

# DECODING INFORMATION IN THE VISUAL CORTEX

SCOTT C. LOWE



Doctor of Philosophy  
School of Informatics  
University of Edinburgh

2016

Scott C. Lowe:  
*Decoding information in the visual cortex*  
Doctor of Philosophy, 2016

SUPERVISORS:

Mark van Rossum  
Stefano Panzeri  
Alex Thiele

## LAY SUMMARY

---

## ABSTRACT

---

Short summary of the contents in English.

## ACKNOWLEDGEMENTS

---

Thank you to Mark van Rossum, Stefano Panzeri, Alex Thiele, Nikos Logothetis, Xing Chen, Mehdi Sanayei, Daniel Zaldivar, Yusuke Murayama,

## DECLARATION

---

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

*Edinburgh, 2016*

---

Scott C. Lowe,

March 10, 2016

## CONTENTS

---

1	INTRODUCTION	1
1.1	Organisation of the Thesis . . . . .	1
2	PERCEPTUAL LEARNING IN V1 AND V4	2
2.1	Background . . . . .	2
2.1.1	Perceptual Learning . . . . .	2
2.1.2	Information Theory, and its applications within Neuroscience . . . . .	5
2.2	Experimental Methods . . . . .	8
2.3	Methods . . . . .	13
2.3.1	Artifact elimination . . . . .	13
2.3.2	Misc . . . . .	18
2.4	Noise correlations . . . . .	18
2.4.1	Background . . . . .	18
2.4.2	Methods . . . . .	19
2.4.3	Results . . . . .	20
2.5	Decoder-based methods . . . . .	21
2.5.1	Probabilistic Model . . . . .	24
2.5.2	Results . . . . .	27
2.6	Information Theory-based methods . . . . .	32
2.6.1	Session-based analysis . . . . .	32
2.6.2	Trial-based analysis . . . . .	36
2.6.3	General Information plots . . . . .	38
2.6.4	Information in fine vs. coarse distinctions . . . . .	43
2.6.5	Information in spike timing code vs. spike count code . . . . .	46
2.6.6	Spontaneous activity . . . . .	49
2.7	Discussions . . . . .	49
2.7.1	Validity of results . . . . .	49
2.7.2	Discussion of results . . . . .	50
2.7.3	Future Work . . . . .	50
2.7.4	Summary . . . . .	53
2.8	Other Discussion . . . . .	54
3	CORTICAL OSCILLATIONS IN CSD ACROSS V1 LAMINA	55
3.1	Introduction . . . . .	55
3.2	Experimental Methods . . . . .	56
3.2.1	Ethics Statement . . . . .	56

3.2.2	Anesthesia for Neurophysiology Experiments . . . . .	56
3.2.3	Visual Stimulation . . . . .	57
3.2.4	Neurophysiology Data Collection and Analysis . . . . .	57
3.2.5	Luminosity Function . . . . .	58
3.2.6	Artefact Removal . . . . .	59
3.2.7	Current Source Density . . . . .	59
3.2.8	Multiunit Activity . . . . .	59
3.2.9	Receptive Field Locations . . . . .	60
3.2.10	Identification of Cortical Laminae . . . . .	60
3.2.11	Identification of cortical laminae . . . . .	60
3.2.12	Power as a function of depth and frequency . . . . .	61
3.2.13	Information as a function of depth and frequency . . . . .	61
3.2.14	Cortical Distribution of Power . . . . .	62
3.2.15	Information Redundancy . . . . .	62
3.2.16	Information at about Spatial Components . . . . .	62
3.2.17	Information about Fine and Coarse Luminance Changes . . . . .	63
3.2.18	Information lag between granular and infragranular regions . . . . .	63
3.2.19	Cross-Frequency Phase-Amplitude Coupling . . . . .	63
3.3	Results . . . . .	64
3.3.1	Distribution of information across depth and frequency . . . . .	64
3.3.2	Information redundancy between frequencies . . . . .	67
3.3.3	Information redundancy across depth . . . . .	67
3.3.4	Information about spatial frequency components of visual stimulus . . . . .	69
3.3.5	Layer 1 6 Hz to 170 Hz amplitude is coupled to L5 4 Hz to 16 Hz phase . . . . .	73
3.4	Discussion . . . . .	73
4	GENERAL DISCUSSION	78
	BIBLIOGRAPHY	79

## LIST OF FIGURES

---

Figure 2.1	Experimental procedure. 1: The monkey fixates upon a central spot. 2: A sample stimulus in the form of either a Gabor patch or a sine grating is presented at the pedestal contrast (e.g. 30 %). 3: Blank sample-test interval. 4: Test stimulus presented. 5: Blank test-target interval. 6: Two target stimuli appear to the left and right of the test stimulus location, signalling that the subject is allowed make a saccade to their chosen target. Subject's overall objective is to make a comparison between the contrasts of the sample and test stimuli (presented during steps 2 and 4, respectively): if the test stimulus was of a higher contrast (e.g. 32 %) than the pedestal contrast, they should saccade to the white target, otherwise if the second patch was of lower contrast (e.g. 28 %), they should saccade to the black target (step 6). Durations shown are approximates as the actual durations vary slightly depending on the animal: see Table. 2.2 for more accurate values. . . . .	10
Figure 2.2	The experimental task during both the non-roving and roving stages. The analysis contained in this report concerns only the first, non-roving, stage. . . . .	12
Figure 2.3	Crosses: Histogram of spikes in bins of width 0.030716 ms. 2.3a: monkey 1 ( <b>M<sub>1</sub></b> ) extrastriate visual cortex area V <sub>4</sub> ( <b>V<sub>4</sub></b> ), channel 4, session 336; a channel and session very strongly exhibiting the monitor artifact. 2.3b: monkey 2 ( <b>M<sub>2</sub></b> ) <b>V<sub>4</sub></b> , channel 41, session 31; a dataset where the artifact is not present. 2.3c: <b>M<sub>2</sub></b> primary visual cortex ( <b>V<sub>1</sub></b> ), channel 9, session 51; data strongly presenting the monitor artifact. 2.3d: <b>M<sub>2</sub> V<sub>1</sub></b> , channel 9, session 56; data mildly showing the effect of the monitor artifact. Green vertical lines: extremities of search window (contained between the lines). Cyan vertical lines: extremities of mean region. Blue line: mean of the data (excluding area between cyan lines). Red lines: mean plus 1, 2 and 3 standard deviations respectively. Gray background: Spikes marked as contaminated and scheduled for redaction. . . . .	15

Figure 2.4	Spikes count during around 500 ms of presentation of the 27 % contrast stimulus, for two channels plotted against each other. There is clearly a weak correlation between the two data-series, but one trial is an outlier with a large number of measured spikes for both channels due to a correlated source of noise in the raw signal. The outlier will cause the correlation to be measured anomalously high. . . . .	19
Figure 2.5	Change in noise correlations with learning. (a): V <sub>1</sub> . (b): V <sub>4</sub> . Blue: M <sub>1</sub> . Red: M <sub>2</sub> . On the x-axis, the number of sessions since the animal began training in the part of the visual field retinotopic to the recording site is shown. Line: pearson r coefficient, averaged across the possible pairings between channels for each of the 14 trial conditions. Shaded region indicates one standard deviation from mean. . . . .	20
Figure 2.6	Distribution of the noise correlations for the pairings across all conditions. Two sessions, one at the beginning of training (black) and one at the end of training (red) are shown for comparative purposes. (a): M <sub>1</sub> V <sub>1</sub> ; sessions 343 and 354. (b): M <sub>2</sub> V <sub>1</sub> ; sessions 51 and 71. (c): M <sub>1</sub> V <sub>4</sub> ; sessions 307 and 338. (d): M <sub>2</sub> V <sub>4</sub> ; sessions 27 and 46. . . . .	22
Figure 2.7	Noise correlations for each pair, meaned across the 14 conditions. (a): M <sub>1</sub> V <sub>1</sub> . (b): M <sub>2</sub> V <sub>1</sub> . (c): M <sub>1</sub> V <sub>4</sub> . (d): M <sub>2</sub> V <sub>4</sub> . Colour is assigned by sorting the pairs into ascending order for one of the sessions near the middle of the training period. The degree of session-to-session correlation of the noise correlation can hence be inferred by visual inspection. . . . .	23
Figure 2.8	(a): Distribution of decoder performance based on spike rate for individual channels, sorted by performance. (b): Decoding performance versus number of channels included in spiking data. Channels were added one at a time, chosen so they maximise the decoder performance for that number of channels whilst keeping all the channels which had come before. . . . .	28

Figure 2.9	Decoding analysis for <b>V<sub>4</sub></b> . Performance of behavioural and decoding predictors by session, averaged across all conditions. Left panels: <b>M<sub>1</sub></b> . Right: <b>M<sub>2</sub></b> . Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis, ‘proportion correct’ is the proportion of trials for which the response is the same as the target. This is presented for behavioural performance (black), decoder performance (blue), and decoder performance when trials are shuffled, destroying noise correlations (red; see text). . . . .	29
Figure 2.10	Decoding analysis for <b>V<sub>1</sub></b> . Performance of behavioural and decoding predictors by session, averaged across all conditions. Left panels: <b>M<sub>1</sub></b> . Right: <b>M<sub>2</sub></b> . Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis, ‘proportion correct’ is the proportion of trials for which the response is the same as the target. This is presented for behavioural performance (black), decoder performance (blue), and decoder performance when trials are shuffled, destroying noise correlations (red; see text). . . . .	29
Figure 2.11	Decoding analysis for <b>V<sub>4</sub></b> . Trial-to-trial agreement between behavioural and decoding predictors. Left panels: <b>M<sub>1</sub></b> . Right: <b>M<sub>2</sub></b> . Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis is the proportion of trials for which the response is the same as the behavioural response. The agreement between behaviour and decoding (unshuffled only) is presented alongside the null hypothesis of completely independent binomial distributions. The dashed line indicates the 95% confidence interval of the null hypothesis. . . . .	30
Figure 2.12	Decoding analysis for <b>V<sub>1</sub></b> . Trial-to-trial agreement between behavioural and decoding predictors. Left panels: <b>M<sub>1</sub></b> . Right: <b>M<sub>2</sub></b> . Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis is the proportion of trials for which the response is the same as the behavioural response. The agreement between behaviour and decoding (unshuffled only) is presented alongside the null hypothesis of completely independent binomial distributions. The dashed line indicates the 95% confidence interval of the null hypothesis. . . . .	31

