeling, it is possible to isolate and estimate the

properties of the sub-components of the local field potential which are of interest in studying neural representations. Furthermore, detailed biophysical simulations can help resolve interpretation problems (Einevoll et al., 2013; Lindén et al., 2011; Logothetis et al., 2001) when studying highly derived signals like LFPs and the BOLD signal.

In neuronal systems identification, we have a clear biophysical understanding of the mechanisms by which the signal – the spike train – is generated (Koch, 1999). However, the models used to estimate neuronal receptive fields are generally not direct reflections of the underlying biophysics of spike train generation. Detailed biophysical models – spatially extended Hodgkin-Huxley neurons with multiple compartments, for instance – are, with current methods, impossible to estimate in systems identification paradigm (Gerstner et al., 2012).

Rather, neuronal receptive field models – in particular, linear-nonlinear-Poisson (LNP) models – are computationally tractable abstractions of the mechanisms by which spike trains are generated (Herz et al., 2006). When more detailed correspondence between the model and the underlying biophysics is desired, the tractable outward shell of LNP and GLM models – the noisy nonlinear transduction mechanism – can be kept in place, while the linear component can be replaced by a more flexible and powerful component (Weber et al., 2010).

In this manner, it is possible to estimate neuronal systems identification models which are faithful to the underlying biology than while retaining the tractable outer shell of generalized linear models. This is the approach that I chose in the analysis of MST neurons in Chapter 4; by representing the stimulus in an MT-like representation, it became possible to relate the tuning of MST neurons to a biologically meaningful concept: their afferent input.

Of course, even with powerful statistical tools it is difficult to unambiguously determine that a given computation is performed by a certain neuron (Gerstner et al., 2012). A cortical visual neuron is embedded within a local network, which is embedded within an area, which is itself embedded with a hierarchy. Under these circumstances, it is challenging, given a systems identification model of a neuron, to make inferences about whether the uncovered computations are performed by that neuron, by the local network, by the area, or by the preceding areas in the visual hierarchy.

This consideration justifies the use of signals defined at multiple spatial scales to triangulate the location of a given computation. By leveraging single neuron, cortical network, behavioural, and simulation data, it becomes feasible to make hypotheses about the location, implementation, and function of a

computation. Of course, these hypotheses are necessarily speculative, but they can be used to judiciously guide experimentally challenging experiments that use activation and inactivation of neurons or neural ensembles to establish causal relationships between neurons and behaviour (Adesnik et al., 2012; Lien and Scanziani, 2013; Olsen et al., 2012).

5.3 Future directions

5.3.1 Local field potentials to study transient visual representation

In chapter 3, I showed that it is possible to faithfully estimate the representation of visual space over a patch of cortex using the local field potential, provided that appropriate care is taken in segregating the different sources of the LFP.

Surprisingly, even on electrodes where multi-unit activity could be recorded, LFP-derived receptive fields were frequently cleaner and more significant than corresponding MUA-derived receptive fields (e.g. Figure 3-2). This implies that local field potentials could be useful in studying transient visual representations, where receptive field estimation is particularly challenging.

Saccades allow relevant stimuli to be brought closer to the fovea, where visual resolution is highest (Schall and Thompson, 1999). This causes a disruption in the appearance of visual stimuli on the retina. Despite this, visual stimuli appear perceptually stable during active vision (Bridgeman et al., 1975; Yarrow et al., 2001). This is in part due to active mechanisms which cause changes in visual sensitivity around the time of saccades (Ross et al., 2001).

Of particular interest here is the perisaccadic remapping of receptive fields, where visual neurons become sensitive, presaccadically, to stimuli which will come into the static receptive field of the neuron following a saccade (Melcher and Colby, 2008). This observation raises many questions: Is remapping driven by a visual or an attentional signal? Where does the remapping signal come from? Are all neurons remapped simultaneously, in a top-down fashion, or does the remapping first take place in a group of seed neurons and then spreads laterally?

These questions may be further elucidated by studying the perisaccadic representation of space in ensembles of neurons. Preliminary results from our lab indicate that is possible to estimate perisaccadic receptive fields across an entire array multi-electrode array with local field potentials, using simple extensions of the methods developed in Chapter 3 (Neupane et al., 2014). This allows for the joint

estimation of the temporal, spatial, and orientation selectivity of remapped receptive fields, which should further constrain our understanding of visual representations around the time of saccades.

In related work, we have found that travelling waves of synaptic activity – reflected in the local field potential - spread across a patch of cortex during saccades (Zanos et al., 2014). The correlated activity helps reset the phase of neurons so that they may respond more strongly following a saccade. Here again, extensions of the methodology introduced in Chapter 3 have helped reveal the link between correlated synaptic activity and active visual representations.

5.3.2 Elucidating the source of nonlinear integration in MST

The analysis presented in Chapter 4 revealed that MST neurons nonlinearly integrate the output of MT neurons. Results could best be explained by the interaction between a compressive nonlinearity at the level of the input and an expansive spike generation nonlinearity.

I identified one candidate mechanism for the compressive input nonlinearity: synaptic depression. Synaptic depression has been hypothesized to have an important role in many neural computations, but physiological investigation of the role of synaptic depression in cortical sensory processing has been limited (Abbott et al., 1997; Chance et al., 1998; Rothman et al., 2009).

While it is difficult to address the role of synaptic depression in nonlinear integration in non-human primates, optic flow selectivity has been described in a wide range of species, from insects to rodents (Gabbiani et al., 2002; Sun and Frost, 1998; Weber et al., 2010; Yilmaz and Meister, 2013). With the genetic, pharmacological and imaging tools available in mice, it should be possible to investigate the role of synaptic depression in the computation of optic flow. I intend to do so as part of my postdoctoral research.

5.3.3 Elucidating hierarchical representations

Hierarchical processing iteratively reformats visual information into a format which is useful for behaviour (DiCarlo and Cox 2007). Visual representations in intermediate and high-level visual areas can therefore be understood as steps in a multi-stage hierarchical process. In Chapter 4, I showed that it was possible to leverage our knowledge of neurons in area MT to better understand the computations performed by MST neurons.

This suggests a bottom-up method to understand hierarchical processing, starting from the retina or the LGN:

1. Measure the responses of visual neurons in a given area to a rich, ecologically-relevant set of stimuli.
2. Find single neuron models which explain responses in the area under study in terms of the population response in the preceding area.
3. Derive a population response model from single neuron models.
4. Repeat this process for the next area.

While this is far from a trivial undertaking (Poggio et al., 2012; Riesenhuber and Poggio, 1999), current methodology allows us to simultaneously measure the responses of many neurons in several areas to a rich set of stimuli (Gill et al., 2006). Furthermore, it is possible to relate these responses to the stimuli which drove them using current systems identification approaches (Wu et al., 2006).

What is missing, then, is a robust method to go from single neuron to population response models. While it is possible to record from a few hundred neurons in a given area, it is improbable that neurons in the next recorded area will receive input from only these few hundred neurons. To avoid this difficulty, we need to build a model which can extrapolate from the responses a few hundred neurons to the responses of the many thousands of unseen neurons in that area.

The problem of inferring population representation from single unit recordings has received little attention in the literature. I propose to address this issue with a deep shared artificial neural network (dsANN) architecture. The underlying idea is directly related to the untangling hypothesis of DiCarlo and Cox (2007): find a function $f(\mathbf{X})$ which linearizes, via a set of weights $\mathbf{W}$, the relationship between the visual input $\mathbf{X}$ and the output of multiple neurons in a given area $\mathbf{Y}$:

$$Y_{ij} \approx \sum_k f(X_{ik}) W_{kj}$$

By leveraging recent advances in machine learning (Bengio, 2009), it is possible to learn a flexible transformation f by stacking physiologically plausible computations in a hierarchical scheme: linear integration within spatially restricted receptive fields, threshold nonlinearities, and normalization (Kouh and Poggio, 2008). $f(X_{ik})$ then captures, up to a linear transformation, the representation of visual stimuli in a given area by a biologically plausible hierarchical structure.

In preliminary work (Mineault and Pack, 2014), I have developed and applied this new methodology to the study of neurons in area V2 (Willmore et al., 2010). Preliminary results indicate that a deep

architecture trained on this data recovers local spatial frequency and orientation in the initial layers; intermediate layers reflect nonlinear combinations of these simple cell-like receptive fields (Mareschal and Baker, 1998); and later layers, like V2 neurons, are sensitive to natural textures (Freeman et al., 2013). I am now in the process of determining whether the inferred representation in V2 is predictive of responses in area V4, and repeating the same process in MT (Cui et al., 2013) to predict MST responses (Mineault et al., 2012).

These encouraging results indicate that it should be possible, by leveraging multi-electrode recordings from multiple areas, to build and constrain a deep hierarchical model of visual cortex. This is theoretically satisfying, as we should be able to fit and predict responses in visual cortex with a model which reflects the hypothesized underlying biology (Hubel and Wiesel, 1962).

More importantly, this could serve as an important in-silico tool to help constrain our understanding of the underlying biology of hierarchical computation (Poggio et al., 2012):

- Do we need more than stacked linear receptive fields and static nonlinearities to explain hierarchical processing (Kouh and Poggio, 2008)? What about normalization and synaptic depression?
- How can a hierarchical architecture solve the problem of the accumulation of neural noise (Kimpo et al., 2003)? Is noise a nuisance or a feature (Hinton et al., 2012)?
- Can the architecture account for the highest levels of visual representation, or is a novel architecture required to account for inferotemporal and parietal representations (Poggio et al., 2012)?

Answering these questions will require a significant research effort, but with the right parametric modelling methods, as I've presented here, I believe it will possible to further our understanding of hierarchical visual processing.

References

Abbey, C.K., and Eckstein, M.P. (2001). Maximum-likelihood and maximum-a-posteriori estimates of human-observer templates. In Proc. SPIE Vol. 4324, P. 114-122.

Abbey, C.K., and Eckstein, M.P. (2002). Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments. J Vis *2*, 66–78.

Abbey, C.K., Eckstein, M.P., Shimozaki, S.S., Baydush, A.H., Catarious, D.M., and Floyd, C.E. (2002). Human-observer templates for detection of a simulated lesion in mammographic images. In Proc. SPIE Vol. 4686, P. 25-36, Medical Imaging 2002.

Abbott, L., Varela, J., Sen, K., and Nelson, S. (1997). Synaptic depression and cortical gain control. Science *275*, 221.

Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. J Opt Soc Am A *2*, 284–299.

Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z.J., and Scanziani, M. (2012). A neural circuit for spatial summation in visual cortex. Nature *490*, 226–231.

Ahrens, M.B., Paninski, L., and Sahani, M. (2008a). Inferring input nonlinearities in neural encoding models. Netw. Comput. Neural Syst. *19*, 35–67.

Ahrens, M.B., Linden, J.F., and Sahani, M. (2008b). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J. Neurosci. Off. J. Soc. Neurosci. *28*, 1929–1942.

Ahumada, A.J., Jr (2002). Classification image weights and internal noise level estimation. J. Vis. *2*, 121–131.

Ahumada, A.J., and Lovell, J. (1971). Stimulus features in signal detection. J. Acoust. Soc. Am. *49*, 1751.

Ahumada Jr, A.J. (1996). Perceptual classification images from Vernier acuity masked by noise. Perception *26*, 1831–1840.

Albright, T.D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. J. Neurophysiol. *52*, 1106–1130.

Allman, J., Miezin, F., McGuinness, E., and others (1985). Direction-and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). Perception *14*, 105–126.

Anastassiou, C.A., Perin, R., Markram, H., and Koch, C. (2011). Ephaptic coupling of cortical neurons. Nat. Neurosci. *14*, 217–223.

Andersen, R.A., Asanuma, C., Essick, G., and Siegel, R.M. (1990). Corticocortical connections of anatomically and physiologically defined subdivisions within the inferior parietal lobule. J. Comp. Neurol. *296*, 65–113.

Anderson, J.C., Binzegger, T., Martin, K.A.., and Rockland, K.S. (1998). The connection from cortical area V1 to V5: a light and electron microscopic study. J. Neurosci. *18*, 10525.

Anderson, J.S., Lampl, I., Gillespie, D.C., and Ferster, D. (2000). The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. Science *290*, 1968–1972.

Andretic, R., Van Swinderen, B., and Greenspan, R.J. (2005). Dopaminergic Modulation of Arousal in Drosophila. Curr. Biol. *15*, 1165–1175.

Anzai, A., Peng, X., and Van Essen, D.C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. Nat. Neurosci. *10*, 1313–1321.

Atallah, B.V., and Scanziani, M. (2009). Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. Neuron *62*, 566–577.

Atick, J.J., and Redlich, A.N. (1992). What does the retina know about natural scenes? Neural Comput. *4*, 196–210.

Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nat. Rev. Neurosci. *7*, 358–366.

Baker, C., Li, G., Wang, Z., Yao, Z., Yuan, N., Talebi, V., Tan, J., Wang, Y., and Zhou, Y. (2013). Second-order neuronal responses to contrast modulation stimuli in primate visual cortex. J. Vis. *13*, 41–41.

Baker Jr, C.L. (1999). Central neural mechanisms for detecting second-order motion. Curr. Opin. Neurobiol. *9*, 461–466.

Bauer, R., Brosch, M., and Eckhorn, R. (1995). Different rules of spatial summation from beyond the receptive field for spike rates and oscillation amplitudes in cat visual cortex. Brain Res. *669*, 291–297.

Bedard, C., Kröger, H., and Destexhe, A. (2006). Model of low-pass filtering of local field potentials in brain tissue. Phys. Rev. E *73*, 051911.

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M.A., Logothetis, N.K., and Panzeri, S. (2008). Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. J. Neurosci. *28*, 5696–5709.

Bengio, Y. (2009). Learning deep architectures for AI. Found. Trends Mach. Learn. *2*, 1–127.

Berens, P., Keliris, G.A., Ecker, A.S., Logothetis, N.K., and Tolias, A.S. (2008). Feature Selectivity of the Gamma-Band of the Local Field Potential in Primate Primary Visual Cortex. Front. Neurosci. *2*, 199–207.

Berkes, P., and Wiskott, L. (2007). Analysis and interpretation of quadratic models of receptive fields. Nat Protoc. *2*, 400–407.

Bishop, C.M. (2006). Pattern recognition and machine learning (Springer New York).

Blasdel, G.G., and Fitzpatrick, D. (1984). Physiological organization of layer 4 in macaque striate cortex. J. Neurosci. *4*, 880–895.

De Boer, E., and Kuyper, P. (1968). Triggered correlation. IEEE Trans. Biomed. Eng. 169–179.

Born, R.T., and Bradley, D.C. (2005). Structure and function of visual area MT. Annu Rev Neurosci *28*, 157–189.

Bottou, L., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Jackel, L.D., LeCun, Y., Muller, U.A., Sackinger, E., and Simard, P. (1994). Comparison of classifier methods: a case study in handwritten digit recognition. In Proceedings of the 12th IAPR International Conference on Computer Vision & Image Processing, (IEEE), pp. 77–82.

Boussaoud, D., Ungerleider, L.G., and Desimone, R. (1990). Pathways for motion analysis: cortical connections of the medial superior temporal and fundus of the superior temporal visual areas in the macaque. J. Comp. Neurol. *296*, 462–495.

Bridgeman, B., Hendry, D., and Stark, L. (1975). Failure to detect displacement of the visual world during saccadic eye movements. Vision Res. *15*, 719–722.

Brincat, S.L., and Connor, C.E. (2006). Dynamic Shape Synthesis in Posterior Inferotemporal Cortex. Neuron *49*, 17–24.

Britten, K.H., and Heuer, H.W. (1999). Spatial Summation in the Receptive Fields of MT Neurons. J Neurosci *19*, 5074–5084.

Bühlmann, P., and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. Stat. Sci. 477–505.

Burt, P., and Adelson, E. (1983). The Laplacian Pyramid as a Compact Image Code. Commun. IEEE Trans. Leg. Pre - 1988 *31*, 532–540.

Butts, D.A., Weng, C., Jin, J., Yeh, C.I., Lesica, N.A., Alonso, J.M., and Stanley, G.B. (2007). Temporal precision in the neural code and the timescales of natural vision. Nature *449*, 92–95.

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

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. Nat. Neurosci. *7*, 446–451.

Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. *13*, 407–420.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C.E., Riesenhuber, M., and Poggio, T. (2007). A model of V4 shape selectivity and invariance. J. Neurophysiol. *98*, 1733.

Carandini, M., and Ferster, D. (2000). Membrane Potential and Firing Rate in Cat Primary Visual Cortex. J. Neurosci. *20*, 470–484.

Carandini, M., and Heeger, D.J. (2011). Normalization as a canonical neural computation. Nat. Rev. Neurosci.

Carandini, M., Heeger, D.J., and Movshon, J.A. (1997). Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. J. Neurosci. *17*, 8621–8644.

Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., and Rust, N.C. (2005). Do We Know What the Early Visual System Does? J. Neurosci. *25*, 10577–10597.

Castro, J.B., and Kandler, K. (2010). Changing tune in auditory cortex. Nat. Neurosci. *13*, 271.

Chance, F.S., Nelson, S.B., and Abbott, L.F. (1998). Synaptic Depression and the Temporal Response Characteristics of V1 Cells. J Neurosci *18*, 4785–4799.

Chauvin, A., Worsley, K.J., Schyns, P.G., Arguin, M., and Gosselin, F. (2005). Accurate statistical tests for smooth classification images. J Vis *5*, 659–667.

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

Chklovskii, D.B., and Koulakov, A.A. (2004). Maps in the brain: What can we learn from them? Annu Rev Neurosci *27*, 369–392.

Colby, C.L., and Goldberg, M.E. (1999). Space and attention in parietal cortex. Annu. Rev. Neurosci. *22*, 319–349.

Courtney, S.M., and Ungerleider, L.G. (1997). What fMRI has taught us about human vision. Curr. Opin. Neurobiol. *7*, 554–561.

Cover, T.M., and Thomas, J.A. (2012). Elements of information theory (John Wiley & Sons).

Cui, Y., Liu, L.D., Khawaja, F.A., Pack, C.C., and Butts, D.A. (2013). Diverse Suppressive Influences in Area MT and Selectivity to Complex Motion Features. J. Neurosci. *33*, 16715–16728.

Das, A., and Gilbert, C.D. (1997). Distortions of visuotopic map match orientation singularities in primary visual cortex. Nature *387*, 594–597.

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

David, S.V., Mesgarani, N., and Shamma, S.A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. Network *18*, 191–212.

Dayan, P., Abbott, L.F., and Abbott, L. (2001). Theoretical neuroscience: Computational and mathematical modeling of neural systems.

DeAngelis, G.C., Ohzawa, I., and Freeman, R.D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. J. Neurophysiol. *69*, 1118–1135.

Derrington, A.M., and Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. J. Physiol. *357*, 219–240.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How Does the Brain Solve Visual Object Recognition? Neuron *73*, 415–434.

Duffy, C.J., and Wurtz, R.H. (1991a). Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. J. Neurophysiol. *65*, 1329–1345.

Duffy, C.J., and Wurtz, R.H. (1991b). Sensitivity of MST neurons to optic flow stimuli. II. Mechanisms of response selectivity revealed by small-field stimuli. J. Neurophysiol. *65*, 1346.

Duffy, C.J., and Wurtz, R.H. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. J. Neurosci. *15*, 5192–5208.

Dumoulin, S.O., and Wandell, B.A. (2008). Population receptive field estimates in human visual cortex. Neuroimage *39*, 647–660.

Eckstein, M.P., and Ahumada, A.J. (2002). Classification images: A tool to analyze visual strategies. J. Vis. *2*.

Eggermont, J.J., Munguia, R., Pienkowski, M., and Shaw, G. (2011). Comparison of LFP-based and spike-based spectro-temporal receptive fields and cross-correlation in cat primary auditory cortex. PLoS One *6*, e20046.

Einevoll, G.T., Kayser, C., Logothetis, N.K., and Panzeri, S. (2013). Modelling and analysis of local field potentials for studying the function of cortical circuits. Nat. Rev. Neurosci. *14*, 770–785.

Elyada, Y.M., Haag, J., and Borst, A. (2009). Different receptive fields in axons and dendrites underlie robust coding in motion-sensitive neurons. Nat. Neurosci. *12*, 327–332.

Engel, S.A., Glover, G.H., and Wandell, B.A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex *7*, 181–192.

Fechner, G. (1860). Elemente der Psychophysik (Elements of psychophysics) (Leipzig: Breitkopf und Härtel).

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex *1*, 1–47.

Field, G.D., and Chichilnisky, E.J. (2007). Information processing in the primate retina: circuitry and coding. Annu Rev Neurosci *30*, 1–30.

Finn, I.M., Priebe, N.J., and Ferster, D. (2007). The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. Neuron *54*, 137–152.

Folk, C.L., Remington, R.W., and Wright, J.H. (1994). The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset, and color. J. Exp. Psychol. Hum. Percept. Perform. *20*, 317.

Freeman, W. (2007). Hilbert transform for brain waves. Scholarpedia *2*, 1338.

Freeman, J., and Simoncelli, E.P. (2011). Metamers of the ventral stream. Nat. Neurosci. *14*, 1195–1201.

Freeman, J., Brouwer, G.J., Heeger, D.J., and Merriam, E.P. (2011). Orientation Decoding Depends on Maps, Not Columns. J. Neurosci. *31*, 4792–4804.

Freeman, J., Ziemba, C.M., Heeger, D.J., Simoncelli, E.P., and Movshon, J.A. (2013). A functional and perceptual signature of the second visual area in primates. Nat. Neurosci. *16*, 974–981.

Freiwald, W.A., and Tsao, D.Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. Science *330*, 845–851.

Freiwald, W.A., Tsao, D.Y., and Livingstone, M.S. (2009). A face feature space in the macaque temporal lobe. Nat. Neurosci. *12*, 1187–1196.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. Ann. Stat. *28*, 337–374.

Fries, P., Reynolds, J.H., Rorie, A.E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. Science *291*, 1560–1563.

Fries, P., Womelsdorf, T., Oostenveld, R., and Desimone, R. (2008). The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. J. Neurosci. *28*, 4823–4835.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. *36*, 193–202.