Figure 2.13	Mutual information between the test stimulus and the neural activity during test presentation. The mutual information with the test stimulus is taken for a spike timing based code for a 20 ms window of spiking activity, sampled with the start of the window offset ( $y$ -axis) from 0 ms up to 500 ms after test stimulus onset (which is slightly more than 20 ms before test stimulus offset). The sampling is in intervals of 5 ms, so any 4 adjacent squares within each session are highly correlated. The recording session number for the data is given along the $x$ -axis, and the number of days the animal has been trained for increases from left to right. Average mutual information across all the channels is denoted by the pseudo-colour of each of the rectangular patches, centred around the $(x, y)$ co-ordinate to which the measurement relates. In each case the average is taken across all available channel data: <a href="#">2.13a</a> 23 channels, <a href="#">2.13b</a> 30 channels, <a href="#">2.13c</a> 25 channels, <a href="#">2.13d</a> 20 channels. No bias correction method was used. . . . .	<a href="#">33</a>
Figure 2.14	Mutual information seems to be anti-correlated with the number of trials in the session. The number of correctly responded trials in each session is plotted on top of the mean information in individual sessions. The mutual information was not corrected for bias, and averaged across the whole period of test stimulus presentation. . . . .	<a href="#">34</a>
Figure 2.15	Mutual information is inversely correlated with the number of trials in the session. The mutual information for a single session is averaged over all channels, and averaged over all times during stimulus presentation, and plotted against the reciprocal of the number of trials in the session ( $1/N$ ). Plug-in: without bias correction. Panzeri-Treves method ( <a href="#">PT</a> ): using the Panzeri-Treves method to correct for bias ( <a href="#">Panzeri and Treves, 1996</a> ). Quadratic Extrapolation method ( <a href="#">QE</a> ): using quadratic extrapolation for bias correction ( <a href="#">Strong et al., 1998</a> ). . . . .	<a href="#">35</a>

Figure 2.16	<b>M<sub>1</sub> V<sub>1</sub>:</b> Mutual information between the test stimulus and 20 ms of spiking activity, averaged across 23 channels. The PT bias correction method was used in all estimates of the information, and the modification to the spiking data to remove the artifact as described in Sec. 2.3.1 was performed. The neural code used in 2.16a, 2.16b is a spike count code, whilst in 2.16c and 2.16d it is a spike timing code where the 20 ms window was subdivided into 5 bins each of 4ms. In 2.16a and 2.16c the spike-train is taken from the test presentation part of the trial; for 2.16b and 2.16d the spike-train is taken from spontaneous pre-stimulus activity. The data is sampled in intervals of 350 trials in the x-direction and 5 ms in the y-direction, so there is significant correlation between any pair of pixels in the image with less than 4 pixels between them in either cartesian direction. . . . .	40
Figure 2.17	<b>M<sub>2</sub> V<sub>1</sub>:</b> Mutual information between the test stimulus and 20 ms of spiking activity, averaged across 25 channels. The PT bias correction method was used in all estimates of the information. Panels 2.17a–2.17d are the same as for Fig. 2.16. . . . .	41
Figure 2.18	<b>M<sub>2</sub> V<sub>4</sub>:</b> Mutual information between the test stimulus and 20 ms of spiking activity, averaged across 20 channels. The PT bias correction method was used in all estimates of the information. Panels 2.18a–2.18d are the same as for Fig. 2.16. . . . .	42
Figure 2.19	<b>V<sub>1</sub>:</b> Fine vs. coarse contrast differences. In the top panels, the six contrasts included are {5, 15, 22, 40, 50, 90}%; bottom panels {22, 25, 28, 32, 35, 40}%. An average of 100 trials per stimulus is used in each of these. Left panels are for M <sub>1</sub> , right are M <sub>2</sub> . In each case, mutual information between the six test stimuli and 20 ms of spiking activity was measured using a spike count code, and bias corrected using the PT method. . . . .	44
Figure 2.20	<b>V<sub>4</sub>:</b> Fine vs coarse contrast differences. In the top panels, the six contrasts included are {10, 15, 20, 40, 50, 60}%; bottom panels {27, 28, 29, 31, 32, 33}%. An average of 100 trials per stimulus is used in each of these. Left panels are for M <sub>1</sub> , right are M <sub>2</sub> . In each case, mutual information between the six test stimuli and 20 ms of spiking activity was measured using a spike count code, and bias corrected using the PT method. . . . .	45

Figure 2.21	<b>V<sub>1</sub></b> : Information in millisecond level spike timing. The information with time-wise shuffled bins was subtracted from information in the spike time code with a 20 ms window subdivided into 5 bins. Information was bias corrected using the PT method. Left panels: M <sub>1</sub> ; Right: M <sub>2</sub> . Top panels: all contrasts, {10, 15, 20, 25, 27, 28, 29, 31, 32, 33, 35, 40, 50, 60}%. Centre panels: {5, 15, 22, 40, 50, 90}%. Bottom panels: {22, 25, 28, 32, 35, 40}%. An average of 100 trials per stimulus is used in the analysis for each. . . . .	48
Figure 3.1	<i>Overview of data collection and example data.</i> A: Illustration of experimental recording setup, showing approximate locations of electrode contacts in relation to a Nissl-stained section of macaque V <sub>1</sub> cortex. Boundaries between cortical laminae are indicated with arrowheads. (Stain reproduced from Tyler et al. 1998, with permission.) (Note: Electrode width is not to scale.) B: Receptive field locations were consistent across the cortical depth. Location of receptive field for each cortical recording site was identified by reverse correlating the multi-unit activity (MUA) with the luminance changes of each pixel in the movie (session E07nm1). C: Example current source density (CSD) traces from simultaneous recordings at three cortical depths for eight repetitions of a movie fragment (session H05nm7). The data are split into three temporal frequency bands (4 Hz to 16 Hz, 28-44 Hz and 6 Hz to 170 Hz, see Methods). . . . .	65

Figure 3.2	<i>Distribution of visual stimulus information across both cortical depth and frequency.</i> A: Distribution of local field potential ( <b>LFP</b> ) power during stimulus presentation. Plot shows the geometric mean power over 6 sessions. Above, mean power within supragranular ( <b>SG</b> ), supragranular ( <b>G</b> ) and infragranular ( <b>IG</b> ) regions. Right, laminar distribution of <b>LFP</b> power in 4 Hz to 16 Hz and 6 Hz to 170 Hz frequency bands. B: Same as A, but distribution of <b>CSD</b> power instead of <b>LFP</b> power. C: Distribution of information about the stimulus contained in <b>LFP</b> power. Plot shows the mean information over 6 sessions. Above, mean information within <b>SG</b> , <b>G</b> and <b>IG</b> regions. Right, cortical distribution of information in the power in 4 Hz to 16 Hz and 6 Hz to 170 Hz frequency bands. D: Same as C, but for information in <b>CSD</b> power instead of <b>LFP</b> power. Note that the information (C+D) is distributed very differently from the <b>LFP</b> and <b>CSD</b> power. . . . .	66
Figure 3.3	<i>CSD information redundancy across frequency bands and laminae.</i> A: Median redundancy between pairs of frequencies over the 12 recording sites, averaged over 6 sessions. B: Redundancy between pairs of recording sites of the information in three frequency bands. Mean of 6 sessions. C: Average of cross-channel redundancy shown in B. Note, that while there is substantial redundancy within bands and between the 60 Hz to 170 Hz and 900 Hz to 3000 Hz bands, there is little redundancy between the 4 Hz to 16 Hz and 6 Hz to 170 Hz band, indicating independent coding. . . . .	68
Figure 3.4	<i>Extraction of spatially filtered luminance components.</i> The luminance of the original video (left) is fast-Fourier transformed in a $224 \times 224$ px square for each frame (top-left: fast-Fourier Transform ( <b>FFT</b> ) of “current frame”). The mask isolates bands of spatial frequencies that are one octave wide (Row 1), yielding the spatially filtered frames (Rows 2 and 3). The stimulus magnitude at each spatial frequency band was obtained by taking the luminance difference of successive frames (Row 4), taking its absolute value (Row 5), and averaging this within the receptive field (Row 6). . . . .	70



## LIST OF TABLES

Table 2.3

## LISTINGS

## ACRONYMS

ACh	acetylcholine
Ag	silver
BOLD	blood oxygen-level dependent
CSD	current source density
EEG	electroencephalography
FFT	fast-Fourier Transform
G	supragranular
IG	infragranular
L <sub>1</sub>	layer 1 of V <sub>1</sub>
L <sub>4</sub>	layer 4 of V <sub>1</sub>
L <sub>5</sub>	layer 5 of V <sub>1</sub>
L <sub>5A</sub>	layer 5A of V <sub>1</sub>

L6	layer 6 of $V_1$
LFP	local field potential
LGN	lateral geniculate nucleus
M1	monkey 1
M2	monkey 2
MEA	multi-electrode array
MRI	magnetic resonance imaging
MSTd	dorsal medial superior temporal area
MUA	multi-unit activity
NaCl	sodium chloride
NSB	Nemenman-Shafee-Bialek entropy estimation method
PSTH	peristimulus time histogram
PT	Panzeri-Treves method
QE	Quadratic Extrapolation method
RF	receptive field
RGC	retinal ganglion cell
SG	supragranular
$V_1$	primary visual cortex
$V_2$	visual area $V_2$
$V_4$	extrastriate visual cortex area $V_4$

## INTRODUCTION

---

### 1.1 ORGANISATION OF THE THESIS

## PERCEPTUAL LEARNING IN V<sub>1</sub> AND V<sub>4</sub>

---

In this chapter, we investigate the neural correlates of perceptual learning within the visual cortical regions primary visual cortex ( $V_1$ ) and extrastriate visual cortex area V<sub>4</sub> ( $V_4$ ).

Perceptual learning is the phenomena in which an individual becomes more adept at fine-grain discrimination of stimuli through repetitive stimulation by the stimulus class in question. Clearly, such changes in perceptual ability are mediated by changes within the brain, but it is not currently known which changes drive the increase of such perceptual abilities.

Moreover, a long-standing question within visual perceptual learning has been whether cortical changes are driven through bottom-up or top-down developments.

Under the bottom-up hypothesis, repetitive stimulation of similar stimuli causes  $V_1$  to change its self-organisation such its representations of these stimuli are more prominent. There is then less loss of this fine-detail information in the first layers of cortical processing, so better quality information about the task at hand can be offered up to the higher-level cortical regions which then make the classification decision with regards to each stimulus.

With the top-down hypothesis, demand for better classification performance from high-level (output) cortical regions triggers an increase in cortical feedback, and the release of neurotransmitters such as acetylcholine ( $ACh$ ), dopamine, or norepinephrine which are all associated with increasing in the rate of change in synaptic connection strengths within the cortical region where they are present.

Using multi-unit spiking data recorded from macaque  $V_1$  and  $V_4$ , recorded by Xing Chen within the lab of Alex Thiele, Newcastle University, I investigated these hypotheses by decoding the information about the sensory stimulus encoded in  $V_1$  and  $V_4$  and comparing the rate of change of this over the course of experimental training.

### 2.1 BACKGROUND

#### 2.1.1 *Perceptual Learning*

When an individual repeatedly performs a sensory perception task they will, over time, demonstrate an improvement in performance. If the task is repeated — fre-

quently and over the course of several weeks — until performance finally saturates, the effect can persist for months. This phenomenon is known as perceptual learning, and its duration sets it apart from shorter term effects such as sensitization (transient increase in sensitivity following a period of stimulation) and priming (change in perception of one stimulus immediately following a different, faint, stimulus).

For the purposes of studying perceptual learning, fine-grained discrimination tasks are appropriate; since they are intrinsically difficult, they can not be immediately solved and there is scope for improvement. For instance, an example of a typical task chosen by neuroscientists to when studying perceptual learning is that of discerning the difference between straight lines of very similar orientations, or the alignment offset between sets of straight lines, known as vernier acuity. If it is trained, perceptual learning can be exhibited across seemingly all sensory modalities; other tasks which have been used for experiments include depth perception, somatosensory spatial resolution, estimation of weight, and discrimination of pitch (Gilbert, 1994; Gilbert et al., 2001).

However, the improvements in sensory discrimination which are made through perceptual learning are highly specific to the task at hand. For instance, training for vernier acuity only gives improvements for stimuli with the same orientation ( $\pm 30^\circ$ ) and spatial frequency ( $\pm 1/2$  octave) (Fiorentini and Berardi, 1980; Poggio et al., 1991), and training on line separation yields no when the lines are replaced with dots (Poggio et al., 1992). Moreover, results are specific to the retinotopic location of the stimulus, with translation through  $<10^\circ$  from the training spot sufficient to remove the effects (Fiorentini and Berardi, 1980, 1981; Poggio et al., 1991; Karni and Sagi, 1991). This said, some studies have found a limited amount of effect-transfer to regions in the opposite hemisphere for timing-dependent tasks (Ball and Sekuler, 1987; Berardi et al., 1987).

There is still some contention over where the physiological changes which lead to perceptual learning are situated in the brain. Consequently, there are several competing models which attempt to explain how perceptual learning arises. The ‘early’ model hypothesises that improvements principally occur at a low level in the sensory cortex (Gilbert et al., 2001; Fahle, 2005). The ‘late’ model states that improvements are in higher level cortical areas related to decision making (Yu et al., 2004). Whilst according to the ‘reverse hierarchy model’, improvements are made first in higher level decision areas, and then these are propagated down the cortical hierarchy to lower levels via top-down feedback signals if the changes at higher levels are insufficient (Ahissar and Hochstein, 2004; Hochstein and Ahissar, 2002).

The most obvious of these is that we would expect the information contained in the population spiking activity to increases over time as perceptual learning occurs.

This is likely to involve an increase in information for some of the individual neurons, but not necessarily all. In line with previous experiments (Raiguel et al., 2006), I also expect to see more of a change in information for neurons in  $V_4$  than  $V_1$ , and also a greater change in the  $V_4$  neurons which are the most informative to begin with (Raiguel et al., 2006). In keeping with the reverse hierarchy model, learning should begin in  $V_4$  first before being propagated down to  $V_1$ , so one would expect to see distinct increases in the mutual information between the stimulus and  $V_4$  on a shorter timescale than between  $V_1$  and  $V_4$ .

Perceptual learning is thought to be connected to cortical remapping and reorganisation in response to similar stimuli (Dinse et al., 2003; Pleger et al., 2003; Polley et al., 2006). In such experiments, the region of the cortex coding for the stimulus is seen to expand. Some researchers in this field have suggested that perceptual learning might be the mechanism which underpins all adult plasticity in the sensory and association cortices (Gilbert et al., 2001).

Neural changes correlated with perceptual learning have been observed at many levels of the cortical hierarchy. Studies have found changes in the orientation tuning curves of neurons in both  $V_1$  (Schoups et al., 2001) and  $V_4$  (Yang and Maunsell, 2004; Raiguel et al., 2006), however the effects are greater in  $V_4$  than in  $V_1$  (Raiguel et al., 2006), and not all studies find neural changes in  $V_1$  and visual area V2 ( $V_2$ ) which relate to perceptual learning, even when the subject has demonstrated psychometric improvement in the task (Ghose et al., 2002).

Due to the specificity of perceptual learning, only neurons in the retinotopic area where the stimulus falls are affected.

When the properties of individual neurons have been observed to change during perceptual learning, their tuning curves for task-relevant features have become sharper (i.e. the bandwidth narrows). Under activity-based models of neural information processing, this will provide more information about the task-relevant stimulus property if it falls on the steeper slope of the tuning curve. Studies have also shown that the effect of perceptual learning is most pronounced on the most relevant neurons from the perspective of information conveyed (Raiguel et al., 2006).

Since all neurons in the visual system have contrast tuning to some degree, one might think a contrast discrimination task a good choice for a perceptual learning study. However, perceptual learning has proven unreliable for such discrimination problems, possibly because contrast sensitivity is already overtrained due to its importance in low-light conditions. Better results have sometimes been found if the contrast test stimulus is accompanied with flanking stimuli (Adini et al., 2002), a phenomenon known as context-dependent learning, though other studies have found

learning occurs at the same rate both with and without flankers (Yu et al., 2004), despite nearly identical setup between the experiments with the conflicting two results.

Since temporal coding, in particular response latency, has been found to be important for subtle contrast differences (Reich et al., 2001; Arabzadeh et al., 2006), I hypothesise that the amount of information in the temporal coding of the spiking data will have increased above and beyond any increase in the information contained in the firing rates alone. Furthermore, I expect to see that response latencies become more stimulus dependent, conveying an increasing amount of information about the stimulus contrast.

Furthermore, since these studies (Reich et al., 2001; Arabzadeh et al., 2006) also found the information contained within firing rate alone was sufficient for gross discrimination of contrast, I hypothesise that information in the latency and temporal code will only increase significantly for test stimuli close in contrast to the sample stimulus (see the following section for an explanation of the experimental setup).

### *2.1.2 Information Theory, and its applications within Neuroscience*

A common method used in neuroscience is to record the extracellular activity of a single neuron under different conditions. Frequently, the approach used is to take many recordings of the same neuron for the same condition, and then take the average across these repetitions (“trials”) to reduce the effects of neuronal variability (also referred to as noise), producing a peristimulus time histogram (PSTH), for instance. However, this is not how the brain processes stimuli — it has access to many neurons simultaneously, but only a single sample of each of these is instantiated at a time since trials are sequential and not concurrent. In contrast to this, using information theory as part of our methodology to study extracellular recordings allows us to consider how much information about a stimulus is present in the system in a the activity of a single-trial.

When applying information theory to neuronal data, we treat the brain as a communication channel, transmitting information about sensory input.

In the perspective of sensory recordings, the different conditions used on the trial are typically different stimuli, and the extracellular recordings provide us with the neuron’s response to the stimuli. For such an experimental setup, let us assume that on each trial the stimulus  $s$  is selected at random with probability  $P(s)$  from a set of stimuli  $S$ , containing  $S$  unique stimuli. For our purposes, we will be considering response given by the neuron in the spike-train it elicits, though an information approach can be performed on data from the local field potential (LFP) of extra-

cellular recordings, or collected by blood oxygen-level dependent (**BOLD**) signal or electroencephalography (**EEG**) (Magri et al., 2009; Quiroga and Panzeri, 2009).

Using information theory, we can also investigate the nature of the neural code used by individual neurons and populations of neurons (Optican and Richmond, 1987). For example, we might look at whether a neuron is conveying information in the millisecond level timing of its spikes, or if all the information is conveyed in its firing rate. This can be done by choosing a way of quantifying the neural activity which reflects what we think are its most salient properties, and then comparing the how much information can be extracted for different codes. Typically (Quiroga and Panzeri, 2009; Brasselet et al., 2012; Panzeri et al., 2007; Arabzadeh et al., 2006; Strong et al., 1998), a certain poststimulus time window of duration  $t \in [20, 40]$  is chosen and a neural code is constructed which forms a discrete, multi-dimensional array  $r = \{r_1, \dots, r_L\}$  of dimension  $L$ , because a discrete response of this form is appropriate for performing an information theoretic analysis on. To look at the information given by the firing rate of a single neuron, we might use a spike-count code where  $L = 1$  and  $r$  is equal to the number of spikes elicited in the window  $t$ . Similarly, to look at the information contained in the firing rate of a many neurons, we would use a spike-count code where  $r_i$  is equal to the number of spikes elicited in the window  $t$  for spiketrain of the  $i$ -th neuron, and let  $L$  equal the number of neurons to be studied. In comparison, to look at the information contained in the millisecond level spike timing of a single neurons, we would divide the time window into  $L$  bins of length  $\Delta t = t/L$  such that  $\Delta t$  is the assumed time precision of the code, and set  $r_i$  to be the number of spikes elicited in  $i$ -th time-bin.

Having chosen a neural code, we can let  $\mathbf{R}$  denote the set of possible response arrays. The relationship between the distribution of responses and stimuli is evaluated by first quantifying the variability of the responses. This can be done with the entropy (Shannon, 1948) of the responses, and we define the *response entropy* to be

$$H(\mathbf{R}) = - \sum_r P(r) \log_2 P(r) \quad (2.1)$$

where  $P(r)$  is the probability of observing the response  $r$  on any trial regardless of the stimulus. However, the responses given by neurons are “noisy”, so they do not give the same response on every trial even if the same stimulus is presented. Consequently, we must also consider the variability due to noise by computing the *noise entropy*, defined as

$$H(\mathbf{R}|\mathbf{S}) = - \sum_{r,s} P(s)P(r|s) \log_2 P(r|s). \quad (2.2)$$

The information about the stimulus which is transmitted in the response is then given by the difference of these, and the *mutual information* between stimulus and response is given by

$$I(\mathbf{S}; \mathbf{R}) = H(\mathbf{R}) - H(\mathbf{R}|\mathbf{S}) = \sum_{r,s} P(r,s) \log_2 \frac{P(r|s)}{P(r)}. \quad (2.3)$$

The mutual information can conceptualised how much an independent observer can expect their uncertainty in the stimulus  $s$  presented on a single trial to be reduced by if they were to observe the neural response. When using base two logarithms, the mutual information is measured in bits, where gaining 1 bit of information about something means a halving in uncertainty about it. The mutual information is zero if and only if responses are completely independent of stimuli. We will frequently abbreviate mutual information to just information.