Gabbiani, F., Krapp, H.G., Koch, C., and Laurent, G. (2002). Multiplicative computation in a visual neuron sensitive to looming. Nature *420*, 320–324.

Gallant, J.L., Braun, J., and Van Essen, D.C. (1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science *259*, 100–103.

Gallant, J.L., Connor, C.E., Rakshit, S., Lewis, J.W., and Van Essen, D.C. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. J. Neurophysiol. *76*, 2718–2739.

Gautama, T., and Van Hulle, M.M. (2001). Function of center-surround antagonism for motion in visual area MT/V5: a modeling study. Vision Res. *41*, 3917–3930.

Geesaman, B.J., and Andersen, R.A. (1996). The analysis of complex motion patterns by form/cue invariant MSTd neurons. J. Neurosci. *16*, 4716–4732.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. J. Am. Stat. Assoc. *70*, 320–328.

Gelfand, A.E., and Dey, D.K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. J. R. Stat. Soc. Ser. B Methodol. *56*, 501–514.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). Bayesian Data Analysis, Second Edition (Chapman and Hall/CRC).

Gerstner, W., Sprekeler, H., and Deco, G. (2012). Theory and Simulation in Neuroscience. Science *338*, 60–65.

Ghose, G.M., and Ts'O, D.Y. (1997). Form Processing Modules in Primate Area V4. J. Neurophysiol. *77*, 2191–2196.

Gibson, J. (1986). The ecological approach to visual perception (Boston: Houghton Mifflin).

Gieselmann, M.A., and Thiele, A. (2008). Comparison of spatial integration and surround suppression characteristics in spiking activity and the local field potential in macaque V1. Eur. J. Neurosci. *28*, 447–459.

Gilbert, C.D., and Wiesel, T.N. (1983). Clustered intrinsic connections in cat visual cortex. J. Neurosci. *3*, 1116–1133.

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

Gold, J.M., Murray, R.F., Bennett, P.J., and Sekuler, A.B. (2000). Deriving behavioural receptive fields for visually completed contours. Curr. Biol. *10*, 663–666.

Goldfine, A.M., Bardin, J.C., Noirhomme, Q., Fins, J.J., Schiff, N.D., and Victor, J.D. (2013). Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study."The Lancet *381*, 289–291.

Goodale, M.A., and Milner, A.D. (1992). Separate visual pathways for perception and action. Trends Neurosci. *15*, 20–25.

Gottlieb, J.P., Kusunoki, M., Goldberg, M.E., and others (1998). The representation of visual salience in monkey parietal cortex. Nature *391*, 481–484.

Gray, C.M., and Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc. Natl. Acad. Sci. *86*, 1698–1702.

Gray, C.M., Maldonado, P.E., Wilson, M., and McNaughton, B. (1995). Tetrodes markedly improve the reliability and yield of multiple single-unit isolation from multi-unit recordings in cat striate cortex. J. Neurosci. Methods *63*, 43–54.

Graziano, M.S., and Gross, C.G. (1995). The representation of extrapersonal space: A possible role for bimodal, visual-tactile neurons. Cogn. Neurosci. 1021–1034.

Graziano, M.S., Andersen, R.A., and Snowden, R.J. (1994). Tuning of MST neurons to spiral motions. J. Neurosci. *14*, 54–67.

Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics (Wiley New York).

Gregoriou, G.G., Gotts, S.J., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. Science *324*, 1207–1210.

Grossberg, S., Mingolla, E., and Pack, C.C. (1999). A neural model of motion processing and visual navigation by cortical area MST. Cereb. Cortex *9*, 878.

Gu, Y., Fetsch, C.R., Adeyemo, B., DeAngelis, G.C., and Angelaki, D.E. (2010). Decoding of MSTd population activity accounts for variations in the precision of heading perception. Neuron *66*, 596–609.

Haider, B., Häusser, M., and Carandini, M. (2013). Inhibition dominates sensory responses in the awake cortex. Nature *493*, 97–100.

Hale, E.T., Yin, W., and Zhang, Y. (2007). A Fixed-Point Continuation Method for l1 -Regularized Minimization with Applications to Compressed Sensing (Rice University).

Ben Hamed, S., Page, W., Duffy, C., and Pouget, A. (2003). MSTd neuronal basis functions for the population encoding of heading direction. J Neurophysiol *90*, 549–558.

Hanazawa, A., and Komatsu, H. (2001). Influence of the direction of elemental luminance gradients on the responses of V4 cells to textured surfaces. J. Neurosci. *21*, 4490–4497.

Hastie, T. (2007). Comment: Boosting Algorithms: Regularization, Prediction and Model Fitting. Stat. Sci. *22*, 513–515.

Hastie, T., and Tibshirani, R. (1990). Generalized additive models (Chapman & Hall/CRC).

Hastie, T., Tibshirani, R., and Friedman, J.J.H. (2001). The elements of statistical learning (Springer New York).

Hatsopoulos, N., Gabbiani, F., and Laurent, G. (1995). Elementary Computation of Object Approach by a Wide-Field Visual Neuron. Science *270*, 1000–1003.

Hattar, S., Liao, H.-W., Takao, M., Berson, D.M., and Yau, K.-W. (2002). Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. Science *295*, 1065–1070.

Heeger (1992). Normalization of cell responses in cat striate cortex. J. Neurosci. 181–197.

Heeger, D.J., and Carandini (1994). Summation and division by neurons in primate visual cortex. Science 1333–1336.

Hegdé, J., and Van Essen, D.C. (2000). Selectivity for complex shapes in primate visual area V2. J. Neurosci. *20*, 61–61.

Henrie, J.A., and Shapley, R. (2005). LFP power spectra in V1 cortex: the graded effect of stimulus contrast. J. Neurophysiol. *94*, 479–490.

Herz, A.V.M., Gollisch, T., Machens, C.K., and Jaeger, D. (2006). Modeling Single-Neuron Dynamics and Computations: A Balance of Detail and Abstraction. Science *314*, 80–85.

Hinkle, D.A., and Connor, C.E. (2002). Three-dimensional orientation tuning in macaque area V4. Nat. Neurosci. *5*, 665–670.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. ArXiv12070580 Cs.

Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. *117*, 500.

Hubel, D.H. (1995). Eye, brain, and vision. (W.H. Freeman).

Hubel, D.H., and Livingstone, M.S. (1987). Segregation of form, color, and stereopsis in primate area 18. J. Neurosci. *7*, 3378–3415.

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

Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. J. Physiol. *195*, 215–243.

Hubel, D.H., and Wiesel, T.N. (1977). Ferrier lecture: Functional architecture of macaque monkey visual cortex. Proc. R. Soc. Lond. B Biol. Sci. 1–59.

Huberman, A.D. (2007). Mechanisms of eye-specific visual circuit development. Curr. Opin. Neurobiol. *17*, 73–80.

Huth, A., Nishimoto, S., Vu, A., and Gallant, J. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron *76*, 1210–1224.

Hwang, E.J., and Andersen, R.A. (2011). Effects of visual stimulation on LFPs, spikes, and LFP-spike relations in PRR. J. Neurophysiol. *105*, 1850–1860.

Jeffreys, H. (1998). The theory of probability (Oxford University Press).

Jia, X., Smith, M.A., and Kohn, A. (2011). Stimulus Selectivity and Spatial Coherence of Gamma Components of the Local Field Potential. J. Neurosci. *31*, 9390–9403.

Johannesma, P.I. (1972). The pre-response stimulus ensemble of neurons in the cochlear nucleus. In Proceedings of the Symposium of Hearing Theory,.

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

Kajikawa, Y., and Schroeder, C.E. (2011). How local is the local field potential? Neuron *72*, 847–858.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. *8*, 679–685.

Kandel, E.R., Schwartz, J.H., and Jessell, T.M. (2000). Principles of neural science (McGraw-Hill New York).

Kaschube, M., Schnabel, M., Löwel, S., Coppola, D.M., White, L.E., and Wolf, F. (2010). Universality in the Evolution of Orientation Columns in the Visual Cortex. Science *330*, 1113–1116.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. J. Am. Stat. Assoc. *90*, 773–795.

Katzner, S., Nauhaus, I., Benucci, A., Bonin, V., Ringach, D.L., and Carandini, M. (2009). Local Origin of Field Potentials in Visual Cortex. Neuron *61*, 35–41.

Khawaja, F.A., Tsui, J.M.G., and Pack, C.C. (2009). Pattern motion selectivity of spiking outputs and local field potentials in macaque visual cortex. J. Neurosci. *29*, 13702–13709.

Khawaja, F.A., Liu, L.D., and Pack, C.C. (2013). Responses of MST neurons to plaid stimuli. J. Neurophysiol. *110*, 63–74.

Kimpo, R.R., Theunissen, F.E., and Doupe, A.J. (2003). Propagation of Correlated Activity through Multiple Stages of a Neural Circuit. J. Neurosci. *23*, 5750–5761.

Knoblauch, K., and Maloney, L. (2008). Classification images estimated by generalized additive models. J Vis *8*, 344–344.

Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol. *71*, 856–867.

Koch, C. (1999). Biophysics of computation: information processing in single neurons (Oxford University Press, USA).

Koenderink, J.J. (1986). Optic flow. Vision Res. *26*, 161–179.

Kouh, M., and Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. Neural Comput. *20*, 1427–1451.

Koulakov, A.A., and Chklovskii, D.B. (2001). Orientation preference patterns in mammalian visual cortex: a wire length minimization approach. Neuron *29*, 519–527.

Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object Selectivity of Local Field Potentials and Spikes in the Macaque Inferior Temporal Cortex. Neuron *49*, 433–445.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., and Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. *12*, 535–540.

Kusunoki, M., Gottlieb, J., and Goldberg, M.E. (2000). The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task relevance. Vision Res. *40*, 1459–1468.

Lagae, L., Maes, H., Raiguel, S., Xiao, D., and Orban, G. (1994). Responses of macaque STS neurons to optic flow components: a comparison of areas MT and MST. J. Neurophysiol. *71*, 1597–1626.

Lappe, M. (2000). Computational mechanisms for optic flow analysis in primate cortex. Int. Rev. Neurobiol. 235–268.

Lashgari, R., Li, X., Chen, Y., Kremkow, J., Bereshpolova, Y., Swadlow, H.A., and Alonso, J.M. (2012). Response Properties of Local Field Potentials and Neighboring Single Neurons in Awake Primary Visual Cortex. J. Neurosci. *32*, 11396–11413.

Lee, T.-W., Wachtler, T., and Sejnowski, T.J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. Vision Res. *42*, 2095–2103.

Levi, D.M., and Klein, S.A. (2002). Classification images for detection and position discrimination in the fovea and parafovea. J. Vis. *2*, 4.

Levitt, J.B., Kiper, D.C., and Movshon, J.A. (1994). Receptive fields and functional architecture of macaque V2. J. Neurophysiol. *71*, 2517–2542.

Li, G., and Baker, C.L. (2012). Functional Organization of Envelope-Responsive Neurons in Early Visual Cortex: Organization of Carrier Tuning Properties. J. Neurosci. *32*, 7538–7549.

Liebe, S., Logothetis, N.K., and Rainer, G. (2011). Dissociable Effects of Natural Image Structure and Color on LFP and Spiking Activity in the Lateral Prefrontal Cortex and Extrastriate Visual Area V4. J. Neurosci. *31*, 10215–10227.

Lien, A.D., and Scanziani, M. (2013). Tuned thalamic excitation is amplified by visual cortical circuits. Nat. Neurosci.

Lindén, H., Tetzlaff, T., Potjans, T.C., Pettersen, K.H., Grün, S., Diesmann, M., and Einevoll, G.T. (2011). Modeling the spatial reach of the LFP. Neuron *72*, 859–872.

Liu, J., and Newsome, W.T. (2006). Local field potential in cortical area MT: stimulus tuning and behavioral correlations. J. Neurosci. *26*, 7779–7790.

Livingstone, M., and Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science *240*, 740–749.

Livingstone, M.S., Pack, C.C., and Born, R.T. (2001). Two-dimensional substructure of MT receptive fields. Neuron *30*, 781–793.

Lochmann, T., Blanche, T.J., and Butts, D.A. (2013). Construction of Direction Selectivity through Local Energy Computations in Primary Visual Cortex. PLoS ONE *8*, e58666.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature *412*, 150–157.

MacCullagh, P., and Nelder, J.A. (1989). Generalized linear models (CRC press).

Mackay, D.J.C. (2002). Information Theory, Inference & Learning Algorithms (Cambridge University Press).

Mangini, M.C., and Biederman, I. (2005). Making the ineffable explicit: estimating the information employed for face classifications. Cogn. Sci. *28*, 209–226.

Mante, V., Bonin, V., and Carandini, M. (2008). Functional Mechanisms Shaping Lateral Geniculate Responses to Artificial and Natural Stimuli. Neuron *58*, 625–638.

Mareschal, I., and Baker, C.L. (1998). Temporal and spatial response to second-order stimuli in cat area 18. J. Neurophysiol. *80*, 2811–2823.

Marmarelis, P.Z., and Marmarelis, V.Z. (1978). Analysis of physiological systems: The white-noise approach (Plenum Press New York).

Marmarelis, P.Z., and Naka, K.I. (1973). Nonlinear analysis and synthesis of receptive-field responses in the catfish retina. I. Horizontal cell leads to ganglion cell chain. J. Neurophysiol. *36*, 605–618.

Marr, D. C. (1982) Vision: a Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press. Cambridge, MA.

Maunsell, J.H., and Van Essen, D.C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. J Neurophysiol *49*, 1127–1147.

Maynard, E.M., Nordhausen, C.T., and Normann, R.A. (1997). The Utah intracortical electrode array: a recording structure for potential brain-computer interfaces. Electroencephalogr. Clin. Neurophysiol. *102*, 228–239.

McFarland, J.M., Cui, Y., and Butts, D.A. (2013). Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs. PLoS Comput Biol *9*, e1003143.

Melcher, D., and Colby, C.L. (2008). Trans-saccadic perception. Trends Cogn. Sci. *12*, 466–473.

Milstein, J., Mormann, F., Fried, I., and Koch, C. (2009). Neuronal shot noise and Brownian 1/f2 behavior in the local field potential. PLoS One *4*, e4338.

Mineault, P.J., and Pack, C.C. (2008). Getting the most out of classification images. J Vis *8*, 271–271.

Mineault, P.J., and Pack, C.C. (2013). The Cerebral Emporium of Benevolent Knowledge. Neuron *79*, 833–835.

Mineault, P.J., and Pack, C.C. (2014). Learning hierarchical representations with deep shared artificial neural nets. (Warwick, UK),.

Mineault, P.J., Barthelmé, S., and Pack, C.C. (2009). Improved classification images with sparse priors in a smooth basis. J. Vis. *9*, 17.

Mineault, P.J., Khawaja, F.A., Butts, D.A., and Pack, C.C. (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. Proc. Natl. Acad. Sci. *109*, 972–980.

Mineault, P.J., Zanos, T.P., and Pack, C.C. (2013). Local field potentials reflect multiple spatial scales in V4. Front. Comput. Neurosci. *7*.

Motter, B.C. (2009). Central V4 Receptive Fields Are Scaled by the V1 Cortical Magnification and Correspond to a Constant-Sized Sampling of the V1 Surface. J. Neurosci. *29*, 5749 –5757.

Movshon, J.A., and Newsome, W.T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. J. Neurosci. *16*, 7733.

Movshon, Adelson, Gizzi, and Newsome (1985). The analysis of moving visual patterns. In Pattern Recognition Mechanisms, (Rome: Vatican Press), pp. 117–151.

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

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science *309*, 951–954.

Münch, T.A., da Silveira, R.A., Siegert, S., Viney, T.J., Awatramani, G.B., and Roska, B. (2009). Approach sensitivity in the retina processed by a multifunctional neural circuit. Nat. Neurosci. *12*, 1308–1316.

Murray, R.F., Bennett, P.J., and Sekuler, A.B. (2002). Optimal methods for calculating classification images: weighted sums. J Vis *2*, 79–104.

Nakamura, H., Gattass, R., Desimone, R., and Ungerleider, L.G. (1993). The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. J. Neurosci. *13*, 3681–3691.

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400–410.

Nauhaus, I., Benucci, A., Carandini, M., and Ringach, D.L. (2008). Neuronal selectivity and local map structure in visual cortex. Neuron *57*, 673–679.

Nauhaus, I., Busse, L., Carandini, M., and Ringach, D.L. (2009). Stimulus contrast modulates functional connectivity in visual cortex. Nat Neurosci *12*, 70–76.

Nelson, M.J., Pouget, P., Nilsen, E.A., Patten, C.D., and Schall, J.D. (2008). Review of signal distortion through metal microelectrode recording circuits and filters. J. Neurosci. Methods *169*, 141–157.

Neri, P., and Heeger, D.J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. Nat. Neurosci. *5*, 812–816.

Neri, P., and Levi, D. (2008a). Temporal dynamics of directional selectivity in human vision. J. Vis. *8*, 22.1–11.

Neri, P., and Levi, D.M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. Vision Res. *46*, 2465–2474.

Neri, P., and Levi, D.M. (2008b). Evidence for joint encoding of motion and disparity in human visual perception. J. Neurophysiol. *100*, 3117–3133.

Neri, P., Parker, A.J., and Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. Nature *401*, 695–698.

Neupane, S., Guitton, D., and Pack, C. (2014). Dynamics of peri-saccadic receptive fields in monkey area V4. (Montréal),.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., and Malach, R. (2007). Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. Curr. Biol. *17*, 1275–1285.

Nishimoto, S., and Gallant, J.L. (2011). A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies. J. Neurosci. *31*, 14551–14564.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. *21*, 1641–1646.

Noë, A. (2004). Action in perception (MIT Press).

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. *10*, 424–430.

Nover, H., Anderson, C.H., and DeAngelis, G.C. (2005). A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. J. Neurosci. *25*, 10049.

Ohki, K., Chung, S., Ch'ng, Y.H., Kara, P., and Reid, R.C. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. Nature *433*, 597–603.

Ohki, K., Chung, S., Kara, P., Hübener, M., Bonhoeffer, T., and Reid, R.C. (2006). Highly ordered arrangement of single neurons in orientation pinwheels. Nature *442*, 925–928.

Olman, C., and Kersten, D. (2004). Classification objects, ideal observers & generative models. Cogn. Sci. *28*, 227–239.

Olsen, S.R., Bortone, D.S., Adesnik, H., and Scanziani, M. (2012). Gain control by layer six in cortical circuits of vision. Nature *483*, 47–52.

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

Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. *14*, 481–487.

Orban, G.A. (2008). Higher order visual processing in macaque extrastriate cortex. Physiol. Rev. *88*, 59–89.

Orban, G., Lagae, L., Verri, A., Raiguel, S., Xiao, D., Maes, H., and Torre, V. (1992). First-order analysis of optical flow in monkey brain. Proc. Natl. Acad. Sci. *89*, 2595.

Orban, G.A., Kennedy, H., and Bullier, J. (1986). Velocity sensitivity and direction selectivity of neurons in areas V1 and V2 of the monkey: influence of eccentricity. J. Neurophysiol. *56*, 462–480.

Pack, C.C., and Born, R.T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. Nature *409*, 1040–1042.

Pack, C.C., Livingstone, M.S., Duffy, K.R., and Born, R.T. (2003a). End-stopping and the aperture problem: two-dimensional motion signals in macaque V1. Neuron *39*, 671–680.

Pack, C.C., Born, R.T., and Livingstone, M.S. (2003b). Two-dimensional substructure of stereo and motion interactions in macaque visual cortex. Neuron *37*, 525–535.

Pack, C.C., Hunter, J.N., and Born, R.T. (2005). Contrast Dependence of Suppressive Influences in Cortical Area MT of Alert Macaque. J. Neurophysiol. *93*, 1809 –1815.

Pagan, M., Urban, L.S., Wohl, M.P., and Rust, N.C. (2013). Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. Nat. Neurosci. *16*, 1132–1139.

Paik, S.-B., and Ringach, D.L. (2012). Link between orientation and retinotopic maps in primary visual cortex. Proc. Natl. Acad. Sci.

Palanca, B.J.A., and DeAngelis, G.C. (2003). Macaque Middle Temporal Neurons Signal Depth in the Absence of Motion. J. Neurosci. *23*, 7647 –7658.

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

Paninski, L. (2005). Asymptotic Theory of Information-Theoretic Experimental Design. Neural Comput *17*, 1480–1507.

Paninski, L., Ahmadian, Y., Ferreira, D.G., Koyama, S., Rad, K.R., Vidne, M., Vogelstein, J., and Wu, W. (2010). A new look at state-space models for neural data. J. Comput. Neurosci. *29*, 107–126.

Paolini, M., Distler, C., Bremmer, F., Lappe, M., and Hoffmann, K.P. (2000). Responses to continuously changing optic flow in area MST. J. Neurophysiol. *84*, 730.

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

Park, M.Y., and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. *69*, 659–677.

Park, I.M., Archer, E.W., Priebe, N., and Pillow, J. (2013). Spectral methods for neural characterization using generalized quadratic models. In Advances in Neural Information Processing Systems, pp. 2454–2462.

Pasupathy, A., and Connor, C.E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. J. Neurophysiol. *86*, 2505–2519.

Pasupathy, A., and Connor, C.E. (2002). Population coding of shape in area V4. Nat. Neurosci. *5*, 1332–1338.

Peirce, J.W. (2007). The potential importance of saturating and supersaturating contrast response functions in visual cortex. J. Vis. *7*, 1–10.

Pena, J.L., and Konishi, M. (2001). Auditory Spatial Receptive Fields Created by Multiplication. Science *292*, 249–252.

Perrone, J.A., and Krauzlis, R.J. (2008). Spatial integration by MT pattern neurons: a closer look at pattern-to-component effects and the role of speed tuning. J Vis *8*, 1 1–14.

Perrone, J.A., and Stone, L.S. (1994). A model of self-motion estimation within primate extrastriate visual cortex. Vision Res. *34*, 2917–2938.

Perrone, J.A., and Stone, L.S. (1998). Emulating the visual receptive-field properties of MST neurons with a template model of heading estimation. J. Neurosci. *18*, 5958.

Pfau, D., Pnevmatikakis, E.A., and Paninski, L. (2013). Robust learning of low dimensional dynamics from large neural ensembles. In Neural Information Processing Systems,.

Pillow, J.W., and Park, I.M. (2011). Bayesian spike-triggered covariance analysis. In Advances in Neural Information Processing Systems, pp. 1692–1700.

Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., and Simoncelli, E.P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature *454*, 995–999.

Poggio, T., Verri, A., and Torre, V. (1991). Green Theorems and Qualitative Properties of the Optical Flow (AI Memo AIM-1289) (MIT).

Poggio, T., Mutch, J., Leibo, J., Rosasco, L., and Tacchetti, A. (2012). The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work).

Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. Int. J. Comput. Vis. *40*, 49–70.

Pouget, A., and Sejnowski, T.J. (1997). Spatial transformations in the parietal cortex using basis functions. J. Cogn. Neurosci. *9*, 222–237.

Priebe, N.J., Cassanello, C.R., and Lisberger, S.G. (2003). The Neural Representation of Speed in Macaque Area MT/V5. J. Neurosci. *23*, 5650–5661.

Priebe, N.J., Mechler, F., Carandini, M., and Ferster, D. (2004). The contribution of spike threshold to the dichotomy of cortical simple and complex cells. Nat. Neurosci. *7*, 1113–1122.

Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature *435*, 1102–1107.

Quiroga, R.Q., Reddy, L., Koch, C., and Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. J. Neurophysiol. *98*, 1997–2007.

Raiguel, S., Hulle, M., Xiao, D.K., Marcar, V., and Orban, G. (1995). Shape and spatial distribution of receptive fields and antagonistic motion surrounds in the middle temporal area (V5) of the macaque. Eur. J. Neurosci. *7*, 2064–2082.

Raiguel, S., Van Hulle, M.M., Xiao, D., Marcar, V.L., Lagae, L., and Orban, G.A. (1997). Size and shape of receptive fields in the medial superior temporal area (MST) of the macaque. Neuroreport *8*, 2803.

Rajan, K., Marre, O., and Tka\vcik, G. (2012). Learning quadratic receptive fields from neural responses to natural stimuli. ArXiv Prepr. ArXiv12090121.

Rajashekar, U., Bovik, A.C., and Cormack, L.K. (2006). Visual search in noise: revealing the influence of structural cues by gaze-contingent classification image analysis. J. Vis. *6*, 379–386.

Rasch, M.J., Gretton, A., Murayama, Y., Maass, W., and Logothetis, N.K. (2008). Inferring spike trains from local field potentials. J. Neurophysiol. *99*, 1461–1476.

Rasmussen, C.E., and Williams, C.K.I. (2006). Gaussian processes for machine learning (MIT press Cambridge, MA).

Ratcliff, R., and Rouder, J.N. (1998). Modeling Response Times for Two-Choice Decisions. Psychol. Sci. *9*, 347–356.

Ray, S., and Maunsell, J.H.R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol. *9*, e1000610.

Reid, R.C., and Alonso, J.-M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. Nature *378*, 281–283.

Reid, R.C., and Shapley, R.M. (1992). Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

Ringach, D.L. (2004). Haphazard Wiring of Simple Receptive Fields and Orientation Columns in Visual Cortex. J. Neurophysiol. *92*, 468–476.

Ringach, D.L., Sapiro, G., and Shapley, R. (1997). A subspace reverse-correlation technique for the study of visual neurons. Vision Res. *37*, 2455–2464.

Ringach, D.L., Shapley, R.M., and Hawken, M.J. (2002). Orientation Selectivity in Macaque V1: Diversity and Laminar Dependence. J. Neurosci. *22*, 5639–5651.

Ross, M.G., and Cohen, A.L. (2009). Using graphical models to infer multiple visual classification features. J. Vis. *9*, 23.

Ross, J., Morrone, M.C., Goldberg, M.E., and Burr, D.C. (2001). Changes in visual perception at the time of saccades. Trends Neurosci. *24*, 113–121.

Rosset, S., Zhu, J., Hastie, T., and Schapire, R. (2004). Boosting as a regularized path to a maximum margin classifier. J. Mach. Learn. Res. *5*, 941–973.

Rothman, J.S., Cathala, L., Steuber, V., and Silver, R.A. (2009). Synaptic depression enables neuronal gain control. Nature *457*, 1015–1018.

Rousche, P.J., and Normann, R.A. (1998). Chronic recording capability of the Utah Intracortical Electrode Array in cat sensory cortex. J. Neurosci. Methods *82*, 1–15.

Rousche, P.J., Petersen, R.S., Battiston, S., Giannotta, S., and Diamond, M.E. (1999). Examination of the spatial and temporal distribution of sensory cortical activity using a 100-electrode array. J. Neurosci. Methods *90*, 57–66.

Royden, C.S., Crowell, J.A., and Banks, M.S. (1994). Estimating heading during eye movements. Vision Res. *34*, 3197–3214.

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

Rust, N.C., Mante, V., Simoncelli, E.P., and Movshon, J.A. (2006). How MT cells analyze the motion of visual patterns. Nat. Neurosci. *9*, 1421–1431.

Sahani, M., and Linden, J.F. (2003a). Evidence Optimization Techniques for Estimating Stimulus-Response Functions.

Sahani, M., and Linden, J.F. (2003b). How linear are auditory cortical responses? Adv. Neural Inf. Process. Syst. 125–132.

Salzman, C.D., Murasugi, C.M., Britten, K.H., and Newsome, W.T. (1992). Microstimulation in visual area MT: effects on direction discrimination performance. J. Neurosci. *12*, 2331–2355.

Schall, J.D., and Thompson, K.G. (1999). Neural selection and control of visually guided eye movements. Annu. Rev. Neurosci. *22*, 241–259.

Schmolesky, M.T., Wang, Y., Hanes, D.P., Thompson, K.G., Leutgeb, S., Schall, J.D., and Leventhal, A.G. (1998). Signal timing across the macaque visual system. J. Neurophysiol. *79*, 3272–3278.

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

Seeger, M.W. (2008). Bayesian Inference and Optimal Design for the Sparse Linear Model. J Mach Learn Res *9*, 759–813.

Seeger, M., Gerwinn, S., and Bethge, M. Bayesian Inference for Sparse Generalized Linear Models.

Sekuler, A.B., Gaspar, C.M., Gold, J.M., and Bennett, P.J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. Curr. Biol. *14*, 391–396.

Selesnick, I.W., Baraniuk, R.G., and Kingsbury, N.C. (2005). The dual-tree complex wavelet transform. Signal Process. Mag. IEEE *22*, 123–151.

Shapley, R., Kaplan, E., Purpura, K., and Lam, D.M. (1993). Contrast sensitivity and light adaptation in photoreceptors or in the retinal network. Contrast Sensit. *5*, 103–116.

Sharpee, T.O. (2013). Computational Identification of Receptive Fields. Annu. Rev. Neurosci. *36*, 103–120.

Shipp, S., and Zeki, S. (1985). Segregation of pathways leading from area V2 to areas V4 and V5 of macaque monkey visual cortex. Nature *315*, 322–324.

Shipp, S., and Zeki, S. (1989). The Organization of Connections between Areas V5 and V1 in Macaque Monkey Visual Cortex. Eur. J. Neurosci. *1*, 309–332.

Siegel, R.M., and Read, H.L. (1997). Analysis of optic flow in the monkey parietal area 7a. Cereb. Cortex *7*, 327–346.

Simoncelli, E.P., and Freeman, W.T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. In ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol. 3)-Volume 3, (Washington, DC, USA: IEEE Computer Society),.

Simoncelli, E.P., and Heeger, D.J. (1998). A model of neuronal responses in visual area MT. Vision Res. *38*, 743–761.

Simoncelli, E.P., Pillow, J., Paninski, L., and Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In The Cognitive Neurosciences, III, M. Gazzaniga, ed. (MIT Press), pp. 327–338.

Solomon, J.A. (2002). Noise reveals visual mechanisms of detection and discrimination. J. Vis. *2*, 7.

Somers, D.C., Nelson, S.B., and Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. J. Neurosci. *15*, 5448.

Srivastava, A., Lee, A.B., Simoncelli, E.P., and Zhu, S.-C. (2003). On Advances in Statistical Modeling of Natural Images. J. Math. Imaging Vis. *18*, 17–33.

Stansbury, D.E., Naselaris, T., and Gallant, J.L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron *79*, 1025–1034.

Sun, H., and Frost, B.J. (1998). Computation of different optical variables of looming objects in pigeon nucleus rotundus neurons. Nat Neurosci *1*, 296–303.

Supèr, H., and Roelfsema, P.R. (2005). Chronic multiunit recordings in behaving animals: advantages and limitations. In Progress in Brain Research, M.K. J. van Pelt, ed. (Elsevier), pp. 263–282.

Van Swinderen, B., Nitz, D.A., and Greenspan, R.J. (2004). Uncoupling of Brain Activity from Movement Defines Arousal States in Drosophila. Curr. Biol. *14*, 81–87.

Tadin, D., Lappin, J.S., and Blake, R. (2006). Fine Temporal Properties of Center–Surround Interactions in Motion Revealed by Reverse Correlation. J. Neurosci. *26*, 2614–2622.

Takemura, A., Inoue, Y., Kawano, K., Quaia, C., and Miles, F.A. (2001). Single-Unit Activity in Cortical Area MST Associated With Disparity-Vergence Eye Movements: Evidence for Population Coding. J Neurophysiol *85*, 2245–2266.

Takemura, A., Murata, Y., Kawano, K., and Miles, F.A. (2007). Deficits in Short-Latency Tracking Eye Movements after Chemical Lesions in Monkey Cortical Areas MT and MST. J Neurosci *27*, 529–541.

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. Curr. Opin. Neurobiol. *2*, 502–505.

Tanaka, K., Hikosaka, K., Saito, H., Yukie, M., Fukada, Y., and Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. J. Neurosci. *6*, 134–144.

Tanaka, K., Fukada, Y., and Saito, H.A. (1989). Underlying mechanisms of the response specificity of expansion/contraction and rotation cells in the dorsal part of the medial superior temporal area of the macaque monkey. J. Neurophysiol. *62*, 642–656.

Thomas, J.P., and Knoblauch, K. (2005). Frequency and phase contributions to the detection of temporal luminance modulation. J. Opt. Soc. Am. A *22*, 2257–2261.

Thomson, A.M., and Bannister, A.P. (2003). Interlaminar Connections in the Neocortex. Cereb. Cortex *13*, 5–14.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.

Tjan, B.S., and Nandy, A.S. (2006). Classification images with uncertainty. J. Vis. *6*.

Tohyama, K., and Fukushima, K. (2005). Neural network model for extracting optic flow. Neural Netw. *18*, 549–556.

Tsui, J.M.G., and Pack, C.C. (2011). Contrast sensitivity of MT receptive field centers and surrounds. J. Neurophysiol. *106*, 1888–1900.

Tsui, J.M.G., Hunter, J.N., Born, R.T., and Pack, C.. (2010). The role of V1 surround suppression in MT motion integration. J. Neurophysiol. *103*, 3123–3138.

Ungerleider, L.G., and Mishkin, M. (1982). Two cortical visual systems. Anal. Vis. Behav. 549–586.

Usrey, M., Reid, C., and Alonso, J.-M. (2001). Rules of Connectivity between Geniculate Cells and Simple Cells in Cat Primary Visual Cortex. J Neurosci *21*, 4002–4015.

De Valois, R.L., William Yund, E., and Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. Vision Res. *22*, 531–544.

Vanduffel, W., Tootell, R.B.H., Schoups, A.A., and Orban, G.A. (2002). The Organization of Orientation Selectivity Throughout Macaque Visual Cortex. Cereb. Cortex *12*, 647–662.

Victor, J.D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. Nat. Neurosci. *8*, 1651–1656.

Vintch, B., Zaharia, A., Movshon, J., and Simoncelli, E. (2012). Efficient and direct estimation of a neural subunit model for sensory coding. NIPS.

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica *57*, 307.

Warren, W.H., Kay, B.A., Zosh, W.D., Duchon, A.P., and Sahuc, S. (2001). Optic flow is used to control human walking. Nat. Neurosci. *4*, 213–216.

Weber, F., Machens, C.K., and Borst, A. (2010). Spatiotemporal response properties of optic-flow processing neurons. Neuron *67*, 629–642.

Wiener, N. (1966). Nonlinear problems in random theory. Nonlinear Probl. Random Theory Norbert Wien. Pp 142 ISBN 0-262-73012-X Camb. Mass. USA MIT Press August 1966Paper *1*.

Willmore, B.D.., Prenger, R.J., and Gallant, J.L. (2010). Neural Representation of Natural Images in Visual Area V2. J. Neurosci. *30*, 2102.

Wilson, N.R., Runyan, C.A., Wang, F.L., and Sur, M. (2012). Division and subtraction by distinct cortical inhibitory networks in vivo. Nature *488*, 343–348.

Wipf, D., and Nagarajan, S. (2008). A New View of Automatic Relevance Determination. In Advances in Neural Information Processing Systems 20, J.C. Platt, D. Koller, Y. Singer, and S. Roweis, eds. (Cambridge, MA: MIT Press), pp. 1625–1632.

Wolpert, D.M., Goodbody, S.J., and Husain, M. (1998). Maintaining internal representations: the role of the human superior parietal lobe. Nat. Neurosci. *1*, 529–533.

Wood, S. (2006). Generalized Additive Models: An Introduction with R (University of Bath, England, UK: Chapman & Hall).

Worsley, K.J., Marrett, S., Neelin, P., and Evans, A.C. (1996). Searching scale space for activation in PET images. Hum. Brain Mapp. *4*, 74–90.

Wu, M.C.., David, S.V., and Gallant, J.L. (2006). Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci *29*, 477–505.

Xiao, D., Raiguel, S., Marcar, V., Koenderink, J., and Orban, G. (1995). Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. Proc. Natl. Acad. Sci. *92*, 11303.

Xing, D., Yeh, C.-I., and Shapley, R.M. (2009). Spatial Spread of the Local Field Potential and its Laminar Variation in Visual Cortex. J. Neurosci. *29*, 11540 –11549.

Yarrow, K., Haggard, P., Heal, R., Brown, P., and Rothwell, J.C. (2001). Illusory perceptions of space and time preserve cross-saccadic perceptual continuity. Nature *414*, 302–305.

Yilmaz, M., and Meister, M. (2013). Rapid innate defensive responses of mice to looming visual stimuli. Curr. Biol. *23*, 2011–2015.

Yu, C.P., Page, W.K., Gaborski, R., and Duffy, C.J. (2010). Receptive Field Dynamics Underlying MST Neuronal Optic Flow Selectivity. J Neurophysiol *103*, 2794–2807.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. *68*, 49–67.

Zanos, T.P., Mineault, P.J., and Pack, C.C. (2011a). Removal of Spurious Correlations Between Spikes and Local Field Potentials. J. Neurophysiol. *105*, 474–486.

Zanos, T.P., Mineault, P.J., Monteon, J.A., and Pack, C.C. (2011b). Functional connectivity during surround suppression in macaque area V4. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. *2011*, 3342–3345.

Zanos, T.P., Mineault, P.J., and Pack, C.C. (2014). Local field potentials reset neurons to their prefered phase during saccades.

Zemel, R.S., and Sejnowski, T.J. (1998). A model for encoding multiple object motions and self-motion in area MST of primate visual cortex. J. Neurosci. *18*, 531–547.

Zhang, K., Sereno, M.I., and Sereno, M.E. (1993). Emergence of position-independent detectors of sense of rotation and dilation with Hebbian learning: an analysis. Neural Comput. *5*, 597–612.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. Ann. Stat. *35*, 2173–2192.

A. Estimating classification images

A.1 Fitting algorithm

A.1.1 Outline

To find a MAP estimate of **w** under a sparse prior in a GLM, we need to minimize an objective function *E*:

$$\text{argmin}_{\mathbf{w}} E(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda ||\mathbf{w}||_1 \quad \text{(A-1)}$$

Here f is a convex, differentiable function of **w**. This optimization is nontrivial as $\|\mathbf{w}\|_1$ is nondifferentiable at the axes, where $w_i = 0$ for some i. The fixed-point continuation (FPC) algorithm (Hale et al., 2007) solves Equation A-1 efficiently by combining two insights. First, the fixed-point iterations:

$$\mathbf{w} \leftarrow \text{shrink}(\mathbf{w} - \tau\nabla f, \tau\lambda) \quad \text{(A-2)}$$

where *shrink* is the soft-threshold operator:

$$\text{shrink}(x, \alpha) = \text{sign}(x)\max(|x| - \alpha, 0) \quad \text{(A-3)}$$

and τ, the step size, is a small number, eventually converge to the solution of Equation A-1, with each iteration coming at a very moderate cost.

Convergence is however quite slow when iterations are started from an arbitrary value of **w**, known as a *cold-start*. The second insight is that if once we have found $\mathbf{w}^* = \text{argmin}_{\mathbf{w}} E(\mathbf{w}, \lambda_1)$, we may use $\mathbf{w}^*$ as a *warm-start* estimate for minimizing $E(\mathbf{w}, \lambda_2)$ when $|\lambda_1 - \lambda_2| / (\lambda_1 + \lambda_2)$ is small. This suggests solving the series of optimization problems:

$$\text{argmin}_{\mathbf{w}} E(\mathbf{w}, \lambda_i) \quad \text{(A-4)}$$

with $\lambda_1 > \lambda_2 > \ldots > \lambda_{\text{end}}$, using the fixed-point iterations (Equation A-2), using the result of the previous optimization to start the next optimization, a process known as *continuation*. The entire regularization path is generated as part of the process, which may be used to determine the optimal λ.

The basic form of this algorithm is elegant and is reasonably fast. Here we add three insights to form a final algorithm which, while a bit less elegant, is frequently an order of magnitude faster than this basic version. First, we add a line search over τ. Second, we use insights from (Park & Hastie, 2007) to find

good values of λ to sample the regularization path. Finally, we use a blockwise implementation of cross-validation which saves iterations when $\lambda_{optimal}$ is unknown and large.

A.1.2 Inner iterations—finding the MAP estimate for fixed λ

Here we present an elementary plausibility argument for the FPC algorithm; rigorous mathematical treatment and proofs of convergence are available in Hale et al. (2007). Consider the derivatives of the objective function E with respect to $\mathbf{w}$:

$$\frac{\partial E}{\partial w} = \nabla f + \lambda \operatorname{sign}(w) \quad \text{(A-5)}$$

Corresponding gradient descent iterations, for a step size τ take the form:

$$\mathbf{w} \leftarrow \mathbf{w} - \tau \frac{\partial \mathbf{E}}{\partial \mathbf{w}} = \mathbf{w} - \tau(\nabla f + \lambda \operatorname{sign}(\mathbf{w})) \quad \text{(A-6)}$$

Here the step size τ is some small number. Such iterations will reduce the objective function E for a sufficiently small τ as long as we avoid the discontinuities in the derivatives of the penalty. One way to avoid these discontinuities is simply to set w_i to 0 as it attempts to pass through the origin:

$$\mathbf{w} \leftarrow \operatorname{sign}(\mathbf{w})\max(\operatorname{sign}(\mathbf{w})(\mathbf{w} - \tau(\nabla f + \lambda \operatorname{sign}(\mathbf{w})), 0) \quad \text{(A-7)}$$

An issue with this new iteration is that for a weight $w_i = 0$ before the iteration, for any $\tau > 0$ the weight leaves the axis and the derivative of the penalty changes. This suggests evaluating the derivative of the penalty at $\mathbf{w} - \tau\nabla f$ instead of at $\mathbf{w}$. We thus obtain:

$$\mathbf{w} \leftarrow \operatorname{sign}(\mathbf{w} - \tau\nabla\mathbf{f})\max(\operatorname{sign}(\mathbf{w} - \tau\nabla\mathbf{f})(\mathbf{w} - \tau(\nabla f + \lambda \operatorname{sign}(\mathbf{w} - \tau\nabla\mathbf{f})), 0)$$

$$= \operatorname{shrink}(\mathbf{w} - \tau\nabla f, \tau\lambda) \quad \text{(A-8)}$$

Thus, the fixed-point iterations can be viewed as gradient descent iterations with special considerations to avoid issues with the discontinuities at the axes. A thorough analysis shows that these iterations indeed converge to the minimum of $E(\mathbf{w}, \lambda)$ (Hale et al., 2007).

A.1.3 Line search

Fixed-point iterations of Equation A2 are guaranteed to converge for fixed τ as long as $\tau < \tau_{max} = 2/\max_i \mathbf{\Lambda}_{ii}$, where $\mathbf{Q\Lambda Q}^T = \mathbf{H}$ is an eigendecomposition of the matrix of second derivatives of the negative log-likelihood with respect to the weights, the Hessian $\mathbf{H}$. Note that this is the same condition as in gradient

descent. The fixed-point iteration's relationship with gradient descent hints that convergence may be substantially faster if τ is chosen optimally on every iteration such that:

$$\tau_{min} = \text{argmin}_{\tau > 0} E(\text{shrink}(\mathbf{w}^0 - \tau\mathbf{g}, \tau\lambda)) \quad \text{(A-9)}$$

Over a series of informal experiments, we have noticed that performing a line search can reduce the number of inner iterations required for convergence by an order of magnitude or more over an optimal, fixed τ. Unfortunately, a naive line-search over τ involves multiple products of the form **Xw** and repeated computations of *E,* which are expensive; hence, overall using a naive line search does not yield a large performance improvement over a fixed τ. Here we propose a line search algorithm which, although rather involved, is much more efficient than a naive line search.

First notice that $E(\tau)$ has a peculiar form: its derivatives of all orders with respect to τ are continuous outside of a finite number of "cutpoints" $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_n$ which occur when a weight enters or leaves the model. Thus, we approximate $E(\tau)$ as a piecewise quadratic, convex function, whose minimum is inexpensively found.