When working with experimental data, the probabilities  $P(s)$ ,  $P(r)$  and  $P(r|s)$  must be estimated from the available data. This presents a major problem, because precise values for the probabilities can only be found exactly from their frequencies in the data if there is an infinite amount number of trials available, and real-world experiments (somewhat inconveniently) contain only a limited number of trials. The estimated probabilities are subject to statistical error, leading to an associated systematic error (bias) and statistical variance in the estimates of the entropies and mutual information. The bias in particular is an issue, causing the mutual information to be upwardly biased, which can lead to incorrect conclusions if not corrected. Conceptually, this is because finite sampling can lead to spurious differences in the response distributions, making them seem more discriminable than they really are. We refer to the bias uncorrected mutual information as  $I_{\text{uncorrected}}(\mathbf{S}; \mathbf{R})$ .

Fortunately, several techniques exist to correct for the bias. Several of these bias correction methods focus on expanding out the measured information as a power series (Miller, 1955; Treves and Panzeri, 1995) in terms of  $1/N$ , where  $N$  is the number of trials in the dataset, though technically the relationship with the power series only holds in the asymptotic sampling regime with a very large number of trials. The first term in the bias, proportional to  $1/N$ , has a coefficient which depends only on the number of stimuli,  $S$ , and possible responses  $\bar{R}$ . However,  $\bar{R} \neq R$  because many responses which are theoretically possible by the construction of the response code may be in fact be impossible to generate. Furthermore,  $\bar{R}$  cannot be found simply by looking at the number of unique responses in the dataset, since low probability responses may not have occurred in the finite number of trials sampled, leading to an underestimation of  $\bar{R}$ . Consequently, one approach to bias correction, a Panzeri-Treves method (PT) (Panzeri and Treves, 1996), uses a Bayesian procedure to estimate

the true number of possible responses and then subtracts the leading term of the bias from the mutual information. A second method of correcting for the bias which makes use of the power series expansion is a Quadratic Extrapolation method ([QE](#)) method ([Strong et al., 1998](#)). Here, the uncorrected mutual information is assumed to be well approximated by

$$I_{\text{uncorrected}}(\mathbf{S}; \mathbf{R}) = I_{\text{true}}(\mathbf{S}; \mathbf{R}) + \frac{a}{N} + \frac{b}{N^2},$$

with the free parameters  $a$  and  $b$  found by computing the information content with fractions of the full available dataset (i.e. using  $N/2$  and  $N/2$  trials).

Other bias correction methods which do not utilise the power series expansion include a Nemenman-Shafee-Bialek entropy estimation method ([NSB](#)). This uses a Bayesian inference approach to entropy estimation with a specially chosen prior probability distribution to make the prior expected entropy distribution uniform ([Nemenman et al., 2004](#)).

A completely different method of bias correction which works for responses with dimension  $L > 1$ , is to compute an estimate of the information known as  $I_{\text{sh}}$ . This is found ([Montemurro et al., 2007](#)) by computing two additional terms:  $H_{\text{ind}}(\mathbf{R}|\mathbf{S})$ , the noise entropy estimate if the dimensions of  $r$  are assumed to be independent from one another; and  $H_{\text{sh}}(\mathbf{R}|\mathbf{S})$ , the noise entropy estimate when bins are shuffled along each dimension,  $r_i$ , of the response to generate pseudo-response arrays. The advantage of this is  $H_{\text{ind}}(\mathbf{R}|\mathbf{S})$  and  $H_{\text{sh}}(\mathbf{R}|\mathbf{S})$  should be equal, but  $H_{\text{sh}}(\mathbf{R}|\mathbf{S})$  has a bias around the same magnitude as  $H(\mathbf{R}|\mathbf{S})$ , so we can compute

$$I_{\text{sh}}(\mathbf{S}; \mathbf{R}) = H(\mathbf{R}) - H_{\text{ind}}(\mathbf{R}|\mathbf{S}) + H_{\text{sh}}(\mathbf{R}|\mathbf{S}) - H(\mathbf{R}|\mathbf{S}),$$

which has a much smaller bias than  $I(\mathbf{S}; \mathbf{R})$ .

Both a [PT](#) and [QE](#) bias correction methods give similar approximations to the true information, whilst [NSB](#) outperforms them with a less biased estimate ([Panzeri et al., 2007](#)). However, [NSB](#) is very computationally intensive ([Panzeri et al., 2007](#)), so we will not be making use of it in the presented body work. All these bias correction methods tend to make a trade off between variability and bias, introducing more terms and hence more variability to reduce the size of bias term.

## 2.2 EXPERIMENTAL METHODS

The experimental data was acquired by Xing Chen, under the supervision of Alexander Thiele, Institute of Neuroscience, Newcastle University.

Extracellular recordings were taken from multi-electrode arrays (**MEAs**) chronically implanted in both **V<sub>1</sub>** and **V<sub>4</sub>** of two macaque monkeys, which we will refer to as monkey **1** (**M<sub>1</sub>**) and monkey **2** (**M<sub>2</sub>**). For each monkey, the two implants were undertaken simultaneously by a fully trained brain surgeon with prior experience with these implants in animals.

The electrode contacts within each **MEA** had a separation distance of XXX, which is sufficiently distant from each other such that neurons can be close enough at most one electrode contact for their spiking activity to be detected.

The number of electrode contact points (channels) which were recorded from for each **MEA** is shown in Table 2.1.

Table 2.1: Number of channels recorded from for each of the monkeys and brain regions. The number of channels on the implanted **MEA** was the same for all 4 implants, but not all electrodes were viable to record from after implantation.

Animal	Region	Number of channels
<b>M<sub>1</sub></b>	<b>V<sub>4</sub></b>	30
	<b>V<sub>1</sub></b>	23
<b>M<sub>2</sub></b>	<b>V<sub>4</sub></b>	20
	<b>V<sub>1</sub></b>	25

The receptive fields of the neurons in the two implant locations did not overlap retinotopically for either monkey. Consequently, the experimental procedure was performed twice for each animal, first with stimuli located within the visual field so as to stimulate the neurons recorded from the **V<sub>4</sub>** implant, and then for the **V<sub>1</sub>** implant.

The experimental design has been described previously by [Chen et al. \(2013, 2014\)](#). The experiment comprised 4 stages, illustrated in Fig. 2.1, each of which conform to the same overarching structure. Each experiment is composed of a series of trials. On each trial, the monkey must first fixate their gaze on a fixation point, which is a black cross on a grey background with 50 % contrast displayed on a monitor in front of the monkey. For **M<sub>1</sub>**, the refresh rate of the monitor was 85 Hz, whilst for **M<sub>2</sub>** it was 75 Hz.

After the monkey has been fixated on the fixation point for a randomised period of time  $t_1$ , a grey-scale Gabor function with a contrast  $C_{\text{sample}}$  appears on the screen. We refer to this as the *sample stimulus*, and it appears in the animal's field of vision such that it is retinotopic to the location of the implanted brain region. This Gabor stimulus remains on screen for  $t_2$ , after which it vanishes. There follows a delay of  $t_3$ , during which the animal must continue to fixate on the fixation point. After this, a second Gabor stimulus with contrast  $C_{\text{test}}$  appears in the same location as the sample

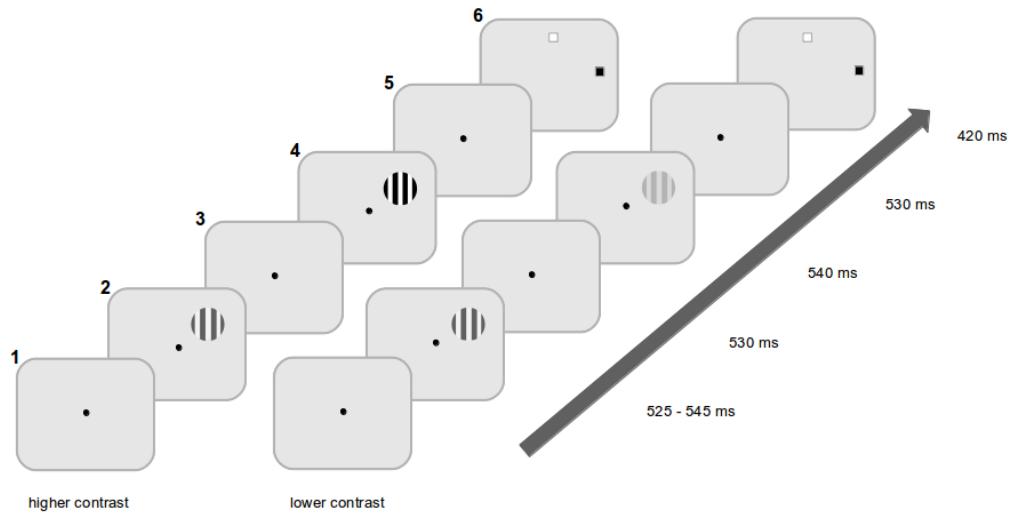


Figure 2.1: Experimental procedure. 1: The monkey fixates upon a central spot. 2: A sample stimulus in the form of either a Gabor patch or a sine grating is presented at the pedestal contrast (e.g. 30 %). 3: Blank sample-test interval. 4: Test stimulus presented. 5: Blank test-target interval. 6: Two target stimuli appear to the left and right of the test stimulus location, signalling that the subject is allowed make a saccade to their chosen target. Subject's overall objective is to make a comparison between the contrasts of the sample and test stimuli (presented during steps 2 and 4, respectively): if the test stimulus was of a higher contrast (e.g. 32 %) than the pedestal contrast, they should saccade to the white target, otherwise if the second patch was of lower contrast (e.g. 28 %), they should saccade to the black target (step 6). Durations shown are approximates as the actual durations vary slightly depending on the animal: see Table. 2.2 for more accurate values.

stimulus, also for  $t_4$ . We refer to this as the test stimulus. There is then a delay of  $t_5$  before the fixation target vanishes and a pair of black and white response target squares appear in a region away from the stimulus location. The monkey is tasked with saccading to the white target if the test stimulus has lower contrast than the sample, or the black target if test has higher contrast. If the monkey responds correctly, it is immediately given a water reward. The experiment is thus a two-alternate forced-choice experiment. After the animal has given its response, the next trial begins without delay, and the fixation target again appears alone on the screen until the monkey has fixated for some randomised duration  $t_1$ . The durations used for each part of the experiment are given in Table. 2.2.

Table 2.2: Durations of each section of the trials.  $t_1$ : interval from fixation-start until sample presentation.  $t_2$ : sample presentation.  $t_3$ : sample-test stimulus interval.  $t_4$ : test presentation.  $t_5$ : test-targets interval. Differing monitor refresh rates for the two monkeys results in different durations because stimuli can only be presented at the start of the monitor refresh cycle. N.B. A random duration for  $t_3$  was used when the experiment was performed on M1 in V4, but this was changed to a static delay for data collected later.