Let $\boldsymbol{\eta}(\mathbf{w}) \equiv \mathbf{Xw} + \mathbf{Uu}$ and $\mathbf{w}(\alpha) = \text{shrink}(\mathbf{w}^0 - \alpha\,\mathbf{g}, \alpha\lambda)$. We may Taylor-expand *f* as a function of α up to second order, which by the chain rule gives:

$$f(\alpha) \approx f(0) - \alpha \frac{\partial f}{\partial \eta} \cdot \frac{\partial \eta}{\partial \alpha} - \frac{1}{2}\alpha^2 \left(\frac{\partial f}{\partial \eta} \cdot \frac{\partial^2 \eta}{\partial \alpha^2} + \frac{\partial^2 f}{\partial n^2} \cdot \left(\frac{\partial \eta}{\partial \alpha} \right)^2 \right) \quad \text{(A-10)}$$

Here $(\cdot)$ denotes the dot product. All the derivatives are computed at $\alpha = \varepsilon$, $\varepsilon > 0$ arbitrarily small, to give right-sided derivatives at $\alpha = 0$. This approximation is valid between $0 < \alpha < \alpha_{\max}$ assuming no weights enter or leave the model in this range. The derivatives of *f* with respect to $\boldsymbol{\eta}$ are straightforward to compute; they also occur in the iteratively reweighted least-squares (IRLS) algorithm often used to fit GLMs (Wood, 2006). The derivatives of $\boldsymbol{\eta}$ are given by

$$\left.\frac{\partial \eta}{\partial \alpha}\right|_{\epsilon} = \mathbf{X} \left.\frac{\partial \mathrm{w}}{\partial \alpha}\right|_{\epsilon} = \mathbf{X}\big((-\mathbf{g} - \lambda\,\text{sign}(\mathbf{w}^0 - \epsilon\mathbf{g})) \cdot (\mathbf{w}(\epsilon) \neq 0)\big)$$

$$\left.\frac{\partial^2 \eta}{\partial \alpha^2}\right|_{\epsilon} = 0 \quad \text{(A-11)}$$

Here $(\mathbf{w}(\varepsilon) \neq 0)$ is a vector with value 1 if $\mathbf{w}_i(\varepsilon) \neq 0$ and 0 otherwise. Similarly, the regularizer $R = \lambda\|\mathbf{w}\|_1$ may be Taylor expanded between $0 < \alpha < \alpha_{\max}$ to give:

$$R(\alpha) = R(0) + \alpha \frac{\partial R}{\partial \alpha} \epsilon$$
$$= R(0) + \alpha\lambda \sum_i (-\mathbf{g} \sin(\mathbf{w}^0 - \epsilon\mathbf{g}) - \lambda) \cdot (\mathbf{w}(\epsilon) \neq 0) \quad \text{(A-12)}$$

Note that this last expansion is exact. Thus, $E(\alpha) = f(\alpha) + R(\alpha) = A + B\alpha + \frac{1}{2}C\alpha^2$ is quadratic in α and it has an extremum at $\alpha_{min} = -B/C$. Now:

$$C = \frac{\partial^2 f}{\partial \eta^2} \cdot \left(\frac{\partial \eta}{\partial \alpha}\right)^2 \quad \text{(A-13)}$$

For GLMs, $\frac{\partial^2 f}{\partial \eta^2} \geq 0$ (Wood, 2006); hence, $C > 0$ and E has a minimum at $\alpha_{min} = -B / C$. Thus, either $0 < \alpha_{min} < \alpha_{max}$, in which case we have found a minimum for E, or not, in which case the minimum of E is located elsewhere and the approximations are no longer valid.

Given the second-order Taylor expansion of E, it follows that E is approximately piecewise quadratic, continuous, and convex away from the cutpoints. At a cutpoint τ_i, $E(\tau_i)$ has a local maximum if and only if the left-sided derivative of $E(\tau_i)$ is positive and the right sided derivative is negative; it is straightforward to show that this is never the case. We conclude that $E(\tau)$ is piecewise quadratic, continuous, and convex, and hence it has a single minimum in the range $0 < \tau < \infty$. This suggests attempting to find a minimum of E between $0 < \tau < \tau_1$. If the minimum is not in that range, we search for the minimum in the range $\tau_1 < \tau < \tau_2$ using a new Taylor expansion at $E(\mathbf{w}(\tau_1))$, and so forth until we find the minimum of E. This gives Algorithm A-1 presented in section A-3.

Notice that the iterations for $i \geq 2$ are rather inexpensive, save for computing $\frac{\partial \eta}{\partial \alpha}$. However, at a cutpoint, $\frac{\partial \mathbf{w}_j}{\partial \alpha}$ changes for exactly one j, corresponding to the weight which enters or leaves the model. This implies that $\frac{\partial \eta}{\partial \alpha}$ changes by a multiple of a single column of $\mathbf{X}$ at cutpoints, and hence $\frac{\partial \eta}{\partial \alpha}$ is updated at very little cost.

The expensive computations in the initial iteration are that of f, $\frac{\partial f}{\partial \eta}$, $\frac{\partial \eta}{\partial \alpha}$, $\frac{\partial^2 f}{\partial \eta}$, $\frac{\partial R}{\partial \alpha}$. $\boldsymbol{\eta}$ is saved from the previous line search, which avoids computing a product of the form $\mathbf{Xw}$. To compute f and its derivatives, roughly N logarithms, N exponentials, and a few element-wise products of vectors of length N must be performed, where N is the number of trials. For $\frac{\partial \eta}{\partial \alpha}$, one product $\mathbf{X}\frac{\partial \mathbf{w}}{\partial \alpha}$ is formed; since the right-hand side is a vector which contains mostly zeroes, this is rather inexpensive. Thus, the proposed

line search algorithm, while rather involved, avoids computing **Xw** and f repeatedly, and thus is much less expensive than a naive line search.

A.1.4 Outer iterations

The FPC algorithm works by solving Equation A1 by iterative thresholding for decreasing values of λ. After each set of inner iterations, **u** is reoptimized and a smaller value of λ is chosen. The next value of λ is set to the largest λ for which a new weight should appear in the model:

$$\tilde{\lambda}_{new} = \max_{i \in \{j \mid wj = 0\}} |(\nabla f(\mathbf{w}_{est}))_i| \quad \text{(A-14)}$$

Simulations show that the optimization can stall when $\tilde{\lambda}_{new}$ is continually chosen to be very close to the current λ, and that the estimate is inaccurate when $\tilde{\lambda}_{new}$ is far from the current λ. We thus bracket $\tilde{\lambda}_{new}$ to be neither too close nor too far from $\lambda_{current}$:

$$\lambda_{new} = \begin{cases} \alpha_{min}\lambda_{curr} \; when \; \tilde{\lambda}_{new} < \alpha_{min}\lambda_{curr} \\ \tilde{\lambda}_{new} \\ \alpha_{max}\lambda_{curr} \; when \; \tilde{\lambda}_{new} > \alpha_{max}\lambda_{curr} \end{cases} \quad \text{(A-15)}$$

We found that using a nonuniform step size is more efficient than any one fixed step size, and used $\alpha_{min} = 0.9$, $\alpha_{max} = 0.98$.

A.1.5 Initial and final iterations

At the very start of the optimization, **w** is set to 0, and **u** is optimized. ∇f is computed, and λ_{start} is set to be slightly smaller than the first value of λ for which a weight should appear in the model:

$$\lambda_{start} = \max_i |\nabla_i f| \quad \text{(A-16)}$$

Outer iterations are stopped when $\lambda_{next} < \varepsilon\lambda_{start}$ for some user-defined value of ε, which is set, by default, to 10^{-3}.

A.1.6 Cross-validation

The number of λ values considered should be minimized within reason to obtain an efficient algorithm. To ensure that fitting is done over a tight range of λ values, we implemented an efficient version of k-fold cross-validation, which we call blockwise cross-validation. Rather than fitting each fold from $10^{-3}\lambda_{start} < \lambda < \lambda_{start}$ in turn, we first fit the model for all folds from $0.7\lambda_{start} < \lambda < \lambda_{start}$ and compute the cross-validated deviance. If a minimum is found, we stop. If not, we restart the fitting process from $0.5\lambda_{start} < \lambda < 0.7\lambda_{start}$, and so on until either $10^{-3}\lambda_{start}$ is reached or a minimum of the cross-validated deviance is

found. All information related to fitting is saved in a structure after each block; hence, restarting an optimization has little overhead. We determine that a minimum has been found by asking that the cross-validated deviance at the minimum λ probed so far is higher than the minimum cross-validated deviance by a certain number of units, by default 10. Finally, we fit the full model down to the minimum λ found by cross-validation.

Our algorithm succeeds in finding the first non-shallow local minimum $D^{CV}(\lambda)$ between λ_{start} and $10^{-3}\lambda_{start}$. While we cannot rule out the possibility that $D^{CV}(\lambda)$ has several non-shallow local minima, this does not appear to be an issue in practice; $D^{CV}(\lambda)$ is typically quite smooth and almost quadratic in shape near its minimum, and we have not observed non-shallow local minima in any of the fits performed for the purposes of this article. In the pathological scenario where $D^{CV}(\lambda)$ has several non-shallow local minima, the cross-validation algorithm will select the one associated with the largest λ, or the largest level of regularization, which we consider to be a safe fallback.

Of crucial importance in cross-validation is that different values of λ across different folds correspond to the same level of regularization. Tibshirani (1996) and Park and Hastie (2007) suggest that corresponding regularization levels are found when the ratio $\lambda/\lambda_{max}^{fold}$ is the same across folds. We thus perform cross-validation to find the optimal ratio $\lambda/\lambda_{max}^{fold}$ rather than λ. The regularization paths are sampled differently across each fold. We solved this issue by linearly interpolating cross-validated deviance values at all $\lambda/\lambda_{max}^{fold}$ used by a fold.

A.1.7 Complexity analysis and memory and time requirements

Computing ∇f, a $\mathcal{O}(mn)$ operation, typically accounts for 50%–90% of the time spent during optimization; roughly 5% is accounted for by miscellaneous overhead; and the remainder time is spent in the line search. The ratio of inner iterations to outer iterations varies with λ, being typically equal to 1 for large λ and 2–4 for small λ. Given the algorithm for determining the next λ and assuming we evaluate the model at λ values from $0.001\lambda_{start}$ to λ_{start}, the number of outer iterations is bounded above by $\log(0.001)/\log(0.98) \approx 340$. Given typical ratios of inner iterations to outer iterations, this might come out to about 600 evaluations of ∇f. This number can be cut down by a factor 2 or 3 by stopping iterations when λ reaches beyond $\lambda_{optimal}$ determined by cross-validation; our software includes a cross-validation implementation, outlined above, that accomplishes this. In total, perhaps 300 evaluations of ∇f may be required in a typical application; with overhead this may balloon up to a cost equivalent to computing ∇f 500–600 times, multiplied by $(k + 1)$ during k-fold cross-validation. This compares favorably and in some cases may be appreciably faster than boosting where the cost of each iteration is

equal to the cost of computing ∇f, in addition to some overhead such as computing f and performing a line search in some variants (Buhlmann & Hothorn, 2007).

Memory requirements are typically equal to the cost of holding two or three copies of the design matrix **X** in memory. With a 64-bit version of Windows or Linux and 4 GBs of RAM, one can thus expect to able work with design matrices as large as 25,000 × 5,000 (1 GB in memory) before running into out-of-memory errors. For the largest design matrices tested here, which are of size 10,000 × 256, optimization including 5-fold cross-validation takes about 40 seconds on our test computer running on 4 Intel Xeon CPUs running at 2 Ghz and 3 GB of RAM.

A.1.8 Software

Our software package, implemented in Matlab, includes two main functions, with function signatures:

```
[thefit] = glmfitsparseprior(y,X,U,stopcrit,varargin)
[thefit] = cvglmfitsparseprior(y,X,U,folds,varargin)
```
(A-17)

Here **y**, **X**, and **U** are response and design matrices and `stopcrit` is the fraction of λ_{start} after which to stop optimization. `folds` is a matrix of booleans which defines which portion of the data is assigned to the fit set versus the validation set for each cross-validation fold. Such a matrix may be generated by the included auxiliary function `getcvfolds`. A structure is returned in both cases which contains the entire regularization path, deviances, AIC values, the values of λ used, and, if applicable, cross-validated deviances and **w** and **u** at the optimal value of λ. In addition to the binomial/logit model presented here, the software supports two other GLMs: a Poisson model with exponential inverse link, for count data, e.g., binned spike trains, and a Gaussian noise model with identity link, e.g., least-squares. These latter may be invoked through supplementary arguments whose calling sequence is available from the Matlab command-line help.

A.2 Inference for the GLM

Goodness-of-fit is assessed in GLMs through the scaled deviance of a fitted model (Wood, 2006):

$$D^* = 2(L - L_{max}) \quad \text{(A-18)}$$

The deviance *proper* is defined as $D = D^*\phi$, where ϕ is a scale parameter that depends on the noise distribution used. For the binomial distribution, $\phi = 1$ (Wood, 2006); hence, we use the terms deviance and scaled deviance interchangeably. L is the negative log-likelihood of the fitted model, while L_{max} is the

negative log-likelihood for a saturated model which contains one parameter per data point y_i. For the logistic regression model used here, $L_{max} = 0$, although this is not always the case, for example with Poisson regression. By construction, $D \geq 0$; a small D implies a good fit to the data while a large D implies a poor fit.

Standard linear regression with Gaussian residuals can be cast as a special case of a GLM (Wood, 2006), for which the deviance is given by

$$D = \sum_{ij}(y_i - X_{ij})^2 \quad \text{(A-19)}$$

Here $\widehat{w}$ is the estimated weight vector. This expression comes from the fact that the negative log of a Gaussian is a sum-of-squares. It is thus helpful to think of deviance as analogous to the residual the sum-of-squares in linear regression. Like the residual sum-of-squares, the deviance always becomes smaller as predictors are added to the design matrix; hence, it is not useful by itself for model selection purposes. Rather, the deviance is used as a basis for model-complexity independent measures of goodness-of-fit, such as the cross-validated deviance, the validated deviance, and the Akaike Information Criterion. The deviance may also be used for classical hypothesis testing in log-likelihood ratio tests.

As explained in the Methods section, *k*-fold cross-validation works by splitting the data into *k* randomly chosen nonoverlapping partitions of equal size, fitting the model with data in all but the partition and computing the likelihood or deviance of the data in the *i*th partition, and repeating the process for $i = 1 \ldots k$. This yields a cross-validated deviance score D^{CV}. D^{CV} is a measure of how well a model predicts out-of-sample observations and therefore estimates its generalization performance.

How do we interpret D^{CV} measured for a model $\mathcal{M}_j$? By Bayes' theorem, the probability of the model given the data **y** is given by

$$p(\mathcal{M}_j|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_j)p(\mathcal{M}_j) \quad \text{(A-20)}$$

Assuming a flat prior on models $p(\mathcal{M}_j) = 1$ for all *j,* we have:

$$p(\mathcal{M}_j|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_j) \quad \text{(A-21)}$$

The probability of a data set **y** under the model is approximately given by cross-validated likelihood of the data (Geisser, 1975; Geisser & Eddy, 1979). Hence, the relative likelihood of two models is approximately given by

$$PSBF(\mathcal{M}_1,\ \mathcal{M}_2)\ = \frac{p(\mathcal{M}_1|\mathbf{y})}{p(\mathcal{M}_2|\mathbf{y})} \approx \exp\left(0.5\left(D_2^{CV} - D_1^{CV}\right)\right) \quad \text{(A-22)}$$

This has the same form as a Bayes factor, hence the name Pseudo-Bayes factor (PSBF) (Gelfand & Dey, 1994). Bayes factors are a common technique of model comparison. An often-used interpretation scale for Bayes factors states that *PSBF* ($\mathcal{M}_1$, $\mathcal{M}_2$) > 150, corresponding to $D_2^{CV} - D_1^{CV} > 10$ is "very strong" (Kass & Raftery, 1995) or "decisive" (Jeffreys, 1961) evidence for $\mathcal{M}_1$ over $\mathcal{M}_2$.

The variance of D^{CV} can be large; its value depends on the exact choice of folds. Therefore, when differences in the cross-validated deviances between two models are relatively small, for example less than 10 units, we do not recommend concluding that either model is better based on these scores; lower variance estimates of D^{CV} can be obtained by repeatedly computing D^{CV} for different folds.

In the same spirit, the deviance of a model based on predictions on a hold-out validation set can be useful to assess model performance. However, since binomial data are quite noisy, much validation data must be collected to obtain reliable results.

The effective degrees of freedom for the MAP estimate of a GLM with a sparse prior is defined as (Park & Hastie, 2007; Hastie, 2007; Zou, Hastie, & Tibshirani, 2007):

$$df = |\{k|\mathbf{w}_k \neq 0\}| \quad \text{(A-23)}$$

That is, it is equal to the number of nonzero elements in the estimated weight vector **w**. For a model with a Gaussian prior, on the other hand (Wood, 2006):

$$df = tr((\mathbf{H} + \lambda\mathbf{A}^T\mathbf{A})^{-1}\mathbf{H}) \quad \text{(A-23)}$$

Here **H** denotes the Hessian of the negative log-likelihood and *tr* the trace. Note that this reduces to *n,* the number of parameters in the model, as $\lambda \rightarrow 0$. Given *df* for the GLM with either a sparse or Gaussian prior, we may define an AIC (Akaike Information Criterion) analogue (Wood, 2006):

$$AIC = D + 2df \quad \text{(A-24)}$$

For the purposes of hyperparameter selection in a sparse GLM, we do not recommend the use of the AIC, as its integer-valued nature means AIC(λ) is discontinuous and has several shallow minima. However, it may be useful in comparing our results with other GLMs with binomial endpoints that, for practical reasons, avoid cross-validation (Knoblauch & Maloney, 2008b; Ross & Cohen, 2009). We note that the AIC is equivalent, up to a linear transformation, to the unbiased risk estimator (UBRE) used in Knoblauch and Maloney (2008b). As with cross-validated deviance, a lower AIC is better, and large differences in AIC values between two models are interpreted as support for the model with the lower AIC. Since the AIC is based on asymptotic results, its validity is dubious when low numbers of trials are used. In the case where the AIC and the cross-validated deviance point towards incompatible conclusions, we recommend averaging cross-validated deviance over several different random choices of folds and base conclusions on this low-variance cross-validated deviance estimate.

It has been shown (Wood, 2006) that for a simple GLM model $\mathcal{M}_1$ nested inside a more complex GLM model $\mathcal{M}_2$:

$$D_1 - D_2 \sim \chi^2_{df2-df1} \quad \text{(A-25)}$$

A *p*-value may be obtained from this expression. This is known as a log-likelihood ratio test. Again, the χ^2 approximation is based on asymptotic results and has dubious validity for a small number of trials (Wood, 2006). It is also important to keep in mind that the test is only valid for *nested* models. We recommend the use of log-likelihood ratio tests when the importance of a single predictor or ensemble of related predictors are in question. For nonnested models, Vuong's test may be used (Vuong, 1989).

A.3 Algorithms

- Compute the ordered set of cutpoints of $E(\tau)$, giving $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_{n-1} < \tau_n = \infty$
- For i from 1 to n
  - If i = 1
    - Compute $\boldsymbol{\eta}$, $f, \frac{\partial f}{\partial \boldsymbol{\eta}}, \frac{\partial^2 f}{\partial \boldsymbol{\eta}}, \frac{\partial \boldsymbol{\eta}}{\partial \alpha}, \frac{\partial R}{\partial \alpha}$ evaluated at $\alpha = \epsilon$
  - Else
    - Let $\Delta\alpha \leftarrow (\tau_{i-1} - \tau_{i-2})$
    - Update $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \Delta\alpha \frac{\partial \boldsymbol{\eta}}{\partial \alpha}$
    - Update $f \leftarrow f + \Delta\alpha \frac{\partial f}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \alpha} + \frac{1}{2}\Delta\alpha^2 \frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \cdot (\frac{\partial \boldsymbol{\eta}}{\partial \alpha})^2$
    - Update $\frac{\partial f}{\partial \boldsymbol{\eta}} \leftarrow \frac{\partial f}{\partial \boldsymbol{\eta}} + \Delta\alpha \frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \frac{\partial \boldsymbol{\eta}}{\partial \alpha}$
    - Update $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}, \frac{\partial R}{\partial \alpha}$
  - End If
  - Compute α_{min}
  - If $0 < \alpha_{min} < (\tau_i - \tau_{i-1})$
    - $\tau_{min} = \tau_{i-1} + \alpha_{min}$; Exit
  - Elseif $\alpha_{min} < 0$
    - $\tau_{min} = \tau_{i-1}$; Exit
  - End If
- End For

Algorithm A-1: Line search for τ_{min}

B. Estimating local field potentials

This section demonstrates a method to estimate spike-free local field potentials by inferring the parameters of a generative model for the wideband signal. This section is excerpted from the supplemental information of Zanos et al. (2011).

B.1 Model and parameter estimation

Our goal is to estimate the local field potential (LFP) based on a measured wideband voltage trace of length n. We assume that this wideband signal $\mathbf{y}$ is the superposition of a low-frequency local field potential $\mathbf{w}$, high-frequency spike components $\boldsymbol{\eta}^k$, an offset μ and white noise $\boldsymbol{\epsilon}$:

$$\mathbf{y} = \mathbf{w} + \sum_{k=1}^{m} \boldsymbol{\eta}^k + \mu + \boldsymbol{\epsilon} \tag{B-1}$$

Here m is the number of sorted neurons emitting spikes. The high-frequency component of the k^{th} neuron, $\boldsymbol{\eta}^k$, is created by the convolution of the neuron's spike train $\mathbf{s}^k$, assumed known, and the neuron's spike waveform $\mathbf{B}\boldsymbol{\varphi}^k$:

$$\boldsymbol{\eta}^k = C'(\mathbf{s}^k)\mathbf{B}\boldsymbol{\varphi}^k \tag{B-2}$$

Here the $C'(\mathbf{a})$ returns a circulant matrix whose first row is $\mathbf{a}$. The product $C'(\mathbf{a})\mathbf{b}$ returns the circular convolution of $\mathbf{a}$ and $\mathbf{b}$:

$$C'(\mathbf{a})\mathbf{b} = (a \circledast b)_i = \sum_j a_{[i-j]} b_j \tag{B-3}$$

Here $a_{[k]} = a_{(k \text{ modulo } n)+1}$. Circulant matrices have a number of properties that are crucial for the tractability of the model parameters; their properties are covered in detail in the last section. We call $\boldsymbol{\eta}^k$ the *waveform train* of the k^{th} neuron.