Animal	Region	Duration (ms)				
		$t_1$	$t_2$	$t_3$	$t_4$	$t_5$
M1	V4	[530.872, 545.526]	529.275	[539.720, 1058.673]	529.275	423.475
	V1	[525.803, 538.983]	529.275	541.164	529.275	423.475
M2	V4	[526.295, 540.641]	529.275	546.632	533.176	426.578
	V1	[525.834, 540.702]	533.176	546.570	533.176	426.640

Following a period of rest after surgery, during which the implants should settle into stable locations, the animals are pre-trained to sit in the chair and fixate on the fixation point. They are then trained in the procedure of the experiment, whilst recordings are taken from V4. Their preliminary training involves a grey disc as the sample, whilst the test is either white or black. This then changes to be a set of Gabor functions.

After the preliminary training, the main perceptual learning experiment is performed, first for V4, then V1. In the main experiment, we keep the contrast of the sample as  $C_{\text{sample}} = 30\%$  throughout.

The test contrast is chosen from a set of 14 contrasts: (10, 15, 20, 25, 27, 28, 29, 31, 32, 33, 35, 40, 50 and 60) % for V1 and (5, 10, 15, 20, 22, 25, 28, 32, 35, 40, 45, 50, 60 and 90) % for V4, for both of the animals. These groups are chosen such that the animal has a similar initial accuracy for both V1 and V4. We note that half of the contrasts are above and half below the sample contrast of 30 %. The contrast is not chosen

randomly at the start of each trial; instead a batch of trials containing a set number of each test contrast is placed into sequence all at once. If the monkey responds incorrectly at the end of the trial, this contrast is added to a list to be repeated at the end of the batch. At the end of the batch, another batch of trials begins in a similar fashion.

The duration of daily training was typically limited by the animal's desire to perform the task. On some days the animal was reluctant to train and only 250 trials could be conducted, whilst on other days the animal would be willing to train for three hours and complete up to 1250 trials. As expected, the monkey exhibited perceptual learning and their psychometric performance increased each day for around 20 days before reaching a plateau. Once performance had not improved for 5 consecutive days, the experiment progressed to the next stage.

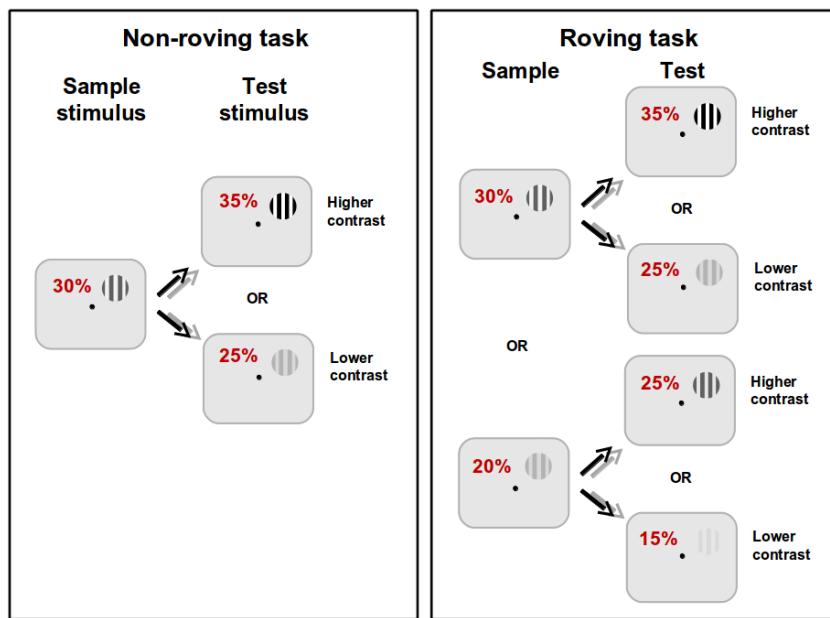


Figure 2.2: The experimental task during both the non-roving and roving stages. The analysis contained in this report concerns only the first, non-roving, stage.

In this study, we will focus on analysing data from the non-roving perceptual learning task as described above. However, it should be noted that once the animals had completed the non-roving stage of the experiment, they progressed to further stages involving a roving task. In the roving task, the sample contrast was chosen at random with  $C_{\text{sample}} \in \{20, 30, 40\}\%$ . Subsequently, once the test stimulus had been shown and the targets appeared, the animal was tasked with responding by comparing the test contrast to whichever sample contrast was presented on that particular trial. During this task, the test contrast would be selected from a set of 12 contrasts specific to

each of the three sample stimuli, such that 6 test stimuli were higher and 6 lower in contrast than the sample. The difference between the non-roving and roving stages is illustrated in Fig. 2.2.

As mentioned in Sec. 2.1.1, some previous research has found flankers around the stimuli are necessary to induce perceptual learning. During this experiment, the researchers found flankers were not necessary for perceptual learning during the first stage, but to subsequently learn to the roving task, flankers around the stimuli were necessary.

## 2.3 METHODS

### 2.3.1 *Artifact elimination*

Looking at rasters generated to include all the trials across all the sessions, an anomaly was apparent. For several of the channels in the data from each of the brain regions of each of the animals, some sessions have spikes which align at the same time relative to the stimulus onset. The temporal alignment is very precise, indicating the effect not part of the animal’s brain activity but instead due to an external source influencing the detected signal in the recordings. Furthermore, the “lines” occurred at regularly spaced intervals, with a period very nearly equal to the refresh rate of the monitor. The spikes across multiple trials line up like this because the experimental equipment will only begin presenting a stimulus when the first pixel on the monitor is being updated, and we of course normalise for the stimulus presentation time across multiple trials.

An empirical estimate of the artifact periodicity was found by choosing an arbitrary channel which strongly expresses the artifact and measuring by eye the duration of the completed artifact cycles within 530 ms (44 for M<sub>1</sub>, 39 for M<sub>2</sub>). Since the artifact is tightly localised in time, this could be done with relatively high accuracy. For M<sub>1</sub>, the period was estimated to be 85.023(1) Hz, whilst for M<sub>2</sub> it was 75.023(1) Hz. The discrepancies from the programmed monitor refresh rates of 85 Hz and 75 Hz respectively can be put down to the specific electronic circuitry used, perhaps issuing the command to the monitor. The discrepancy is small enough to be of little consequence, except we will need 5 sf of accuracy rather than 2 sf in the following treatment of the artifact. Why it should be out by exactly 0.023 Hz for both animals remains unclear, since this corresponds to the refresh cycle running 0.0031 ms fast for M<sub>1</sub> and 0.0042 ms fast for M<sub>2</sub>. An even bigger mystery is how the artifact has made its way into the recordings. Although it remains possible that the cause is some other piece of equipment also locked to the same refresh cycle, in the following we will refer to

this artifact as the “monitor artifact”. When a collection of data points exhibits the monitor artifact, we refer to the data as “contaminated”.

From the rasters, it seemed as if three-quarters of the channels in **M<sub>1</sub>** **V<sub>4</sub>** were contaminated for at least one session; over two-thirds of the channels for **M<sub>1</sub>** **V<sub>1</sub>** and **M<sub>2</sub>** **V<sub>4</sub>** were contaminated for at least one session; and around a third of the channels for **M<sub>2</sub>** **V<sub>4</sub>** were contaminated for at least one session. The contaminated channels include both high-quality channels with a lot of detected spikes, and low-quality channels with fewer detected spikes. For some of the lowest quality channels and sessions, the artificial spikes were clearly more numerous than the genuine ones. It was considered paramount that the effects of the artifact be corrected for.

To clean up the contamination, a more rigorous method of evaluating the problem was required. For the collection of spike times from a single channel across multiple trials during a single session, we perform the following steps:

1. Consider the set of all spike times where the visual stimulus has been the same for at least the last 150 ms.
2. From each spike time,  $t$ , subtract time of nearest stimulus onset/offset,  $T_{\text{onset}}$ . (Both onset and offset are synchronised with the monitor cycle, and the nearest one offers greatest accuracy.)

$$t \leftarrow t - T_{\text{onset}}$$

3. Take the modulo of the spike times with respect to the monitor period,  $\tau_m$  (11.7616 ms for **M<sub>1</sub>**, 13.3292 ms for **M<sub>2</sub>**).

$$t \leftarrow t \bmod \tau_m$$

4. Take a histogram of the spike times over bins with the width of the reciprocal of the sampling frequency of the spikes (sampling frequency 32 556.000 Hz; bin width 0.030 716 ms).

Conceptually, this is equivalent to stacking all the “lines” in the raster on top of each other and seeing how thick the resulting line is. When the visual stimulus has remained unchanged for at least 150 ms, the neurons in **V<sub>1</sub>** and **V<sub>4</sub>** settle down to a steady firing rate, so we expect there to be about the same number of spikes in each of these bins. However, as the monitor artifact increases the number of spikes at set intervals after the monitor refresh commences, there will be an increase in the number of spikes in these bins when the monitor artifact is manifest in the data.

As shown in Fig. 2-3, the analysis conforms to our expectations. However, doing the analysis in this more rigorous manner revealed the contamination was more