$\mathbf{B}$ is a matrix of basis functions which map the spike parameters $\boldsymbol{\varphi}^k$ onto a spike waveform. Typically, the number of parameters that describe the spike waveforms is much smaller than the length of the signal, which implies that $\mathbf{B}$ is much higher than it is large.

Assumptions are as follows:

$$p(\mathbf{w}) = N(0, \gamma^2\mathbf{\Gamma})$$
$$p(\boldsymbol{\epsilon}) = N(0, \sigma^2\mathbf{I}) \tag{B-40}$$

Here $N(\mathbf{a}, \mathbf{\Sigma})$ represents a multivariate Gaussian with mean $\mathbf{a}$ and covariance $\mathbf{\Sigma}$. $\mathbf{\Gamma} = C'(F^{-1}(\mathbf{g}))$ is a matrix that embodies an assumption of smoothness, and $\mathbf{\Gamma x}$ produces a low-pass filtered version of $\mathbf{x}$.

By Bayes' theorem, we have that the posterior probability of the parameters is, up to an additive constant:

$$-\log p(\mathbf{w}, \boldsymbol{\varphi}^k, \boldsymbol{\mu}|\mathbf{y})$$
$$= \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{w} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\boldsymbol{\varphi}^k - \mu\right)'\left(\mathbf{y} - \mathbf{w} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\boldsymbol{\varphi}^k - \mu\right) + \frac{1}{2\gamma^2}\mathbf{w}'\Gamma^{-1}\mathbf{w} \tag{1}$$

We can solve for the MAP estimate of the parameters by taking partial derivatives of the posterior and setting derivatives to zero. The MAP estimates are given by:

$$\bar{\mathbf{w}} = (\gamma^2\mathbf{\Gamma} + \sigma^2\mathbf{I})^{-1}\gamma^2\mathbf{\Gamma}\left(\mathbf{y} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\bar{\boldsymbol{\varphi}}^k - \bar{\mu}\right)$$
$$\bar{\boldsymbol{\varphi}}^k = (C'(\mathbf{s}^k)\mathbf{B})^+\left(\mathbf{y} - \bar{\mathbf{w}} - \sum_{j\neq k} C'(\mathbf{s}^j)\mathbf{B}\bar{\boldsymbol{\varphi}}^j - \bar{\mu}\right) \tag{B-6}$$
$$\bar{\mu} = \frac{1}{n}\hat{\mathbf{1}}'\left(\mathbf{y} - \bar{\mathbf{w}} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\bar{\boldsymbol{\varphi}}^k\right)$$

Here $\mathbf{A}^+$ is the pseudoinverse $\mathbf{A}^+ = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, and $\hat{\mathbf{1}}$ is a vector of ones. Although it is possible to solve this system by taking initial guesses of the model parameters and computing the parameters using these equations in sequence and iterating, in practice we found convergence to be too slow for this to be a practical solution. Instead we isolate each variable individually, starting with $\bar{\boldsymbol{\varphi}}^k$. Because most spike removal algorithms do not deal with overlapping spikes, the waveform trains $\boldsymbol{\eta}^k$ have little overlap, and hence we isolate $\bar{\boldsymbol{\varphi}}^k$ on the assumption that the other waveforms are known. We thus define $\mathbf{v} = \mathbf{y} - \sum_{j\neq k} C'(\mathbf{s}^j)\mathbf{B}\bar{\boldsymbol{\varphi}}^j$ and drop the indices in the equations to lighten the notation:

$$\bar{\mathbf{w}} = \mathbf{M}(\mathbf{v} - \mathbf{D}\bar{\boldsymbol{\varphi}} - \bar{\mu})$$
$$\bar{\boldsymbol{\varphi}} = \mathbf{D}^+(\mathbf{v} - \bar{\mathbf{w}} - \bar{\mu}) \tag{B-7}$$

$$\bar{\mu} = \frac{1}{n}\widehat{\mathbf{1}}'(\mathbf{v} - \bar{\mathbf{w}} - \mathbf{D}\bar{\boldsymbol{\varphi}})$$

Here $\mathbf{D} \equiv C'(\mathbf{s})\mathbf{B}$ and $\mathbf{M} = (\gamma^2\boldsymbol{\Gamma} + \sigma^2\mathbf{I})^{-1}\gamma^2\boldsymbol{\Gamma}$, a low-pass filter. $\mathbf{D}'\mathbf{x}$ computes the spike-triggered sum of the signal $\mathbf{x}$ while $\mathbf{D}\mathbf{x}$ computes the waveform train associated with the spike waveform $\mathbf{x}$. By substituting $\bar{\mu}$ into the first two equations, then isolating $\bar{\mathbf{w}}$ in the first equation and finally substituting it in the second equation, we obtain the normal equation:

$$\mathbf{D}'\mathbf{J}(\mathbf{I} - \mathbf{M})\mathbf{D}\bar{\boldsymbol{\varphi}} = \mathbf{D}'\mathbf{J}(\mathbf{I} - \mathbf{M})\mathbf{v} \tag{B-8}$$

Here $\mathbf{J} = \mathbf{I} - \frac{1}{n}\widehat{\mathbf{1}}\widehat{\mathbf{1}}'$ is a centering matrix which when applied to a vector yields the same vector but with its mean set to 0. This equation can be interpreted as follows: the STA ($\mathbf{D}'$) of the centered ($\mathbf{J}$) high-pass filtered ($\mathbf{I} - \mathbf{M}$) waveform train ($\mathbf{D}\bar{\boldsymbol{\varphi}}$) is equal to the STA of the centered high-pass filtered wideband signal.

While this equation is satisfyingly compact and intuitive, multiplication by $\mathbf{M}$ must be done through multiplication in the Fourier domain, and hence a Fourier and an inverse Fourier transform need to be performed for every column of $\mathbf{D}$; this operation is thus a bottleneck. Since spikes are assumed to be of finite duration, however, it follows that we can write $= [\mathbf{B}_s; \mathbf{0}]$, where $\mathbf{0}$ is a matrix of zeros of size $(n - q)$ by p, where $q \ll n$ is the length of a spike and $p \leq q$ is the number of basis functions. Note that this requires time shifting the spike train $\mathbf{s}$ so that an entry equal to 1 indicates the beginning of a spike rather than its peak. Substituting and simplifying (B-8), we find:

$$\mathbf{D}'\mathbf{J}(\mathbf{I} - \mathbf{M})\mathbf{D} = \mathbf{B}_s'(\mathbf{k} * \mathbf{B}_s) - \frac{1}{\mathrm{n}}(\mathbf{s}'\widehat{\mathbf{1}})^2\left(F^{-1}\left(\frac{\sigma^2}{\sigma^2 + \gamma^2\mathbf{g}}\right)'\widehat{\mathbf{1}}\right)(\mathbf{B}_s'\widehat{\mathbf{1}})(\mathbf{B}_s'\widehat{\mathbf{1}})' \tag{B-9}$$

Here $*$ denotes linear convolution with zero-padding, applied column-wise, and $\mathbf{k}$, the convolution kernel, is derived from $\mathbf{a} = \mathrm{F}^{-1}\left(\frac{\sigma^2}{\sigma^2+\gamma^2\mathbf{g}}|\mathrm{F}(\mathbf{s})|^2\right)$ by circular time shifting, with $k_i = a_{[i-q]}$. Since q is small, the convolution in (9) is inexpensive and the equation can be solved by preconditioned conjugate gradients at negligible cost. After solving equation (9) sequentially for every neuron, we plug these estimates in the system of equations (6) and solve to find:

$$\bar{\mu} = \frac{1}{n}\hat{\mathbf{1}}' \left(\mathbf{y} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\bar{\boldsymbol{\varphi}}^k \right) \tag{B-10}$$

Thus $\bar{\mu}$ is equal to the mean of the wideband signal minus the mean of the spike contribution. Finally, the LFP $\bar{\mathbf{w}}$ is given by the first equation in the system (6); it is the low-pass filtered wideband signal minus the spike contribution. In practice, the experimenter will probably want to use his or her own filter or filterbank on the despiked wideband signal to obtain the LFP. Hence $\bar{\mathbf{w}}$ **is** never actually computed by the despiking algorithm; instead the algorithm works with and returns the despiked wideband signal, defined as:

$$\bar{\mathbf{z}} = \mathbf{y} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\bar{\boldsymbol{\varphi}}^k - \bar{\mu} \tag{B-11}$$

Because we ignored cross terms when solving for the spike waveforms $\bar{\boldsymbol{\varphi}}^k$, and also because of possible numerical instability, solutions must be checked for convergence. It can be verified using equation (6) that for every k, the following auxiliary equation holds:

$$\mathbf{B}_s[\mathbf{B}_s; \mathbf{0}]' C(\mathbf{s}^k)\mathbf{J}(\mathbf{I} - \mathbf{M}) \left(\mathbf{y} - \sum_j C'(\mathbf{s}^k)\mathbf{B}\bar{\boldsymbol{\varphi}}^j - \mu \right) = \mathbf{0} \tag{B-12}$$

This states that the STA of a high-pass filtered, centered despiked wideband signal projected onto the basis in which we express spike waveforms is 0. Convergence is attained when the largest deviation from 0 observed is smaller than some fraction of the standard deviation of $\mathbf{y}$. When convergence is not attained, the process is repeated. In practice, it rarely takes more than 2 iterations to reach convergence.

Figure B-1A shows an example of a wideband signal before and after spike removal. The two signals have been offset vertically to facilitate comparison. By construction, the proposed method removes only the mean spike waveform around the time of every spike, and hence does not remove all traces of spikes when the spike waveform is variable. It also relies on the information given by the experimenter about the timing of spikes, and does not remove spikes which have not been detected. These limitations can be overcome in large part with good spike detection, alignment, and sorting. In these cases the method performs admirably, as seen in this figure, and results in an appreciable change in the wideband

signal. The changes are much less conspicuous when looking at the LFP, shown in Figure B-1B. The difference between the two signals consists of a barely visible artifact around the time of every spike. Although the artifact is very small for each spike, it is highly stereotyped: it always occurs at the same time relative to spikes, and always has the same shape, sign and relative phase. Thus, any technique that looks at temporal relationships between spikes and LFPs will amplify the artifact

B.2 Hyperparameter estimation

Recall that our model assumptions are that:

$$\begin{gathered} p(\mathbf{w}) = N(0, \gamma^2 \mathbf{\Gamma});\ \mathbf{\Gamma} = C'(F^{-1}(\mathbf{g})) \\ p(\boldsymbol{\epsilon}) = N(0, \sigma^2 \mathbf{I}) \end{gathered} \tag{B-13}$$

Up to now, we have assumed that σ and γ are known. The strength of the prior relative to the noise is important as it determines what the model considers as signal and what it discounts as noise. These hyperparameters, which control the regularization of the model, cannot be determined by MAP, unlike regular parameters (Wu et al. 2006). Here we determine these parameters by optimizing the marginal likelihood of the model, a metric which takes into account both the quality of the model fit to the data and the number of degrees of freedom in the model to determine the optimal degree of regularization. This method is also known as evidence optimization (see chapter 3 of Bishop 2006 for a detailed introduction to this subject). We begin by ignoring the uncertainty in the model due to the parameters of the spike. $\mathbf{z}$ is defined as before as the despiked wideband signal $\mathbf{z} = \mathbf{y} - \sum_k C'(\mathbf{s}^k)\mathbf{B}\bar{\boldsymbol{\varphi}}^k - \bar{\mu}$. $\bar{\boldsymbol{\varphi}}^k$ and $\bar{\mu}$ are assumed to have both been estimated according to the methods of the Model and Parameter Estimation section. The marginal likelihood of the model is defined as:

$$p(\mathbf{z}|\sigma, \gamma, \mathbf{g}) = \int p(\mathbf{z}|\mathbf{w}, \sigma) p(\mathbf{w}|\gamma, \mathbf{g}) d\mathbf{w} \tag{B-14}$$

The marginal likelihood is thus the likelihood of the data (and therefore the model) with the uncertainty in the model parameters $\mathbf{w}$ marginalized out by integration. Unlike in MAP estimation, normalization constants that ensure that probabilities integrate to 1 are of crucial importance and thus are not ignored. The marginal likelihood is then:

$$p(\mathbf{z}|\sigma, \gamma, \mathbf{g}) = \frac{1}{\sqrt{2\pi|\sigma^2|}} \frac{1}{\sqrt{2\pi|\gamma^2 \mathbf{\Gamma}|}} \int \exp\left(-\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{w})'(\mathbf{z}-\mathbf{w}) - \frac{1}{2\gamma^2}\mathbf{w}'\mathbf{\Gamma}^{-1}\mathbf{w}\right) d\mathbf{w} \tag{B-15}$$

Here $|\mathbf{M}|$ is the determinant of the matrix $\mathbf{M}$. The integral is performed by completing the square inside the exponential. Taking the negative log of the integral and grouping terms which are independent of the hyperparameters into a constant term k (compare eq. 3.86 in Bishop 2006), we find:

$$\begin{aligned} -\log p(\mathbf{z}|\sigma,\gamma,\mathbf{g}) &= k + n\log\sigma + n\log\gamma + \frac{1}{2}\log|\mathbf{I}| + \frac{1}{2}\log|\mathbf{\Gamma}| + \frac{1}{2}\log|\sigma^{-2}\mathbf{I}^{-1} + \gamma^{-2}\mathbf{\Gamma}^{-1}| \\ &+ \frac{1}{2\sigma^2}(\mathbf{z}-\bar{\mathbf{w}})'(\mathbf{z}-\bar{\mathbf{w}}) + \frac{1}{2\gamma^2}\bar{\mathbf{w}}'\mathbf{\Gamma}^{-1}\bar{\mathbf{w}} \\ &= k + \frac{1}{2}\log|\sigma^2\mathbf{I} + \gamma^2\mathbf{\Gamma}| + \frac{1}{2\sigma^2}(\mathbf{z}-\bar{\mathbf{w}})'(\mathbf{z}-\bar{\mathbf{w}}) + \frac{1}{2\gamma^2}\bar{\mathbf{w}}'\mathbf{\Gamma}^{-1}\bar{\mathbf{w}} \end{aligned} \tag{B-16}$$

The leading term measures the model complexity, while the trailing terms measure the misfit of the model to the data; the optimal hyperparameters strike the best balance between model fit and complexity by minimizing their sum. To find optimal hyperparameters, the negative log marginal likelihood is minimized numerically.

Here circulant matrices are particularly useful in two ways. First, the normally problematic log determinant appearing in the marginal likelihood has the special form $\log|\sigma^2\mathbf{I} + \gamma^2\mathbf{\Gamma}| = \sum_i \log(\sigma^2 + \gamma^2 g_i)$ (see Circulant Matrices section for derivation), and is thus inexpensive to compute. Secondly, in the usual approach to evidence optimization (Bishop 2006), we find a MAP solution based on fixed hyperparameters and then determine optimal hyperparameters based on a fixed MAP solution, iterating until convergence. This iterative approach is taken because computing a MAP solution is usually expensive.

In contrast, the error term $\frac{1}{2\sigma^2}(\mathbf{z}-\bar{\mathbf{w}})'(\mathbf{z}-\bar{\mathbf{w}}) + \frac{1}{2\gamma^2}\bar{\mathbf{w}}'\mathbf{\Gamma}^{-1}\bar{\mathbf{w}}$ can be computed entirely in the Fourier domain as the discrete Fourier transform is an orthogonal transform, and thus preserves inner products up to a constant: $\mathbf{a}'\mathbf{b} = \frac{1}{n}F(\mathbf{a})'F(\mathbf{b})$. Remarkably, during evidence optimization, we do not need to perform any forward or inverse Fourier transforms. We found that this non-iterative approach, enabled by the choice of circulant matrices, was more computationally efficient than the usual iterative solutions by almost an order of magnitude.

We do, however, neglect the derivatives of $\mathbf{z}$ with respect to the hyperparameters (recall that $\bar{\boldsymbol{\varphi}}^k$ is a function of the hyperparameters), and we need to compensate for this fact. The complete algorithm is as follows:

```
Find optimal hyperparameters based on z = y
While convergence of hyperparameters and evidence is not reached
       While convergence of auxiliary equation is not reached
              Estimate each φ̄^k, μ̄
       End while
       Set z = y − Σ_k C′(s^k)Bφ̄^k − μ̄
       Recompute hyperparameters based on z
End while
Return z
```

Convergence of hyperparameters is then usually reached in less than 5 iterations, and our implementation of this algorithm is highly optimized and fast enough to be of practical use in day-to-day research. For instance, a wideband signal lasting about 3 minutes can be despiked in about 20 seconds on a medium-powered computer running 32-bit Matlab or in about 7 seconds on a high-powered computer running 64-bit Matlab on the Intel Core i7 platform. Computational times rise as $n \log n$, where n is the length of the wideband signal, because of the use of FFTs by the algorithm. The method can scale to arbitrarily long recordings by performing the despiking on short segments of data, an approach we detail in the Chunking section.

B.3 Empirical estimate of g

We have now shown how to optimize σ and γ using evidence optimization. There remains a vector of free hyperparameters $\mathbf{g}$ which controls our assumptions on the frequency content of the LFP. Choosing this vector properly is crucial, since in essence the assumed frequency content of the LFP is the only means through which the model can discriminate which portion of the STA around the time of a spike is artifactual and which portion is due to legitimate spike-LFP correlations.

If $p(\mathbf{w}) = N(\mathbf{w}|0, \gamma^2\mathbf{\Gamma})$, then the covariance of $\mathbf{w}$ is $\gamma^2\mathbf{\Gamma}_{\mathrm{ij}} = E_{\mathbf{w}}(w_i w_j)$, where $E_{\mathbf{w}}$ denotes the expected value over all $\mathbf{w}$. But because we constrain the prior matrix to be circulant, the covariance of $\mathbf{w}$ is completely described by its autocovariance, $\gamma^2\mathbf{\Gamma}_{\mathrm{ij}} = E_{\mathbf{w}}\left(E_k\left(w_k w_{[k+i-j]}\right)\right)$. The Fourier transform of this autocovariance is the expected power spectral density (PSD) of $\mathbf{w}$, and thus we have:

$$E_{\mathbf{w}}(|F(\mathbf{w})|^2) \propto \mathbf{g} \tag{B-17}$$

Thus, $\mathbf{g}$ should be matched to the expected PSD of the LFP. Here we have two difficulties. First, we never actually observe the LFP, only the wideband signal. Second, we typically observe only a handful of

such wideband signals, thus even if $E_{\mathbf{w}}(|F(\mathbf{w})|^2) = E_{\mathbf{y}}(|F(\mathbf{y})|^2)$ because there are no spikes or noise in our recording, the mean empirical PSD of a handful of wideband signals is very noisy.

We resolved these issues by using our knowledge of the properties of the LFP and the wideband signal. We know that in a certain range of frequencies where the LFP has most of its power, say 1-150 Hz, it account for most of the power in the PSD of the wideband signal and therefore $E_{\mathbf{w}}(|F(\mathbf{w})|^2) \approx E_{\mathbf{y}}(|F(\mathbf{y})|^2)$ in this frequency range. We therefore fit a function to the PSD of the wideband signal in the range of 1 to 150 Hz and extrapolated this function at lower and higher frequencies to obtain $\mathbf{g}$. Extrapolation with highly nonlinear functions is unadvisable, so we used functions which were constant at the lowest frequencies and linear in log-log space at higher frequencies, consistent with previous reports (Bédard and Destexhe 2009).

We found that the function $-\exp(1 + \log x)$, which is constant for small x and decreases linearly for large x to provide an excellent fit to the PSD of the wideband signal within the range of 1 to 150 Hz. A procedure for fitting this function to a PSD is implemented in the Matlab function `fitLFPpowerSpectrum`. An example of such a fit is shown in Figure B-2. Note that we purposefully set $\mathbf{g}$ to be lower than the empirical PSD of the wideband signal at the highest frequencies, as we know that most of the power at these frequencies is actually due to spikes. Some recordings may require a different function to be fit to the PSD of the wideband signal, for example when there is a peak in the PSD in a range of frequencies. Such a peak could happen for a variety of reasons, for example because of a low-pass filter in the recording system whose cutoff overlaps the PSD of the LFP or because of intrinsic properties of the recorded brain region. In that case, a sum of the function $-\exp(1 + \log x)$ and a logistic function could be used to fit the PSD of the LFP.

We obtained excellent results with this method of choosing $\mathbf{g}$. Other functional forms that closely followed the envelope of the PSD of the wideband signal also performed well. Choosing a loose $\mathbf{g}$, on the other hand, yielded less satisfactory results. For example, a choice of $\mathbf{g} = 1$ for frequencies smaller than a cutoff of 200 Hz and a vanishingly small value elsewhere performed poorly. We therefore highly recommend that $\mathbf{g}$ be selected on the basis of empirical PSD of the wideband signal.

B.4 Choise of basis

A final implicit set of hyperparameters is the basis $\mathbf{B}$. Our algorithm assumes that this basis has the form $[\mathbf{B}_s; \mathbf{0}]$, which as we showed earlier is an appropriate form when we assume that spikes are finite. The height of $\mathbf{B}_s$ corresponds to the duration of spikes measured in samples. As spike-sorting algorithms

traditionally use snippets ranging in duration from about 1.5 to 3 ms, it is safe to assume that 3 ms is an upper bound for the duration of spike waveforms. Note that this duration does not correspond to the physical duration of a spike, which is shorter, but rather to the duration of the measured spike waveform, which is affected by the filters of the recording system. We chose the spike length to be equal to 3 ms (30 samples), and aligned spikes so that their peak was located at the 11th sample.

Thus, the basis $\mathbf{B}_s$ was taken to be the identity matrix of size 30x30. As we recorded in areas where neurons fire at high rates, and our sampling rate was relatively low, the spike waveforms $\boldsymbol{\varphi}^k$ were well constrained in this basis. When recording at higher sampling rates, or in areas with low firing rates, however, spike waveforms may be poorly constrained. In this case, $\mathbf{B}_s$ can be chosen to be undercomplete, thus parametrizing spike waveforms in a low-dimensional subspace. For example, we could express the waveforms in a spline basis with a higher density of knots around the time of the peak of spikes than elsewhere.

We must note, however, that this method has its limitations. Our algorithm is not well adapted to short recordings that contain a handful of spikes (say, less than 100), as it needs a sufficient amount of data to constrain the spike waveforms. When despiking trial data, therefore, one should perform the despiking on a continuous wideband signal, splitting the data into smaller chunks for trial analysis afterwards, if necessary.

Our implementation of the despiking algorithm automatically multiplies the chosen basis by a whitening matrix $\mathbf{W}$ obtained through a singular value decomposition, so that the basis internally by the algorithm is orthogonal, $\mathbf{B}_s'\mathbf{W}'\mathbf{W}\mathbf{B}_s = \mathbf{I}$. This tends to improve numerical conditioning appreciably. The implementation then expresses the spike waveforms $\boldsymbol{\varphi}^k$ in the original basis, so this is completely transparent to the end-user.

B.5 Chunking

When the signal is too long, it becomes inconvenient to perform the matrix operations required to estimate the local field potential. In addition, in long recordings electrode drift can cause spike waveforms to shift. We addressed these issues by splitting the signal into overlapping chunks, estimating the local field potential for each chunk, then stitching the results back to obtain the complete despiked signal.

The chunking scheme is illustrated graphically in Figure B-3. Here we show a signal which is much shorter than one would use in reality for ease of visualization. The signal is split into overlapping chunks. Within each chunk, the signal is multiplied by an analysis window and added to a time reversed version of itself multiplied by 1 minus this window, thus creating a composite signal. The analysis window has a trapezoid shape:

$$f(i) = \begin{cases} \frac{1}{2} + \frac{1}{2}\frac{i}{N \cdot \text{overlap}} & i < N \cdot \text{overlap} \\ 1 & N \cdot \text{overlap} < i < N - N \cdot \text{overlap} \\ 1 - \frac{1}{2}\frac{i - (N - N \cdot \text{overlap})}{N \cdot \text{overlap}} & i > N - N \cdot \text{overlap} \end{cases} \quad \text{(B-18)}$$

Here N is the length of a segment and `overlap` is a variable that controls the degree of overlap between chunks. The edges are thus blended together to avoid discontinuities. The composite signal within each chunk is then despiked as in the previous sections. The despiked signals are put back together by multiplying each signal by a synthesis window and summing the windowed signals together. The synthesis window also has a trapezoid shape:

$$h(i) = \begin{cases} 0 & i < 2N \cdot \text{overlap} \\ \frac{(i - 2N \cdot \text{overlap})}{N \cdot \text{overlap}} & 2N \cdot \text{overlap} < i < 3N \cdot \text{overlap} \\ 1 & 3N \cdot \text{overlap} < i < N - 3N \cdot \text{overlap} \\ 1 - \frac{i - (N - 3N \cdot \text{overlap})}{N \cdot \text{overlap}} & N - 3N \cdot \text{overlap} < i < N - 2N \cdot \text{overlap} \\ 0 & i > N - 2N \cdot \text{overlap} \end{cases} \quad \text{(B-19)}$$

The support of the synthesis window is smaller than the size of the analysis window, thus discarding the edges within each chunk. For the initial segment the analysis and synthesis windows are of a different shape to avoid artifacts at the beginning of the recording signal. The analysis window is given by:

$$f_{\text{first}}(i) = \begin{cases} 1 & i < N - 4N \cdot \text{overlap} \\ 1 - \frac{1}{2}\frac{i - (N - 4N \cdot \text{overlap})}{N \cdot \text{overlap}} & N - 4N \cdot \text{overlap} < i < N - 2N \cdot \text{overlap} \\ 0 & i > N - 2N \cdot \text{overlap} \end{cases} \quad \text{(B-20)}$$

And the synthesis window is:

$$h_{\text{first}}(i) = \begin{cases} 1 & i < N - 6N \cdot \text{overlap} \\ \dfrac{i - (N - 6N \cdot \text{overlap})}{N \cdot \text{overlap}} & N - 6N \cdot \text{overlap} < i < N - 5N \cdot \text{overlap} \\ 0 & 0 \end{cases} \tag{B-21}$$

The analysis and synthesis windows for the last chunk are mirror inverses of those of the first chunk.

The number of chunks is determined by N, which is given by the user implicitly through the $\mathbf{g}$ parameter, and the variable `overlap`, which is given explicitly. Because the number of chunks must be an integer, however, the method automatically adjusts the overlap upwards so that chunks are equispaced and that recovery windows add up to 1 everywhere. The method is implemented in Matlab as the function `despikeLFPbyChunks`.

In simulations where ground truth was available, despiking by chunks gave essentially the same results as despiking an entire signal provided that chunks were large enough to obtain reliable estimates of spike waveforms. The chunk size should be a power of 2 for the fastest speeds as FFT routines are typically optimized for such cases. We recommend using chunks that are a few minutes long (say 2-5 minutes) so that several hundred spikes will be present in each chunk.

B.6 Properties of circulant matrices

A circulant matrix is defined as:

$$C(\boldsymbol{x}) = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_n & x_1 & x_2 & & x_{n-1} \\ x_{n-1} & x_n & x_1 & & x_{n-2} \\ \vdots & & & \ddots & \vdots \\ x_2 & x_3 & x_4 & \cdots & x_1 \end{bmatrix} \tag{B-22}$$

By this definition, circulant matrices are closed under addition and $C(\boldsymbol{a}) + C(\mathbf{b}) = C(\mathbf{a} + \mathbf{b})$. The transpose of a circulant matrix is circulant. The product of two circulant matrices is circulant. Note also that the identity matrix $\mathbf{I}$ is circulant. Multiplication of the transpose of a circulant matrix by a vector corresponds to a circular convolution:

$$C'(\mathbf{a})\mathbf{b} = (a \circledast b)_i = \sum_j a_{[i-j]} b_j \tag{B-23}$$

Here $a_{[k]} = a_{(k \text{ modulo } n)+1}$ with n being the length of $\mathbf{a}$. By the circular convolution theorem we have that:

$$C'(\mathbf{a})\mathbf{b} = F^{-1}\big(F(\boldsymbol{a})F(\boldsymbol{b})\big) \tag{B-24}$$

Here $F(\mathbf{a})$ is the discrete Fourier transform of $\mathbf{a}$. This implies the following properties:

$$\begin{aligned} C'(\mathbf{a})\mathbf{b} &= C'(\mathbf{b})\mathbf{a} \\ C(\mathbf{a})\mathbf{b} &= \mathrm{F}^{-1}\left(\overline{F(\boldsymbol{a})}F(\boldsymbol{b})\right) \\ \big(C'(\mathbf{a})\big)^{-1} &= C'\left(F^{-1}\left(\frac{1}{F(\mathbf{a})}\right)\right) \end{aligned} \tag{B-25}$$

The determinant of a circulant matrix can be found by noting that:

$$\begin{aligned} C'(\mathbf{a})\mathbf{b} &= \mathrm{F}^{-1}\big(F(\boldsymbol{a})F(\boldsymbol{b})\big) \\ &= \hat{\mathbf{F}}^{-1}\mathrm{diag}(\hat{\mathbf{F}}\mathbf{a})\hat{\mathbf{F}}\mathbf{b} \end{aligned} \tag{B-26}$$

Here $\hat{\mathbf{F}}$ is the discrete Fourier transform matrix (DFT matrix) which maps a vector onto its discrete Fourier transform. Hence:

$$\begin{aligned} |C'(\mathbf{a})| &= \left|\hat{\mathbf{F}}^{-1}\mathrm{diag}\big(\hat{\mathbf{F}}\mathbf{a}\big)\hat{\mathbf{F}}\right| \\ &= \left|\hat{\mathbf{F}}^{-1}\right|\left|\mathrm{diag}\big(\hat{\mathbf{F}}\mathbf{a}\big)\right|\left|\hat{\mathbf{F}}\right| \\ &= \left|\mathrm{diag}\big(\hat{\mathbf{F}}\mathbf{a}\big)\right|\left|\hat{\mathbf{F}}\hat{\mathbf{F}}^{-1}\right| \\ &= \prod_i \mathrm{abs}\big(\hat{\mathbf{F}}\mathbf{a}\big)_i \end{aligned} \tag{B-27}$$

This expression is valid for $\mathbf{a} \in \mathbb{R}^N$. Therefore the log-determinant of a circulant matrix is:

$$\log|C'(\mathbf{a})| = \sum_i \log|F(\mathbf{a})_i| \tag{B-28}$$

A symmetric circulant matrix corresponds to circular convolution by a symmetric kernel, and its corresponding Fourier coefficients are real. A symmetric circulant matrix whose corresponding Fourier coefficients are positive is positive definite and therefore is a valid covariance matrix.

B.7 Figures

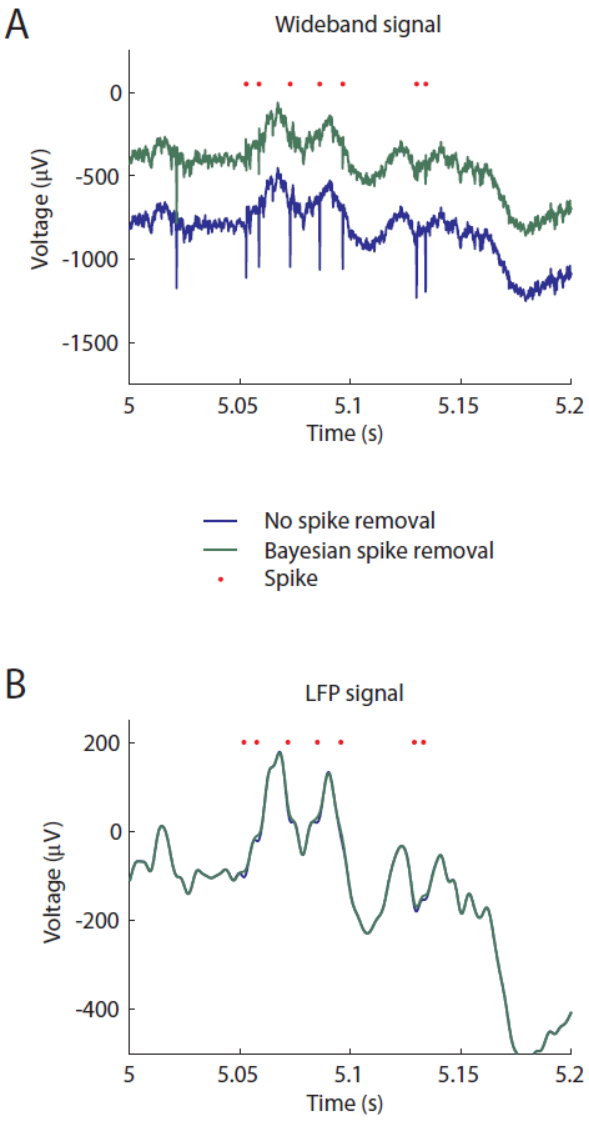


Figure B-1: Effect of spike removal in an example recording

A) wideband signal before and after despiking B) low-pass filtered wideband signal before and after despiking. Note that although the spike artifact is barely visible after low-pass filtering, it is highly stereotyped and always in sync with spike times. Therefore it can seriously bias analyses which look at spike-LFP temporal relationships, as we show in the main text.

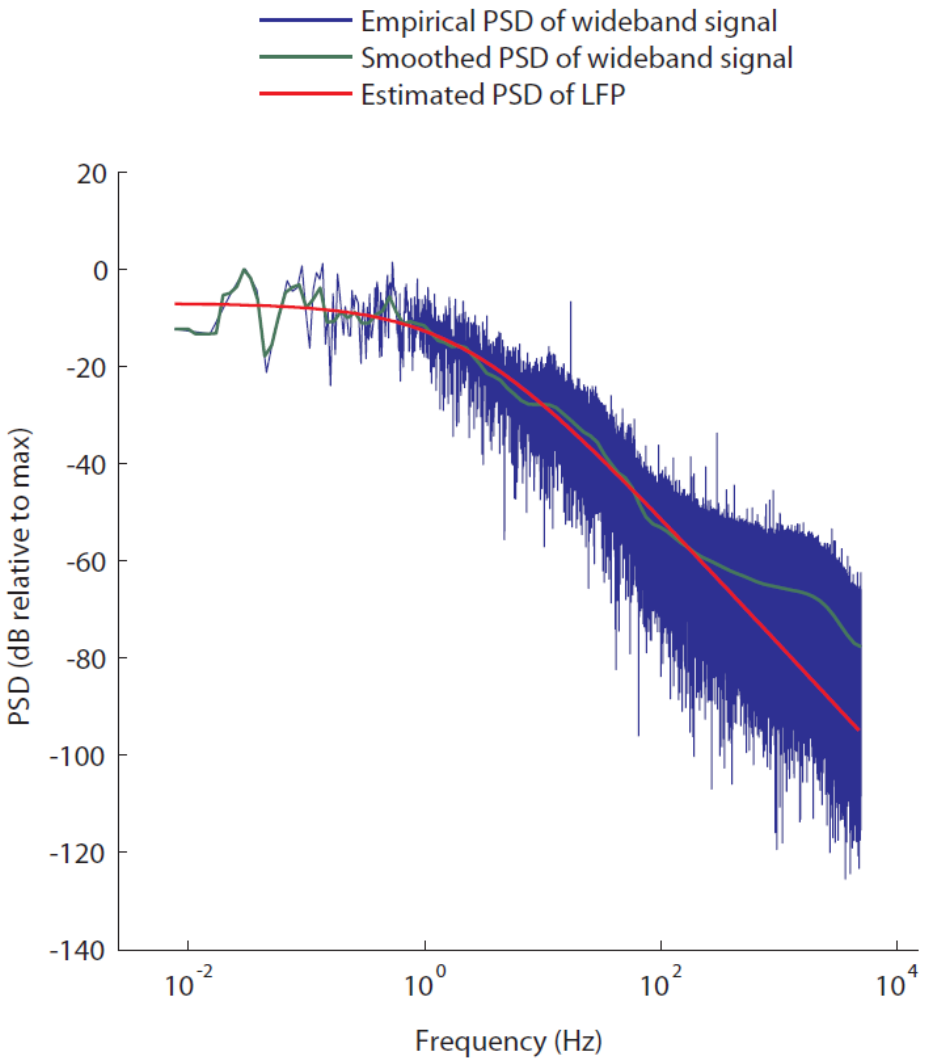


Figure B-2: Choice of **g**

g is constrained to be a low-complexity parametric function. It should match the PSD of the wideband signal in the range of 1 to 150 Hz. At higher frequencies, **g** can undershoot the PSD of the wideband signal; much of the power at these frequencies is attributable to noise and spikes.

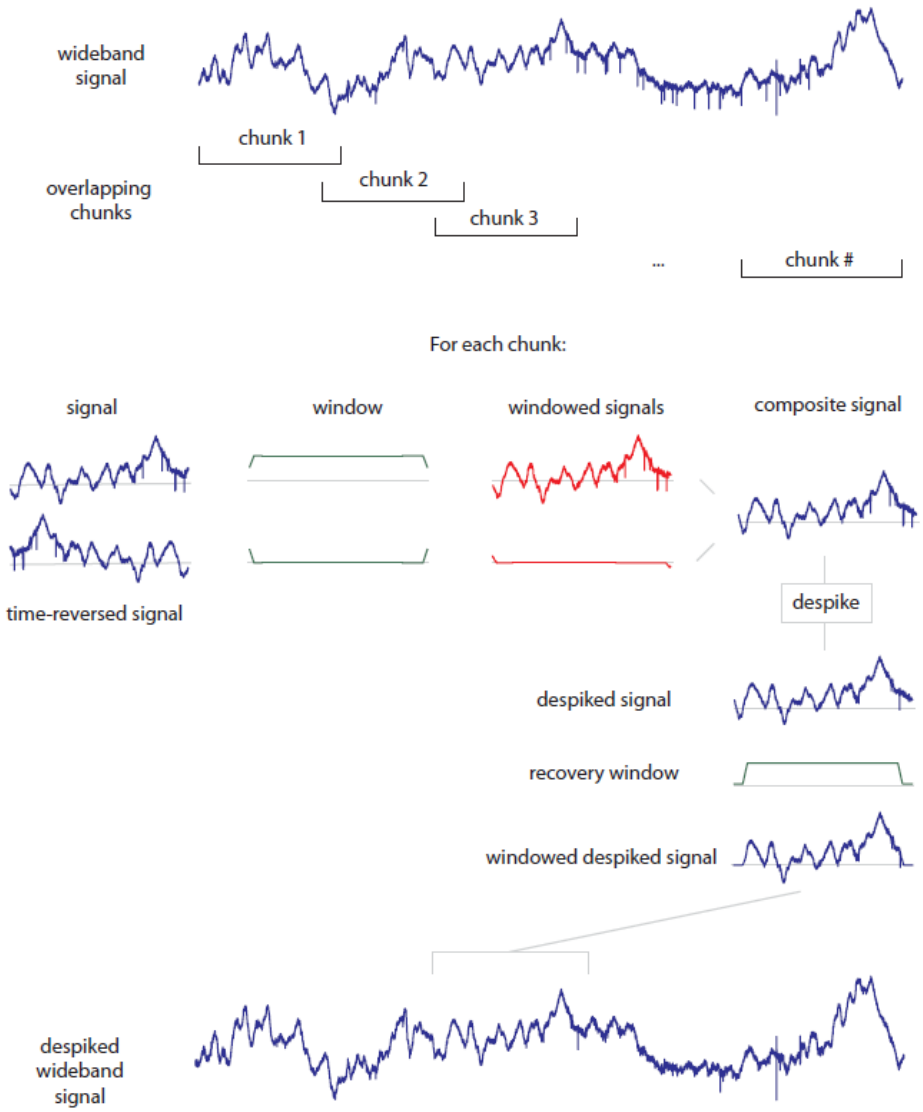


Figure B-3: Chunking procedure

In the chunking procedure, the signal is split into overlapping chunks (top). For each chunk, a composite signal is formed by the addition of a windowed signal and time-reversed version of this signal windowed with a complementary window. The composite signal is fed into the despiking algorithm. The recovered signal chunk is then windowed through a synthesis window. These windowed despiked chunks are added together to recover the complete signal. The recovery windows add up to 1 at every point in time.

C. Estimating receptive fields in MST

C.1 Stimulus generation

Dots within the aperture of the random-dot patterns were assigned an instantaneous velocity in the horizontal and vertical directions (u,v) depending on their position (x,y) within their aperture (Koenderink, 1986):

$$(u,v) = v_0(\cos\theta_1, \sin\theta_1) \text{ (for translation)} \quad (1)$$

$$(u,v) = \omega_0\left(x\cos\theta_2 - y\sin\theta_2, x\sin\theta_2 + y\cos\theta_2\right) \text{ (for expansion/rotation)} \quad (2)$$

$$(u,v) = \omega_0\left(x\cos\theta_3 + y\sin\theta_3, x\sin\theta_3 - y\cos\theta_3\right) \text{ (for deformation)} \quad (3)$$

We set v_0 = 40 deg/s and $\omega_0 = 2\,Hz$, as these values elicited robust responses from most MST neurons (Duffy and Wurtz, 1991; Tanaka and Saito, 1989). The values of $\theta_1, \theta_2, \theta_3$ were sampled at 45 degree intervals, yielding a basic set of 24 stimuli. Stimuli were presented in 500 ms trials in pseudorandom order, with 5 repeats per stimulus.

The continuous optic flow stimulus was generated according to:

$$u(x,y,t) = s_1(t)\cos\theta + \left(s_3(t) + s_5(t)\right)x\cos\theta + \left(-s_3(t) + s_5(t)\right)y\ \sin\theta \quad (4)$$

$$v(x,y,t) = s_2(t)\ \sin\theta + \left(s_4(t) + s_6(t)\right)x\sin\theta + \left(s_4(t) - s_6(t)\right)y\ \cos\theta \quad (5)$$

where s_1 and s_2 correspond to the translation speeds in the x and y directions, s_3 and s_4 to the magnitude of expansion and contraction, and s_5 and s_6 to the components of deformation. These values were determined by low-pass filtering independent streams of Gaussian-distributed values, with a cutoff of 2Hz. The magnitude of each component was scaled so that the distribution of optic flow components (expansion, rotation, and deformation) had a standard deviation of 1 s^{-1}, while that of the translation components was 20 deg./second. The position of the aperture was determined by another pair of low-pass filtered Gaussian variables with a cutoff of 0.05-0.10 Hz and standard deviation of 10 – 15 deg.

C.2 Model fitting and validation

C.2.1 Gradient boosting and cross-validation

The form of the hierarchical models, in which subunits are combined additively, makes them amenable to an estimation procedure known as gradient boosting (Buhlmann and Hothorn, 2007), which is a stepwise fitting procedure that introduces an assumption of sparseness (Friedman et al., 2000). Briefly, gradient boosting starts with an empty model (consisting entirely of a constant firing rate), and iteratively adds subunits whose output is most similar to the current model residual (i.e., the difference between actual and predicted firing rates). Early subunits tend to fit to prominent effects while later tend to fit to noise. The process is deliberately slowed by setting a subunit's gain to a fraction α of its optimal value when it is added. This which makes the procedure less greedy and allows subsequently added subunits to account for subtler features of the data.

In order to limit the number of degrees of freedom in the model, we determine the optimal number of boosting iterations by 5-fold cross-validation (Bishop, 2006). Here the data are split into 5 non-overlapping subsets of equal size, and a model is fit to the data in 4 of these subsets and used to predict the data in the leave-aside set. This process is repeated for the 5 different partitions of the data, and the prediction scores are averaged to form the cross-validated goodness-of-fit. The optimal number of iterations is then defined as the one that maximizes this cross-validated goodness-of-fit; an example run is shown in Figure S1C. Importantly, increasing the number of parameters beyond the optimal value decreases the quality of the predictions, as the extra parameters fit to noise in the training set. Thus cross-validated goodness-of-fit measures are largely insensitive to the number of model parameters.