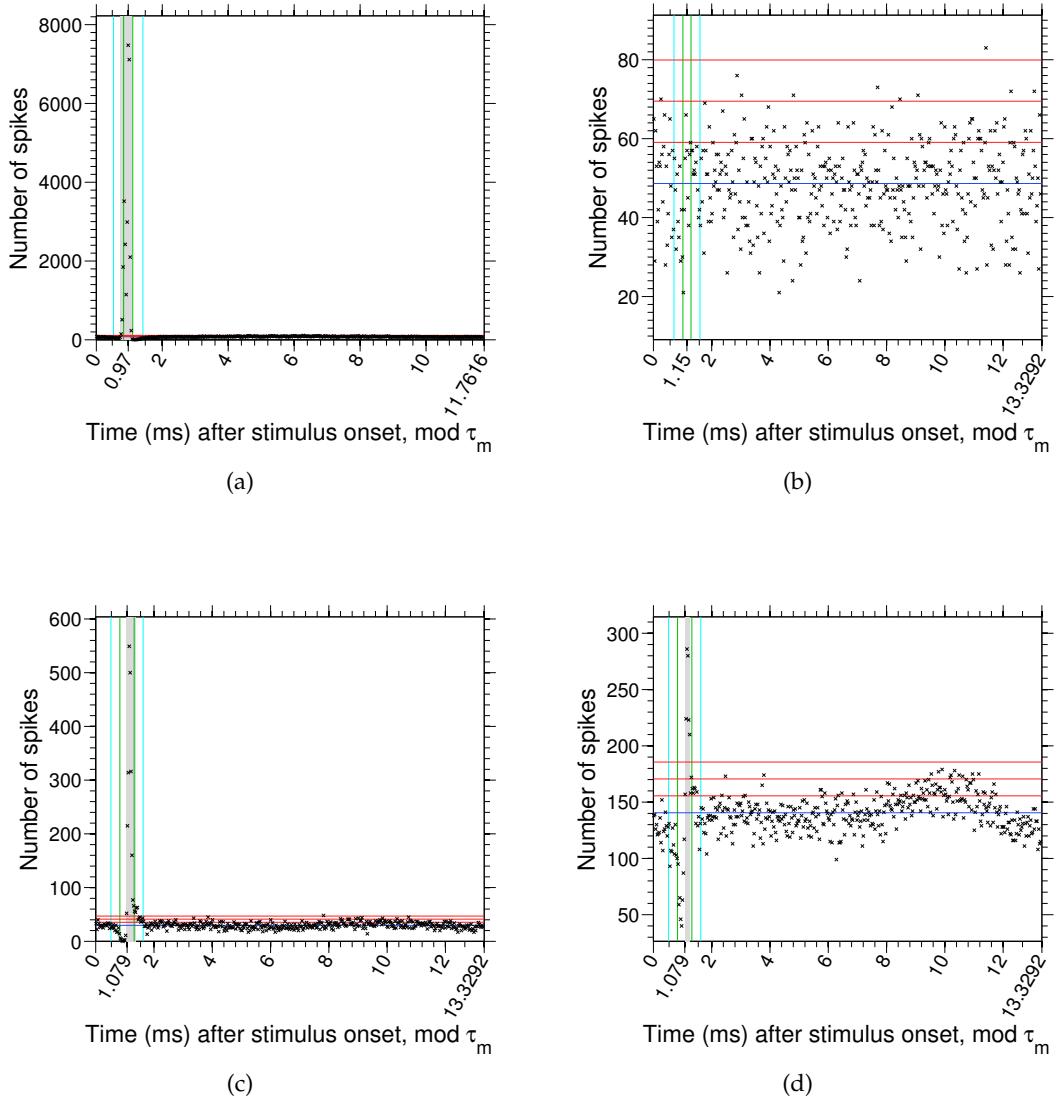


Figure 2.3: Crosses: Histogram of spikes in bins of width 0.030716 ms. 2.3a: M1 V<sub>4</sub>, channel 4, session 336; a channel and session very strongly exhibiting the monitor artifact. 2.3b: M2 V<sub>4</sub>, channel 41, session 31; a dataset where the artifact is not present. 2.3c: M2 V<sub>1</sub>, channel 9, session 51; data strongly presenting the monitor artifact. 2.3d: M2 V<sub>1</sub>, channel 9, session 56; data mildly showing the effect of the monitor artifact. Green vertical lines: extremities of search window (contained between the lines). Cyan vertical lines: extremities of mean region. Blue line: mean of the data (excluding area between cyan lines). Red lines: mean plus 1, 2 and 3 standard deviations respectively. Gray background: Spikes marked as contaminated and scheduled for redaction.

widespread than previously assumed. The histograms of spike times indicate that more sessions are contaminated than not, and only a couple of channels are completely clean for every session.

For the most part, the timing of the monitor artifact mod  $\tau_m$  is very reliable and it affects the same 6 or so bins whenever the data is contaminated. The contaminated part of each monitor cycle is thus restricted to at most 0.2 ms, and, assuming this period of time contains both genuine spikes and artificial spikes, if we simply delete all data points with this 0.2 ms window, we will lose at most less than 2 % of the genuine spikes. This level of data loss is not terribly significant, and justifiable for the gain in reliability of the remaining dataset, particularly for sessions where artificial spikes seem to be more common than genuine ones.

Table 2.3: Typical time of the peak in number of spikes due to the monitor artifact effect. The time is given relative to the start of the monitor refresh as given by the stimulus onset times. Monitor artifacts occur at  $t = t_m^* + k\tau_m$ , for  $k \in \mathbb{Z}$ . For channels recording from M2 V1, the artifact will occur at one of the two stated times throughout a given session, but which of the two varies from channel-to-channel for any individual session, and from session-to-session for any individual channel. Why this happens is unclear.

Animal	Region	Time of artifact peak $t_m^* \pm 0.01$ (ms)
M1	V1	0.95
	V4	0.97
M2	V1	either 0.97 or 1.19
	V4	1.15

However, it is possible to do better than this and to isolate and remove only the contaminated bins. As shown in Fig. 2.3, the effect of the artifact is greater for some sessions than others and From observations on the typical width time of the width and peak time (Table. 2.3) of the artifact, along with its periodicity ( $\tau_m$ ), the following method of removal was devised and implemented.

1. Perform the modulo  $\tau_m$  and take the histogram as described above.
2. Define the “search region” as bins which contain spike within  $t_m^* \pm 0.130$  ms, so we have one bin more than the anticipated artifact width in either direction. (For M2 V1, we use the range (0.968 - 0.130 , 1.190 + 0.130)ms.)
3. Find the mean,  $\mu$ , and standard deviation,  $\sigma$ , of all the bins which are at least 10 away (0.3 ms away) from the search region. We also exclude the final bin since  $0.030716 \bmod \tau_m \neq 0$  and it is only part-full.

4. If any one of the bins in the search region exceeds  $\mu + 3\sigma$ , we declare the session contaminated and proceed with the following steps. If not, it is declared intact and left as it is.
5. All bins in the search region which contain more than  $\mu + 3\sigma$  spikes are declared contaminated.
6. Let  $t_m$  denote the time of the centre of the bin in the search region containing the most spikes.
7. All bins which contain spikes within  $t_m \pm 0.1$  ms are placed in “quarantine”.
8. All quarantined bins which contain more than  $\mu + 2.5\sigma$  spikes are declared contaminated.
9. All bins between the first and last contaminated bins are declared contaminated.
10. Consider immediate neighbours of the first bin and if they are also contain more than  $\mu + 2.5\sigma$  spikes, they are declared contaminated.
11. Repeat the above step until either a neighbour is found which is not contaminated, or until a maximum of three bins have been added to the contaminated region.
12. Consider immediate neighbours of the last bin and if they are also contain more than  $\mu + 2.5\sigma$  spikes, they are declared contaminated.
13. Repeat the above step until either a neighbour is found which is not contaminated, or until a maximum of three bins have been added to the contaminated region.
14. Consider the full dataset of spikes for this particular channel and session.
15. For each spike, consider which bin its spike time would be arranged into.
16. If it is a contaminated bin, the spike is removed from the dataset.

This allows us to have prior assumptions about the location and width of the monitor artifact effect, but be flexible about exactly where it falls and which sets of bins are affected. It is possible that the dynamically targeted removal method may cause issues due to inconsistency in the data across different sessions, but there is a clearly a gain in the amount of preserved data and a gain in the reliability of the data which remains.

### 2.3.2 Misc

Neurolynx NSE files were loaded into MATLAB using the Neurolynx supplied code when operating on Windows, and an version 6 of an unofficial port for Linux (recommended by Neurolynx) otherwise.<sup>1</sup>

## 2.4 NOISE CORRELATIONS

### 2.4.1 Background

The firing rate of neurons in the sensory system will vary trial-to-trial, even if the stimulus presented is the same. The manner in which these responses fluctuate together is known as the noise correlation between neurons. If there are positive noise correlations, the changes in neural activity across a population will increase and decrease together. Long-standing theoretical work has indicated such correlations will hamper the ability of upstream neurons to accurately decode the stimulus, which is in keeping with experimental results demonstrating attention reduces noise correlations (Cohen and Maunsell, 2009). This fits with intuition — if stimulus A is encoded with one firing rate across a population and stimulus B with a slightly higher rate, a systematic increase or decrease in the population firing rates can easily cause A and B to be confused with one another.

However, recent theoretical results indicate that for a more realistic non-homogeneous population of neurons, neural correlations do not limit the population-wide information as the number of neurons increases (Oram et al., 1998; Averbeck et al., 2006; Ecker et al., 2011).

Other studies have indicated that noise correlations are beneficial to population encoding, provided that the structure of the correlations means the correlations are orthogonal to the direction in which stimuli are encoded.

A recent paper by Gu et al. (2011) on neurons recorded from the macaque dorsal dorsal medial superior temporal area (**MSTd**) before and after training in a head direction discrimination task, found that although pairwise noise correlations between neurons are reduced with training, this does not yield an increase in performance in a decoder. We will see if these findings hold here as well.

---

<sup>1</sup> Available at  
<http://www.urut.ch/new/serendipity/index.php?/pages/downloads.html>

### 2.4.2 Methods

The noise correlation was investigated by considering how strongly correlated the spike counts during test stimulus presentation.

To evaluate the strength of noise correlations, the Pearson  $r$  correlation coefficient was used. For two random variables  $X$  and  $Y$ , this is given by

$$r_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

and provides a version of the covariance between the  $X$  and  $Y$  which is normalised again their standard deviations. This means  $-1 \leq r \leq 1$  indicates how well  $X$  and  $Y$  fit a straight line relationship of either positive or negative slope, regardless of the gradient of the line, which could be changed by rescaling either  $X$  or  $Y$ . If  $r = \pm 1$ , there is a perfect linear relationship between  $X$  and  $Y$ , whilst if  $r = 0$  then  $X$  and  $Y$  are completely independent.

For each individual session, the Pearson  $r$  correlation coefficient was computed between the spike counts during test stimulus presentation between pairs of channels, across all the trials in the session.  $r$  was computed for each pair and each condition, and the mean was taken across all of these.

As noted previously, on some trials movements of the subject produced a recording artifact across multiple channels which was incorrectly registered as genuine spikes by the spike detection algorithm. When comparing the spike counts for a pair of channels, to trials with erroneously high detected spike rates for both channels, as shown in Fig. 2.4.

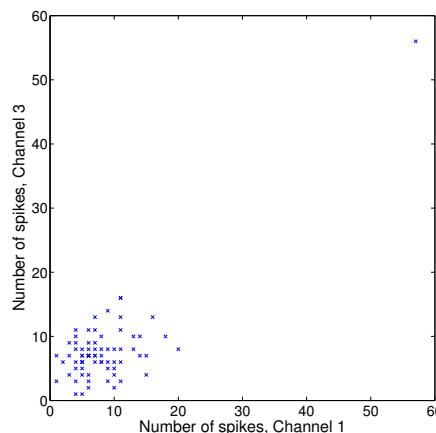


Figure 2.4: Spikes count during around 500 ms of presentation of the 27 % contrast stimulus, for two channels plotted against each other. There is clearly a weak correlation between the two data-series, but one trial is an outlier with a large number of measured spikes for both channels due to a correlated source of noise in the raw signal. The outlier will cause the correlation to be measured anomalously high.

To counter this problem, all trials where at least a quarter of the channels had a number of spikes more than 2.5 standard deviations above their mean spike count were removed. This should be effective at fixing the problem because the motion-triggered artifact occurs always effects multiple channels simultaneously with the same artifact.