C.2.2 Model fitting algorithm

The main optimization problem in boosting is to find, at each iteration, the parameters of the subunit whose output is most similar to the current model residual; similarity here is measured by the absolute value of the correlation between a unit's output and the current residual. A direct maximization of the correlation with respect to the parameters of MT-like units would have been challenging to perform rapidly (thousands of these optimizations have to be performed for a given model fit). Alternatively, choosing a unit out of a pool of precomputed units entails storing the output of a high-resolution filter bank of MT cells which would have stretched the memory capacity of the desktop computers performing the optimizations. We instead adopted a hybrid approach to find the optimal subunit for the linear and nonlinear integration hierarchical models, first finding the approximate optimal parameters

out of a low-resolution pool of pre-computed subunits (3 speeds, 8 directions of motion, 144 unit centers, one unit size) and refining the parameters through numerical gradient ascent. In all cases, the unit size p_σ was constrained to be ≥ 1.5 grid units (3 or 3.75 degrees). We fit the spatial and temporal structure of each model in an alternating fashion, according to the following algorithm (Ahrens et al., 2008):

1. Assume an initial temporal filter
2. Do 3 times:
    a) Boost model for 25 iterations, $\alpha = 0.5$
    b) refine temporal filter (see next section)
3. Boost model for up to 1000 iterations, $\alpha = 0.1$, with 5-fold cross-validation.

Each model took on average about one hour to fit on a recent desktop computer. For the nonlinear MT model, we fit the model using this procedure for $\beta = 0.2$ to $\beta = 1.4$ in steps of 0.2 and defined the optimal exponent as the one which yielded the model with the highest cross-validated likelihood.

As an additional validation of this estimation procedure we performed simulations in which the parameters were optimized with respect to simulated neurons with fixed collections of subunits (see Figure C-1).

C.2.3 Fitting the temporal filter

Temporal processing is linear and separable with respect to spatial processing in our model. Thus, If the spatial parameters of the model are fixed, refining the temporal filter can be done by fitting a standard generalized linear model (GLM) with one parameter per time lag and an offset (Ahrens et al., 2008). We assumed that the temporal filter was smooth through the use of a Gaussian smoothness prior (Wu et al., 2006). We fit this penalized GLM using standard methods (Paninski, 2004), using cross-validation to adjust the strength of the prior. Example resulting time filters are shown in Figure C-6.

C.2.4 Application of the model to simulated data

To verify the validity of our fitting procedures, we applied them to simulated neuronal responses, for which the subunits and nonlinearities were known. Testing was done on four spatial receptive field profiles (Figure C-1A), which were endowed with subunits corresponding to translation and expansion selectivity, with and without suppression in the anti-preferred direction. For each simulated neuron the

temporal filters were characterized by five time points [.25,1,.25,-.1,0] (from shortest to longest lag). Responses were normalized to have a standard deviation of 10 Hz, then passed through an exponential output nonlinearity, then normalized again to have a mean firing rate of 10 Hz, and finally truncated to have a maximum peak rate of 140Hz. These values were chosen so that the simulated cells were driven relatively weakly compared to the observed distribution of responses in our sample cells. The cells' outputs were then transduced to a firing rate by a Poisson noise generator. We fit the responses of the neurons with the same continuous optic flow stimulus used in the main text with the hierarchical model, and ran our subunit visualization procedure on the fitted model neurons. The results are shown in Figure C-1B. The model and visualization procedures are able to recover the correct receptive fields within the limits imposed by noise.

C.2.5 Validation metrics

Given a predicted response $\mathbf{r}$ and an observed response $\mathbf{y}$, the quality of the prediction may be assessed using the standard R^2 metric of variance accounted for):

$$R^2 = \frac{\mathrm{Var}(\mathbf{y}) - \mathrm{Var}(\mathbf{y}-\mathbf{r})}{\mathrm{Var}(\mathbf{y})} \tag{6}$$

In practice the value $R^2 = 1$ cannot be attained, as $\mathrm{Var}(\mathbf{y}-\mathbf{r})$ for a perfect prediction is the variance of the noise, which is non-negligible in physiological measurements. To recover a natural scale we thus used a corrected R^2 metric, also known as predictive power (Sahani and Linden, 2003):

$$\bar{R}^2 = \frac{\mathrm{Var}(\mathbf{y}) - \mathrm{Var}(\mathbf{y}-\mathbf{r})}{\mathrm{Var}(\hat{\mathbf{y}})} \tag{7}$$

Here $\mathrm{Var}(\hat{\mathbf{y}})$ is the variance of the unobserved noiseless signal $\hat{\mathbf{y}}$. The explainable signal variance $\mathrm{Var}(\hat{\mathbf{y}})$ is estimated from the pattern of disagreement between responses in different presentations of the same stimulus (equation 1 in Sahani and Linden, 2003).

To determine whether the relative level of responses to different classes of optic flow (translation, spirals, deformation) was correctly accounted for by the different models, we also computed a *stimulus class* $\bar{R}^2$ which introduced a free gain per optic flow type. As the model underlying this second prediction is a superset of the one-gain model, a likelihood ratio test was used to establish the significance of this second set of predictions over the first. The magnitudes of the gains were then compared to determine whether the model under- or overestimated the responses of one stimulus type

over another. In the case of the hierarchical model with linear integration, we found that the relative level of responses across stimulus types was misestimated for 70% of cells (25/36, likelihood ratio test, $p<0.001$), and in a majority of these cases (84%, 21/25) predicted responses were too weak for spiral stimuli relative to translation stimuli. In addition, stimulus class predictions were compared for the linear and nonlinear hierarchical models to determine whether the increase in quality of fit was due to better accounting of gain. We emphasize that the stimulus class metric is not an accurate reflection of the quality of the model predictions, but rather is an artifice that allowed us to isolate one mechanism underlying quality of fit.

C.2.6 Visualization of subunits

For most cells the model recovered many subunits that were similar. Displaying all the subunits recovered by the model made it difficult to discern the structure of the receptive fields because many subunits overlap. Thus, we used a second procedure to select a subset of these subunits which could account for 80% of the likelihood captured by the full model. We refit the weights of the subunits of each cell, imposing a Laplace prior on these weights (Mineault et al., 2009; Tibshirani, 1996). This yields even sparser models than those achieved with boosting. We then adjusted the strength of the prior for each cell in order to obtain the sparsest model that accounted for 80% of the likelihood relative to the full model. We then plotted the resulting subunits (Figure 4, for example), modulating the opacity and color of the subunits in proportion to the weight of the recovered subunits.

C.2.7 Analysis of subunit overlap

To gain more insight into the degree to which subunits overlap in space and in tuning, we computed the pairwise normalized distance between subunits with positive weights discovered by the visualization procedure (see previous section), and correlated it with the pairwise difference in direction and speed tuning. The normalized distance was defined as the spatial distance between subunits divided by the sum of their radii. A normalized distance > 1 indicates that the subunits are non-overlapping. An example of this analysis is shown in Figure S8A for the direction tuning of the subunits of the cell originally shown in Figure 4B. In this case, we see that differences in direction selectivity build up gradually with normalized distance.

We repeated this analysis for every cell and compiled all the pairwise differences in Figures S8B and S8C. Again, we see that differences build up gradually with normalized distance; this is true for both direction and speed tuning. This is most clearly seen with the blue trend line, which computes the running median pairwise difference in selectivity.

Figure S8D and S8E replots the same data, zooming in on the x axis (normalized distances <1.5). It is clear that strongly overlapping subunits (normalized distance <.1) generally have very similar direction and speed tuning, indicating that they correspond to the same input. On the other hand, median pairwise differences in direction selectivity reach values > 45 degrees around a normalized distance of .5, where the overlap is substantial. We conclude that MST receptive fields are densely tiled by subunits whose tuning varies rapidly as a function of position.

C.3 Alternative models

C.3.1 Linear model

The simplest model that we explored performs a linear match between the local velocity of the observed optic flow field and a preferred template. This model has linear speed tuning and cosine direction tuning, and so it is not tuned in the same sense as the hierarchical models explored in the main text. Rather it is most similar to a linear receptive field model in the luminance domain, as is often used to model LGN or V1 simple cells (Carandini et al., 2005; Chichilnisky, 2001). While this model, endowed with an exponential output nonlinearity and Poisson noise, can be fit directly through maximum likelihood methods (Paninski, 2004), we used the same boosting methodology we applied to our other models, so that the results could be directly compared. In the boosting formulation, a model cell contains subunits whose activation is proportional to $u(x,y)$ and $v(x,y)$, the horizontal and vertical components of the velocity of the stimulus inside their receptive fields:

$$f(u(x,y),v(x,y),\mathbf{p}) = p_g \sum_{xy} \left(\cos p_\theta \, u(x,y) + \sin p_\theta \, v(x,y)\right) G(x,y,p_x,p_y,p_\sigma) \quad (2)$$

Here p_g denotes the gain of a unit, p_θ denotes its preferred direction of motion, and G denotes a Gaussian as in the main text (equation 11).

The resulting tuning curve predictions for the cells originally shown in Figure 1B and 1C are presented in Figure C-2A and C-2B. The same patterns of approximately correct predictions for tuning to translation and inadequate predictions for complex optic flow were visible in most cells. Figures C-2C and C-2D compare the cross-validated likelihood and prediction $\bar{R}^2$ for the linear and linear hierarchical models. While the linear model is attractive because of its mathematical tractability, it fails to capture the more complex selectivity seen in MST responses.

C.3.2 Unrestricted nonlinear MT model

In the nonlinear MT model considered in the main text, all subunits shared a single power law nonlinearity for a given MST cell; the shape of this nonlinearity was determined by an exhaustive search. In the unrestricted nonlinear MT model, this constraint was relaxed such that each subunit had its own power-law nonlinearity selected out of a range from 0.2 to 1.4 in steps of 0.2.

Because of the added computational burden of this model, we used only precomputed MT subunits during fitting (3 speeds, 8 directions of motion, 144 unit centers, one unit size, 7 exponents for ~24000 precomputed subunits). Other aspects of the fitting procedure were similar to the models presented in the main text. The resulting quality of fits (Table 1) indicate that the added flexibility does not lead to enhanced fits.

C.3.3 Divisive center-surround model

In this model, the output of a MT subunit is divided by a weighted sum of the output of other units in its neighborhood. Calling $U(p_x, p_y, p_\theta, p_\rho)$ the raw output of an MT subunit with center at (p_x, p_y) and tuned to direction and speed (p_θ, p_ρ), a corresponding surround-suppressed subunit $N(p_x, p_y, p_\theta, p_\rho)$ is given by:

$$N(p_x, p_y, p_\theta, p_\rho) = \frac{U(p_x, p_y, p_\theta, p_\rho)}{1 + \alpha \sum_{\text{all } \Delta} S(\Delta x, \Delta y) T(\Delta\theta, \Delta\rho) U(p_x + \Delta x, p_y + \Delta y, p_\theta + \Delta\theta, p_\rho + \Delta\rho)} \quad (3)$$

Here $S(\Delta x, \Delta y)$ is the spatial weighting function, a 2D Gaussian centered at the origin with a width σ_s, while $T(\Delta\theta, \Delta\rho)$ is the tuning weighting function, given by:

$$T(\Delta\theta, \Delta\rho) = \exp(-(c\Delta\rho^2 + (1 - cos\,\Delta\theta)^2/4)/2\rho_t^2) \quad (4)$$

c was chosen so that the range of $c\Delta\rho^2$ was equal to 1. By varying α, σ_s and σ_t, we obtained several distinct center-surround models ranging in suppression strength (α), size of the spatial integration pool (σ_s), and tuning strength (σ_t).

We preset the integration size of the raw units to $p_\sigma = 1.8$ grid units. We manually picked 5 values for σ_s ranging from a small to a large surround, $\sigma_s = [0.5,1,2,3,4.5]$. We also manually picked σ_t= $[.13, .3, .42, .72, 2]$ corresponding to angular bandwidths (full-width at half max) of roughly [90,145,180,270, ∞] degrees. For each pair of parameters, we wished to find values of α corresponding to "weak" and "strong" suppression. For every parameter pair, we thus computed $N\left(p_x, p_y, p_\theta, p_\rho\right)$ for a typical stimulus sequence for a unit in the center of the screen over a large range of suppression

strengths α. For small α, the output of $N\left(p_x, p_y, p_\theta, p_\rho\right)$ was highly correlated $(r \approx 1)$ with the raw unit $U\left(p_x, p_y, p_\theta, p_\rho\right)$, while for large α this correlation reached an asymptote $r_{\min}$. We chose α values corresponding to $r = r_{\min} + [.25, .5, .7, .9, .95] \cdot (1 - r_{\min})$. We thus obtained 125 triplets of parameters (σ_s, σ_t, α) in addition to a reference triplet corresponding to $\alpha = 0$.

Because of the large number of model fits (126 per cell) involved, we used only precomputed MT subunits while fitting the divisive center-surround models (3 speeds, 8 directions of motion, 144 unit centers, one unit size). By examining the cross-validated likelihood of the fits to the continuous optic flow stimulus, we determined the optimal parameters of the surround for each cell (Figure S4A).

C.3.4 Symmetric and asymmetric subtractive surround models

In this model, we considered the possibility that the tuned, asymmetric surrounds of MT cells could contribute significantly to the optic flow selectivity of MST neurons. Rather than a single stereotyped Gaussian envelope, MT cells came in different varieties: no surround (as before), asymmetric one-sided surround, and bilaterally symmetric surround (Raiguel et al., 1995; Xiao et al., 1997). One-sided surrounds were created by the difference of two Gaussians: the centre was a positive symmetric Gaussian, while the surround was created by a spatially offset, larger Gaussian with negative weight. The output of the center and surround were combined before the half-rectifying nonlinearity.

The surround was offset from the center by a distance 1.5 times the radius at half-height of the center Gaussian (Raiguel et al., 1995). The radius of the surround was 1.5 times the radius of the center (Raiguel et al., 1995). The surround could be located at 0, 90, 180 or 270 degrees with respect to the preferred direction of the MT cell. Bilateral surrounds were created similarly by the difference of a center Gaussian and two lateral Gaussians.

We considerd 3 different surround strengths (50%, 100%, 150%). At 100% surround strength, a full screen homogeneous stimulus yielded a net null response in surround-suppressed MT cells; the weight of the surround corresponding to 100% strength was scaled by .5 or 1.5 to yield the 50% and 150% surround strengths. Models were fit using the same method as the divisive surround model.

While this change improved the quality of fits on the continuous stimulus in a manner comparable to the addition of nonlinear integration, predictions on the tuning curve stimulus set were poorer (Table 1). This latter test is a more stringent test than the first, since it contains stimuli not found in the initial set.

We also considered a model with a subtractive *symmetric* surround. This surround was created by summing the output of 8 Gaussians surrounding the center; the parameters of these Gaussians were as in the asymmetric surround model. MST cells had access to both these surround-suppressed MT cells and cells with no surrounds. Other aspects of the model were identical to the asymmetric surround model. These yielded essentially identical fits to the subtractive *asymmetric* surrounds.

Finally, we considered the possibility that the combination of a subtractive surround and an output nonlinearity could act synergistically to explain MST selectivity. In this case we considered power-law nonlinearities (exponents of .2, .4, .6, and 1.4) interacting with either symmetric or asymmetric surrounds of 3 different strengths. This more complex model yielded very similar fits to the more parsimonious single-parameter nonlinearity, regardless of whether the subtractive surround was symmetric or asymmetric (Table 1).

In summary we find that the addition of a subtractive surround (whether symmetric or asymmetric) can improve the performance of the model relative to a simple model comprised only of excitatory subunits. However, none of the subtractive surround models consistently outperformed the simple model with a simple output nonlinearity. Our conclusions are therefore that the single-parameter model provides a powerful and parsimonious account of MST selectivity.

C.3.5 Multiplicative interaction model

The results described in the main text suggest that a multiplicative interaction among inputs is important for explaining MST responses. To test this idea explicitly, we fit the 22 least noisy cells in our sample with a model that contained explicit pairwise multiplicative interactions. The functional subunits $f(\rho, \theta, \mathbf{p})$ of the models computed the sum of a pair of MT cells $M_1(\rho, \theta)$ and $M_2(\rho, \theta)$ and their multiplicative interaction:

$$f(\rho, \theta, \mathbf{p}) = p_a M_1(\rho, \theta) + p_b M(\rho, \theta) + p_c \sqrt{M_1(\rho, \theta), \cdot\, M_2(\rho, \theta)} \quad (5)$$

Here p_a, p_b and p_c are gains. The square root of the product of the subunits is taken to compress the dynamic range of the interaction term.

This multiplicative interaction model was compared against a baseline model similar to the hierarchical model with linear integration used in the main text. We fitted these models through boosting. The parameters of the multiplication model to be fitted in each boosting iteration were the 3 gains as well as the parameters of the MT cells M_1 and M_2. To make the problem tractable, the parameters of MT cells

were restricted to discrete values along a grid (5 speeds, 8 directions of motion, 144 unit centers, one unit size). As an exhaustive search over all interactions was impractical, we used a greedy algorithm to determine the parameters of each subunit:

1. Find the MT filter M_1 whose output is most similar to the current residual.
2. Project out the output of M_1 from the residual to obtain a second residual.
3. Find the MT filter M_2 such that $\sqrt{M_1(\rho,\theta),\cdot M_2(\rho,\theta)}$ is most similar to the second residual computed in step 2.
4. Fit all gains through least squares.

The subunits of the baseline linear model were also restricted to discrete values along a grid to facilitate comparisons.

We compared the cross-validated likelihood of these models on the continuous optic flow stimulus; results are shown in Figure C-5A. The model with multiplicative interactions performed better than the linear integration in 100% (22/22) of cases. For those cells for which the *continuous optic flow* stimulus spanned the spatial range of the *tuning curve* stimulus (14/22), we evaluated the quality of the model predictions, the results of which are presented in Figure C-5B. The multiplicative interaction model performed better than the linear integration model in 78% of cells (11/14, $p < .05$, binomial test). These results are consistent with the notion that multiplicative input interactions are an important property of MST cells.

C.4 Linear scaling

The need for nonlinear integration in our MST model might seem inconsistent with the results of a recent study (Heuer and Britten, 2007), which found that MST neurons, like those in MT (Britten et al., 1993), respond linearly as a function of the coherence of optic flow stimuli. Here we demonstrate that linear scaling as a function of coherence may be achieved in a model which includes nonlinear interactions. Assume that an MST neuron's output is $f(a,b)$ in response to two MT inputs a and b (this generalizes to an arbitrary number of inputs). a and b are in turn assumed to be linear as a function of the coherence c (Britten et al., 1993). Without loss of generality, we let the firing rate at zero coherence for all cells be 0. Then $a(c) = a_0 c$, $b(c) = b_0 c$. An MST cell is linear as a function of coherence if and only if the following relationship holds for all values of $a_0, b_0, c \geq 0$:

$$f(ca_0, cb_0) = cf(a_0, b_0) \tag{6}$$

The family of "power law" integration rules (Britten and Heuer, 1999):

$$f(a, b) = \left(a^\beta + b^\beta\right)^{\frac{1}{\beta}} \tag{7}$$

Satisfy this linearity condition for all β. For example, when $\beta \to \infty$, the integration rule becomes:

$$f(a, b) = \lim_{\beta \to \infty} \left(a^\beta + b^\beta\right)^{\frac{1}{\beta}} = \max(a, b) \tag{8}$$

Clearly, $\max(ca, cb) = c \max(a, b)$, and a linear response to coherence is obtained. Hence nonlinear integration is not inconsistent with linear responses to coherence. In our framework, a linear response to coherence could be achieved by replacing the output linearity $\exp(x)$ with $(x^+)^{\frac{1}{\beta}}$; this would considerably complicate the fitting procedure, however.

C.5 Figures

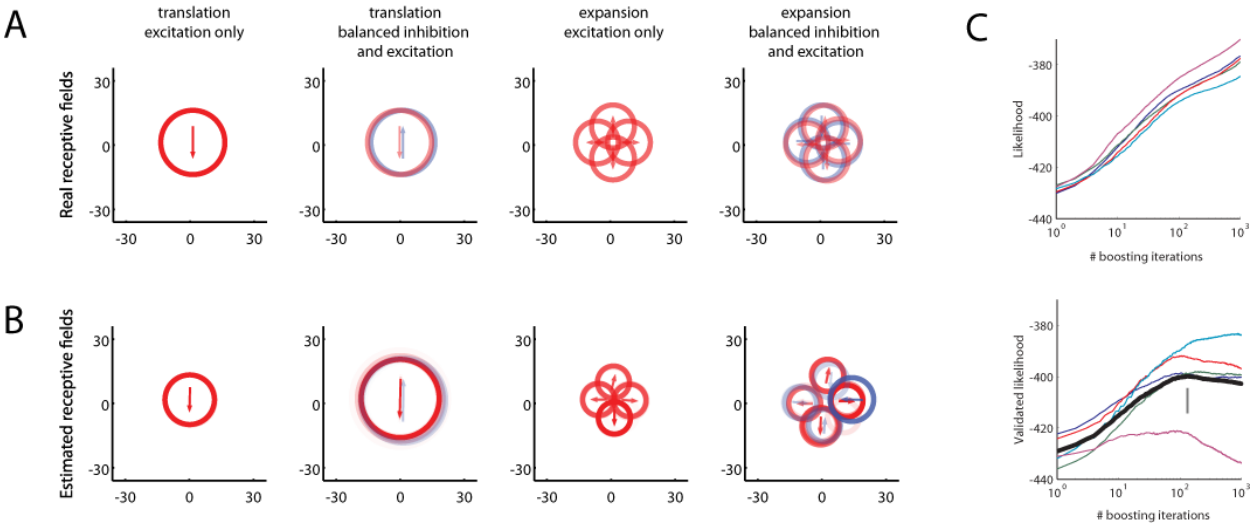


Figure C-1: Our methods can estimate veridical receptive fields. 0-4

(A) Receptive fields of simulated neurons. (B) Estimated receptive fields based on hierarchical model and subunit visualization procedure. Our methods are able to estimate veridical receptive fields within the limits imposed by noise. (C) Cross-validation example. Top: evolution of quality of fit for each fold as a function of number of boosting iterations. The likelihood of the data always increases as more parameters are added into the model. Bottom: evolution of the validated goodness-of-fit. Thin lines: evolution of the quality of the predictions for each leave-aside fold as the number of boosting iterations increases. As more parameters are added, the model starts overfitting to noise, and beyond a certain point the predictions on the leave-aside fold become worse. The optimal number of iterations varies from fold to fold, as more or less noisy data is placed at random in the fit and validation folds. Thick line: the average of the validation scores is used to determine the number of boosting iterations (indicated by a gray arrow) to be used for the final model fit, which uses all data. The number of boosting iterations is not equal to number of degrees of freedom in the model, because of the use of a damping parameter $\alpha < 1$ (see Main Methods in Chapter 4).