### 2.4.3 Results

For both **V<sub>1</sub>** and **V<sub>4</sub>**, the average noise correlation between pairs of channels seem to remain stable for **M<sub>1</sub>** and decrease by a small margin for **M<sub>2</sub>** (see Fig. 2.5). The increase in noise correlation for **M<sub>1</sub> V<sub>1</sub>** for the last two sessions (session numbers 358 and 359) is due to a decline in data quality and an increase in noise from artificial sources. There is a large amount of variance in the noise correlations for different pairs, so the decline in mean correlation for **M<sub>2</sub>** does not seem very large considering the amount of variance.

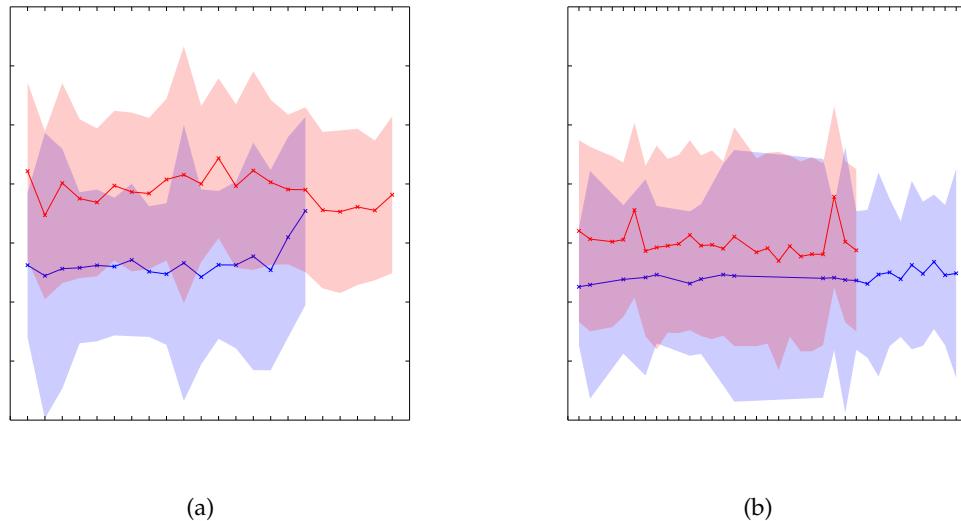


Figure 2.5: Change in noise correlations with learning. (a): **V<sub>1</sub>**. (b): **V<sub>4</sub>**. Blue: **M<sub>1</sub>**. Red: **M<sub>2</sub>**. On the x-axis, the number of sessions since the animal began training in the part of the visual field retinotopic to the recording site is shown. Line: pearson  $r$  coefficient, averaged across the possible pairings between channels for each of the 14 trial conditions. Shaded region indicates one standard deviation from mean.

Fig. 2.6 is intended to reproduce Figure 2C from Gu et al. (2011), with the distribution of  $r$  shown across pairs for one pre-training and one post-training session. These sessions were chosen from a restricted set of sessions which did not have problems with artificially high correlations from the motion artifact, and selected from this set such that they were as close to the start and end of the training period as possible.

However, this selection was made before the set of trials was redacted as mentioned in §2.4.2, so the sessions selected could possibly be made further apart.

The data presented in Fig. 2.6 shows that the distribution of noise correlation pairs does not move significantly for  $M_1 V_1$ , which is contrasted by a clear decrease for  $M_2 V_1$ . It should be noted though that the distribution for  $M_2 V_1$  begins higher than  $M_1$  and decreases to a similar value as  $M_1 V_1$ . For  $V_4$ , there is an increase in mean noise correlation for  $M_1$  and a decrease for  $M_2$ , though as  $M_2$  begins higher than  $M_1$  the two do tend toward one another.

Cherry-picking is a significant problem here, as there is sizable day-to-day variation in the noise correlation across the pairs. Choosing a session where there is more noise correlation than neighbouring sessions at the beginning of training and less at the end of training will give the impression that there is a more significant decrease in noise correlations. More effort should be made to counter inadvertently cherry-picking in the session selection, as Fig. 2.7b indicates this may be one reason for such a sizable decrease in noise correlation for  $M_2 V_1$ .

A pairs of channels which have higher noise correlations in one session are likely to have higher noise correlation in other sessions. However, this effect is not present for  $M_1 V_1$ , Fig. 2.7a, and could suggest there is less session-to-session consistency for this dataset.

This figure provides an easy way of visually inspecting whether noise correlations are conserved, but does not allow us to quantify this.

## 2.5 DECODER-BASED METHODS

The input data to the decoding algorithm was chosen to be the total spike count during stimulus presentation from each trial. This total is taken only for channels known to offer a consistent firing rate across sessions after the spontaneous activity matching process. If there are 20 such channels in the dataset, taking the total number of spikes during stimulus presentation for each of the channels results in a 20-dimensional vector of population activity on each trial.

It should be noted that using all the spikes during the stimulus presentation period is likely to be sub-optimal, as not necessarily all of this period of time is carrying useful information about the stimuli. The amount of spikes before the onset response, for instance, is clearly not stimulus dependent as it precedes the neural response to the stimulus, so it will only be a source of noise and not useful for the decoding. Furthermore, it is possible that the firing rate during the onset response might be more informative than the average firing rate across the whole presentation period.

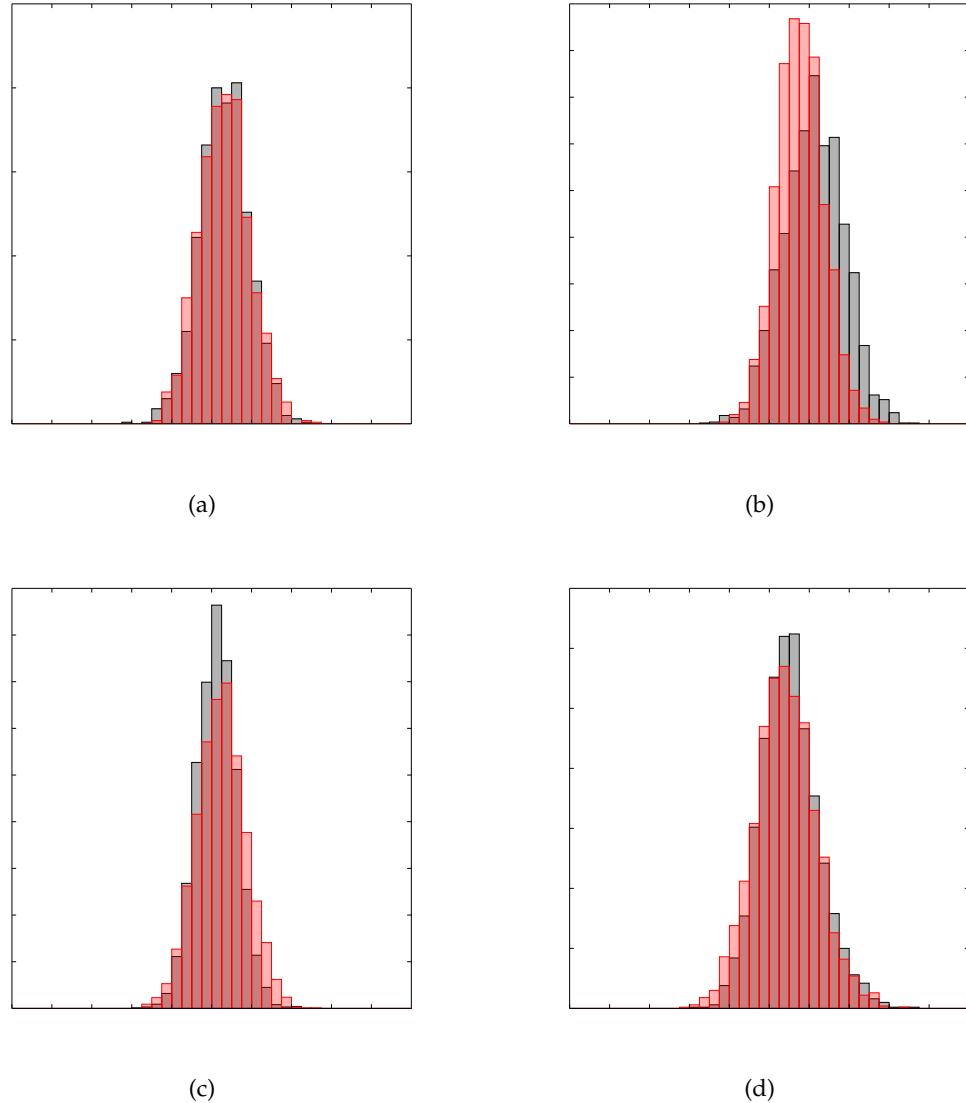


Figure 2.6: Distribution of the noise correlations for the pairings across all conditions. Two sessions, one at the beginning of training (black) and one at the end of training (red) are shown for comparative purposes. (a): M<sub>1</sub> V<sub>1</sub>; sessions 343 and 354. (b): M<sub>2</sub> V<sub>1</sub>; sessions 51 and 71. (c): M<sub>1</sub> V<sub>4</sub>; sessions 307 and 338. (d): M<sub>2</sub> V<sub>4</sub>; sessions 27 and 46.