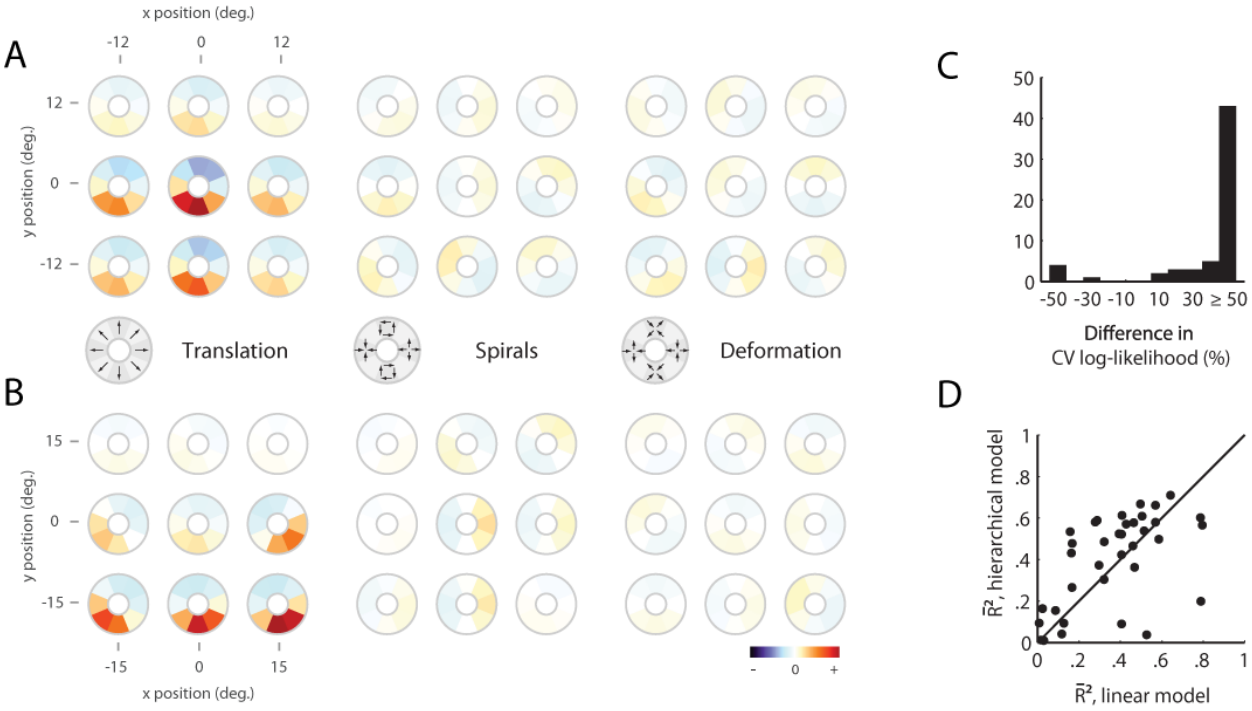


Figure C-2: Failure of linear model to account for MST responses.

(A) and (B): Predicted responses to tuning curve stimuli based on linear model for cells depicted in Figure 1B and 1C. The predicted responses to translation are approximately correct, but responses to spirals and deformation are not captured. (C) percentage difference in cross-validated log-likelihood between hierarchical and linear models. (D) $\bar{R}^2$ of tuning curve predictions compared between hierarchical and linear models. Despite its higher dimensionality, the hierarchical model performs better on validation sets than the linear model.

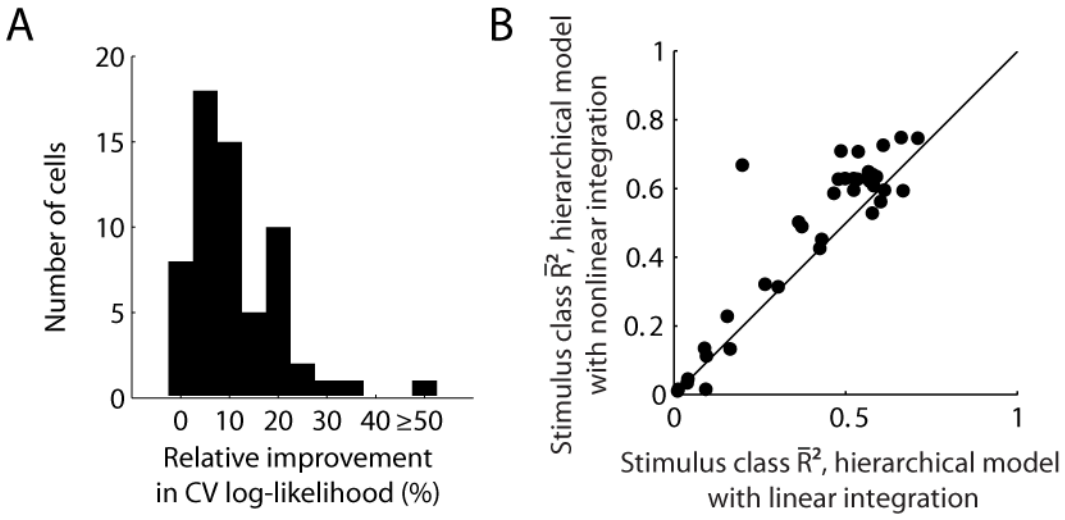


Figure C-3: Analysis of relative goodness-of-fit of linear and nonlinear integration models.

(A) Relative cross-validated log-likelihood between nonlinear and linear hierarchical models. Note the sizable number of cells showing improvements of 20% or more after the addition of a single parameter (β). (B) Stimulus class $\bar{R}^2$ of tuning curve predictions compared between linear and nonlinear hierarchical models. The stimulus class $\bar{R}^2$ metric measures fraction of variance accounted for assuming independent gains for different stimulus classes (translation, spirals, deformation). By construction, it is insensitive to how well each model predicts the relative gain of responses between different stimulus classes. The hierarchical model with nonlinear integration retains better predictive ability according to this metric, indicating that the better fits are not entirely due to better accounting of relative gain between stimulus classes.

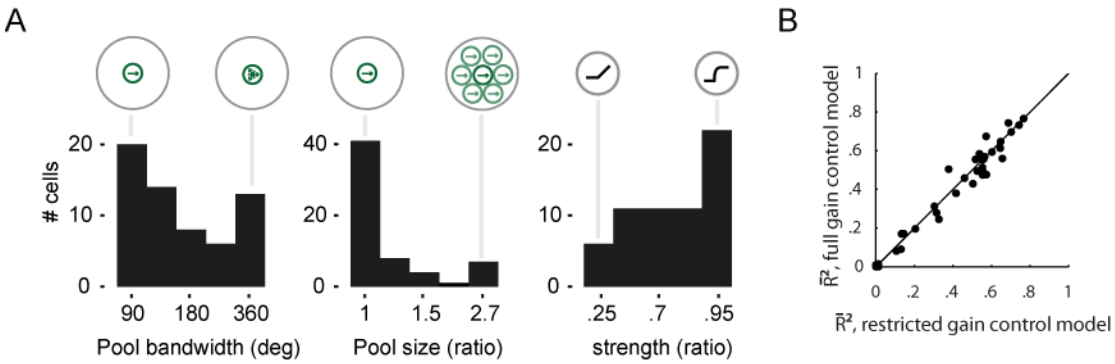


Figure C-4: Gain control model results.

(A) Histogram of optimal parameters for gain control pool. Left: optimal bandwidth was skewed towards small bandwidths (tuned gain control), although some cells preferred untuned gain control pools. Middle: optimal pool size was the same as that of the subunits themselves, ruling out strong center-surround effects. Right: optimal gain control strength was skewed towards strong gain control, which gives more compressive effects, consistent with results in Figure 6. Overall, these results are consistent with a strong, tuned and spatially limited gain control mechanism mathematically equivalent to a static compressive nonlinearity (B) Quality of predictions of gain control model with optimal unconstrained pool (full model) compared with that of gain control model with pool constrained to be tuned and of limited spatial extent (restricted model). The full model does not lead to predictions appreciably better than the restricted model whose gain control pool has an effect equivalent to a static compressive nonlinearity.

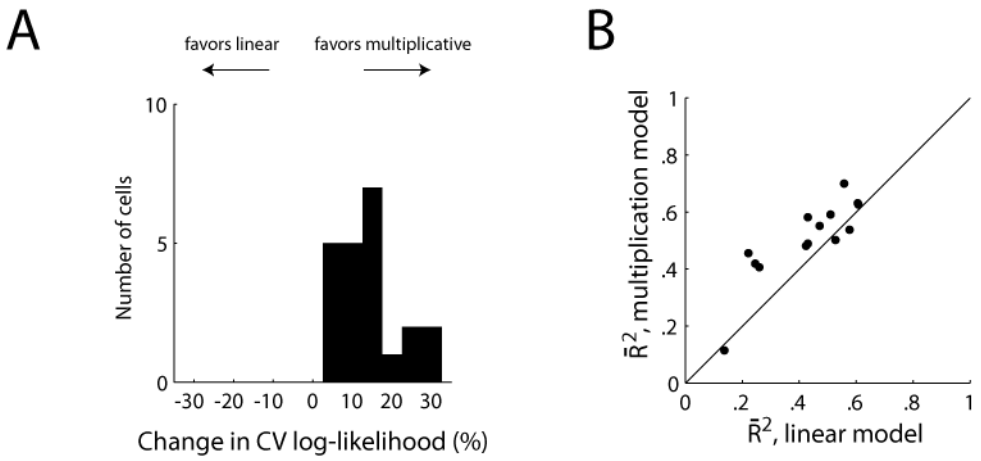
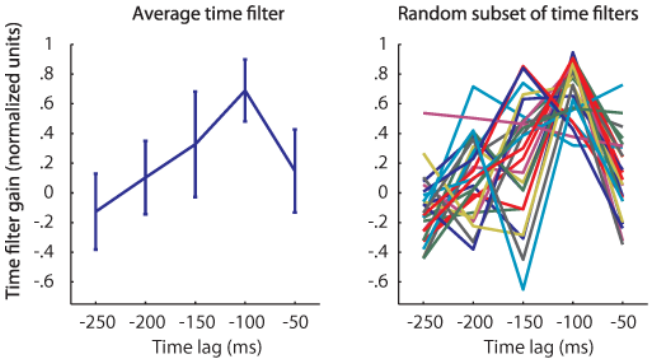
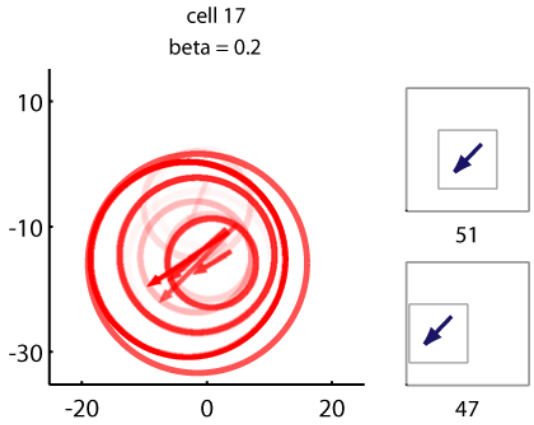
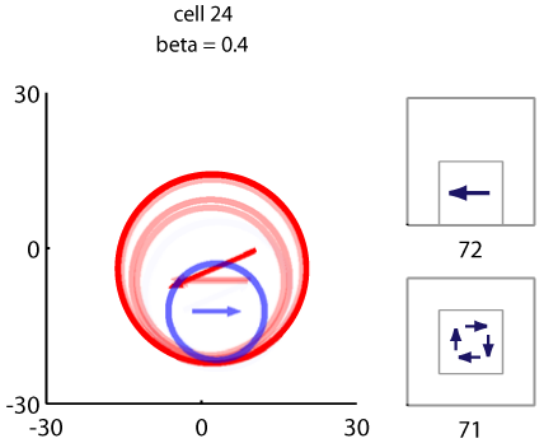


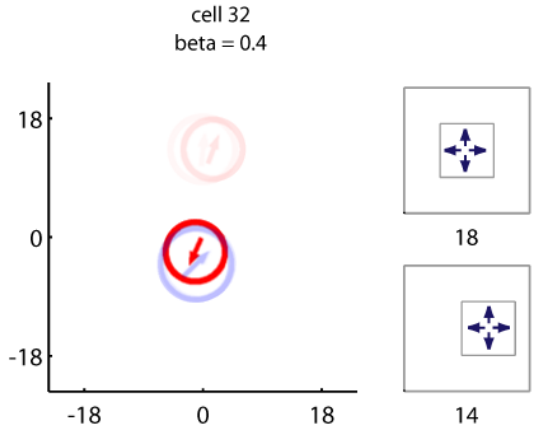
Figure C-5: Multiplicative interaction model confirms existence of nonlinear integration mechanism.

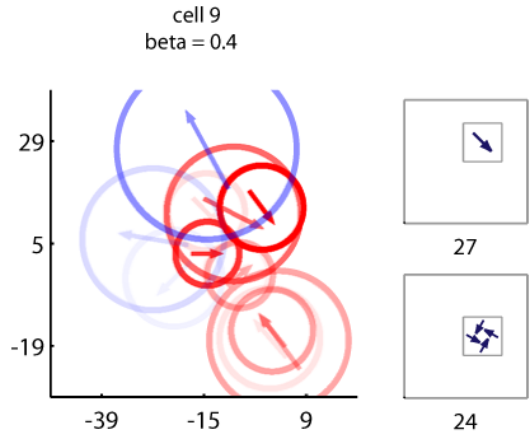
(A) Cross-validated log-likelihood of linear hierarchical model versus a model with multiplicative pairwise interactions. Note that only the most responsive cells were used for these fits because the multiplicative interactions model has a high number of degrees of freedom. (B) Relative $\bar{R}^2$ of tuning curve predictions. A model with explicit multiplicative interactions provides a better description of the data, consistent with a prominent nonlinear integration mechanism.

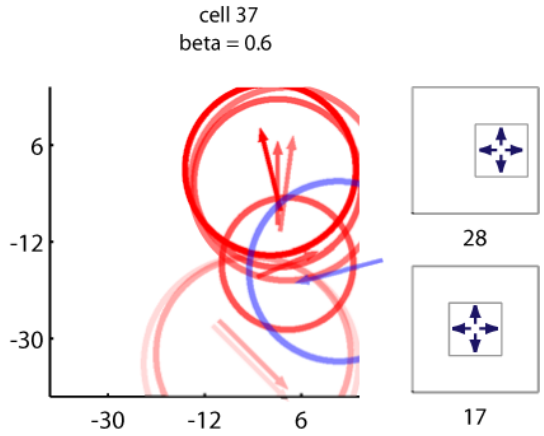

Average time filter
Random subset of time filters
Time filter gain (normalized units)
1
.8
.6
.4
.2
0
-.2
-.4
-.6
-250 -200 -150 -100 -50
Time lag (ms)

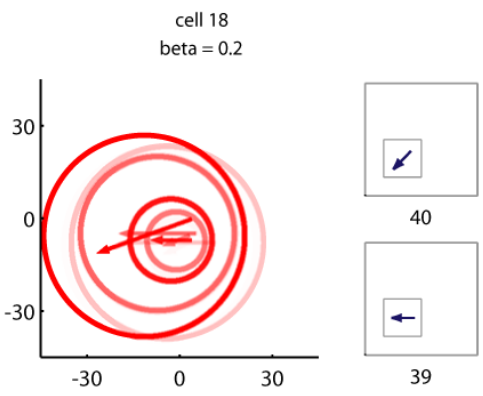

cell 17
beta = 0.2
10
-10
-30
-20 0 20
51
47

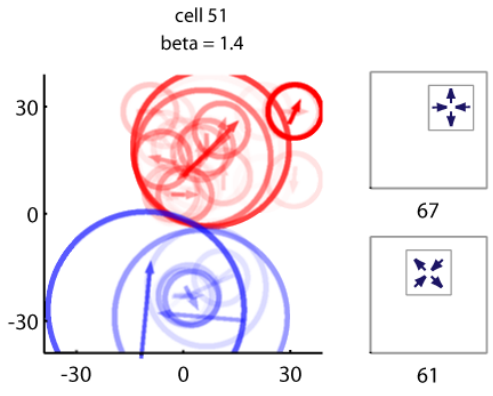

cell 24
beta = 0.4
30
0
-30
-30 0 30
72
71

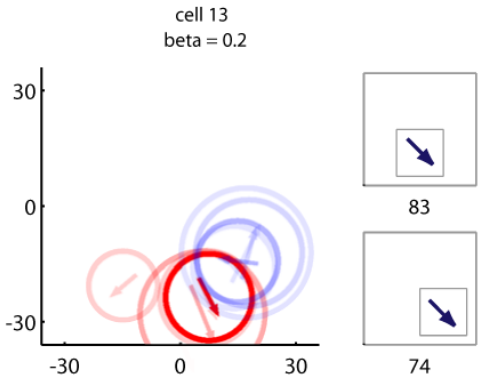

cell 32
beta = 0.4
18
0
-18
-18 0 18
18
14

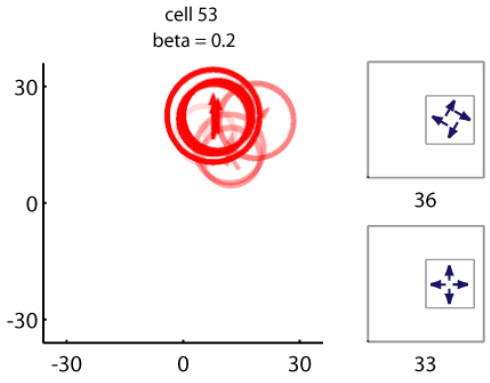

cell 9
beta = 0.4
29
5
-19
-39 -15 9
27
24

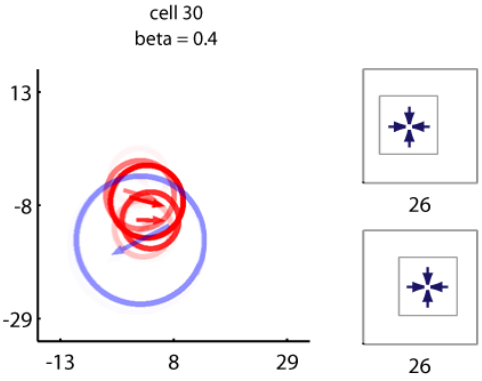

cell 37
beta = 0.6
6
-12
-30
-30 -12 6
28
17

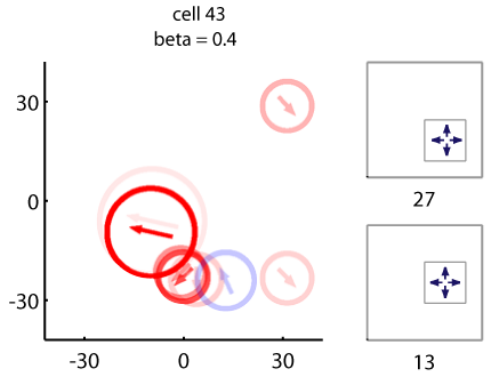

cell 18
beta = 0.2
30
0
-30
-30
0
30
40
39

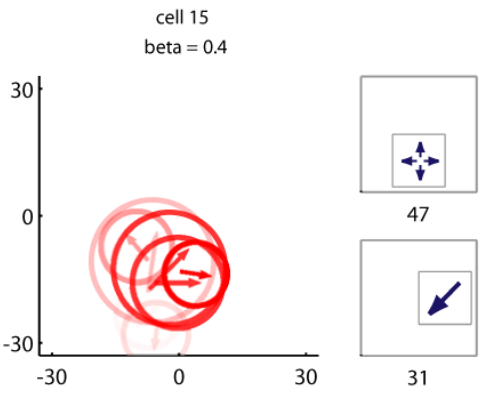

cell 51
beta = 1.4
30
0
-30
-30
0
30
67
61


cell 13
beta = 0.2
30
0
-30
-30
0
30
83
74


cell 53
beta = 0.2
30
0
-30
-30
0
30
36
33


cell 30
beta = 0.4
13
-8
-29
-13
8
29
26
26


cell 43
beta = 0.4
30
0
-30
-30
0
30
27
13


cell 15
beta = 0.4
30
0
-30
-30
0
30
47
31

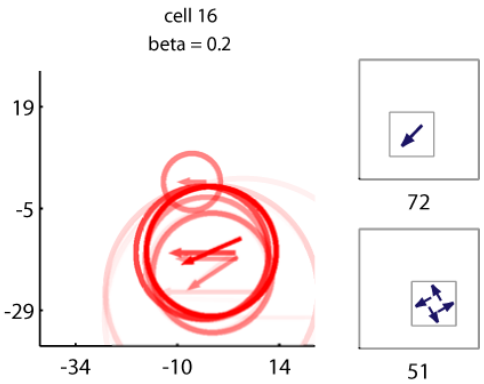

cell 16
beta = 0.2
19
-5
-29
-34
-10
14
72
51

Figure C-6 (two pages): Receptive field parameters for sample cells.

Page 1, top left: mean temporal filter and random subset of temporal filters found for the population of MST cells. In most cases, temporal filters are integrative and peak at 100 ms. Page 1, other positions and Page 2: Subunits of 13 sample cells, including rotation, translation, contraction, and deformation tuned cells. To the right of the receptive fields are pictured the two tuning curve stimuli eliciting the greatest response in the cell; the number underneath these diagrams is the measured firing rate in Hertz.