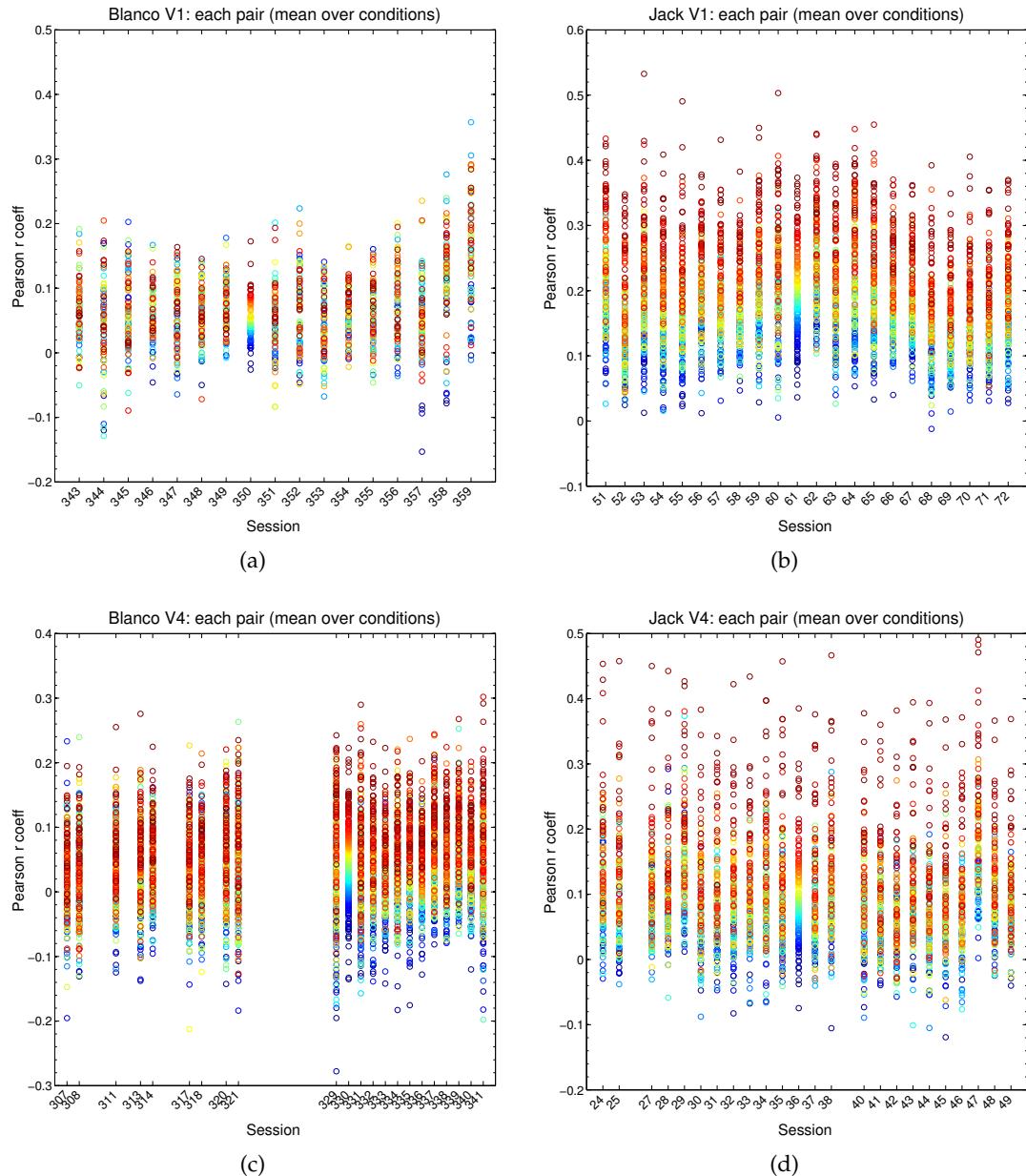


Figure 2.7: Noise correlations for each pair, meaned across the 14 conditions. (a): M<sub>1</sub> V<sub>1</sub>. (b): M<sub>2</sub> V<sub>1</sub>. (c): M<sub>1</sub> V<sub>4</sub>. (d): M<sub>2</sub> V<sub>4</sub>. Colour is assigned by sorting the pairs into ascending order for one of the sessions near the middle of the training period. The degree of session-to-session correlation of the noise correlation can hence be inferred by visual inspection.

However, finding the optimal window from which to take the spikes is an exercise in of itself.

The decoder used was a Fisher linear discriminant classifier. Given a training dataset of labelled data points in 20D, the classifier will find the optimal 19D hyperplane such that it separates the two classes of points optimally, under the assumption that the two clusters to be separated are multivariate normal distributions. The vector normal to the hyperplane is

$$\vec{w} = \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_0) \quad (2.4)$$

where  $\Sigma$  is the covariance matrix between the two populations, as determined from the available labelled data, and  $\vec{\mu}_0$  and  $\vec{\mu}_1$  are the computed means of the two distributions.

Once a classifying hyperplane has been found, test data points can be classified by observing the side of the hyperplane on which they fall. For a new data point,  $\vec{x}$ , if

$$\vec{w} \cdot \vec{x} > c \quad (2.5)$$

then we classify  $\vec{x}$  as group 1 (higher than 30 % contrast), otherwise we classify it as 0 (lower).

This was performed using the function `classify` with type ‘linear’ in MATLAB. For training and testing of the decoder, leave-one-out classification was used across each session individually. This means that for a session of 500 trials, the decoder is trained on 499 trials with the spike counts for all the trials and the correct response (higher or lower than 30 % contrast) and we then check whether the decoder identifies the remaining trial correctly. This is repeated for all the 500 trials, and then the performance is defined as the proportion of trials which are identified correctly. Training the decoder in this way means the decoding hyperplane used for classification is optimal for each individual session, but also different for each session.

The impact of noise correlations on the decoder is considered by comparing the regular decoder performance with the performance of a decoder applied to data which has been shuffled across trials under the same stimulus conditions. After shuffling, specific neural response vectors no longer correspond to any particular trial, since the spike counts for each channel are taken from different trials.

### 2.5.1 Probabilistic Model

As well as comparing the behavioural and decoding responses with the target responses to find a value for predictor performance, we can compare the behavioural

and decoding responses with one another and find a value for their agreement, the proportion of trials on which the two responses are the same.

We would like to evaluate whether the behavioural responses are dependent on the firing rates of the recorded population of neurons. This can be done by considering the behavioural and decoding response agreement, but in order to do this we need to construct a model of how much agreement we would expect to find if the two were completely independent, which will form our null hypothesis (NH) to compare against.

Throughout this work, it is assumed that the behavioural and decoding responses are given by a Bernoulli distribution. In this model, there is a true probability,  $p$ , of the animal giving the correct answer to the task for each trial, which is sampled from on each of the trials. When this is sampled over  $n$  trials, we expect to find  $np$  trials have been responded correctly and  $n(1 - p)$  incorrectly.

Hence the underlying true probability of a correct response can be approximated by taking the number of correct responses and dividing them by the total number of trials. Using one of several statistical methods, we can find an estimate for a percentage confidence interval (PCI). It is from this that error bars, indicated by the shaded regions surrounding line plots in Figs. 2.10 and 2.9, are computed. These are presented as the standard error, equivalent to a 31.8 % confidence interval within which 68.2 % of data points are expected to lie. (This has been done using the function `binofit` in MATLAB, which utilises the Clopper-Pearson method. Although this is not the best method, it should be sufficient for our purposes.)

The behavioural performance is given by the estimated probability of the behavioural response being the same as the correct response, simply found by dividing the number of times the behavioural response matched the target response by the number of presentations. The target response is of course whether or not the test stimulus is actually higher or lower than the sample. Similarly, the decoding performance is the proportion of times the decoding response is the same as the target response; and the behavioural and decoding agreement is the proportion of times the behavioural and decoding responses are the same.

Using the estimated probabilities of the behavioural and decoding responses being correct, we can find an expected agreement between the behavioural and decoding responses if they are completely independent of one another, which constitutes the null hypothesis model (NH). By comparing the observed agreement with the agreement we expect for independent variables, we can see whether or not behavioural responses are correlated with the decoded neural responses.

The theory behind the estimation of the agreement for independent binomial processes is described as follows. Let us take two independent binomial variables,  $X_1$  and

$X_2$ , which can either be in state  $s_1$  with probabilities  $p_1$  and  $p_2$  respectively, or in state  $s_2$  with probability  $(1 - p_1)$  and  $(1 - p_2)$ . If  $X_1$  and  $X_2$  are completely independent, we expect to find them in the same state with probability  $p_1 p_2 + (1 - p_1)(1 - p_2)$ . Obviously, if  $p_1$  and  $p_2$  are both 0.5, there is a 0.5 chance of them being in the same state. If  $p_1$  and  $p_2$  are both either higher or lower than 0.5, there is a higher chance of  $X_1$  and  $X_2$  being in the same state, and conversely if one is above and one below 0.5, there is a lower chance of them being in the same state. This means that if the probabilities of the behavioural and decoding responses being correct rises during the perceptual learning process, we expect the proportion of trials on which they agree to rise as well, even if the two distributions are completely independent.

The initial model for the expected agreement between behaviour and decoding used a simple model as described above, where  $p_1$  and  $p_2$  are taken to be the overall probability of a correct response across all the test stimuli for behaviour and decoding respectively. We will refer to this model as the single binomial model, as each predictor is modelled with a single binomial. However, since there are 14 possible stimuli with varying difficulty, a more detailed model can be conceived in which the probability of a correct response for each predictor depends on the stimulus presented. In this “multi-binomial” model, the set of conditions (different stimuli) is given by  $C$ , and each predictor,  $i \in \{b, d\}$ , can either be in the “lower” or “higher” state ( $s_l$  or  $s_h$ ). For each  $c \in C$ , there is some probability  $p_{il|c} := P(x_i = s_l|c)$  of predictor  $i$  being in the lower state. Consequently, the probability of the two predictors being in the same state is given by

$$p_a = \sum_{c \in C} P(c) (p_{bl|c} p_{dl|c} + (1 - p_{bl|c})(1 - p_{dl|c})),$$

where  $P(c)$  is the probability of condition  $c$  being presented to the animal on any given trial.

We will assume that  $c$  is a random variable, which is sampled from for each trial independent from prior trials. This is a simplification of the actual condition selection process, “delayed repetition”. In this, the conditions for a set of trials are chosen in advance with  $n$  appearances of each condition in a random order, after which any trial on which the animal responded incorrectly is repeated. This means the single random variable approximation  $P(c)$  is not uniform across the conditions, as harder conditions are presented more frequently.

For a given group of trials (such as, for instance, a single session),  $P(c)$  can be estimated by dividing the number of times each appears by the total number of trials in the group. Similarly,  $p_{il|c}$  can be estimated by dividing the number of trials on

which the predictor was in the lower state for a given condition by the number of trials where that condition was presented.

It is not immediately clear whether the single- or multi-binomial models will yield similar or different results. Obviously, the multi-binomial model has more accurate detail and so should give a better result, but whether or not the single binomial should give a reasonable approximation is not obvious, though one would assume it would. Applying the two models to the data and comparing them, we found the single binomial significantly underestimates the agreement which would really be found if the two predictors were completely independent. Henceforth, only the more accurate multi-binomial model is considered.

The expected agreement under the assumption of two independent binomial processes gives us the null hypothesis model to compare against. If the results of the analysis deviate significantly from this model, we can conclude the behavioural and neural decoded responses are correlated. A  $p < 0.05$  significance boundary was found by sampling the NH model 100,000 times and noting the agreement for the fifth-percentile of samples. During this sampling, the number of trials for each test condition was fixed with the same values as used in the experiment. Since we are only interested in sessions where the behavioural and decoded responses agree more than they would if independent, a single-tailed test was used. (NB: in the figures below, the single-tailed lower bound is also included for symmetry, so only 90 % of the samples lie within the shaded region).

The multi-binomial model should not be confused with a multinomial distribution, which would map one random variable to  $n$  possible responses. The model used here is instead a two-step process where there are a set of binomial distributions, which binomial is used is determined by one random variable (the output of which is the same for both behaviour and decoder), and the output of the selected binomial is given by a second random variable. In essence, the simple, single binomial model is similar to shuffling the responses across all trials regardless of condition, and the multi-binomial model we use is similar to shuffling the responses with the same condition presented.

### 2.5.2 Results

Fig. 2.8a shows how well the decoder does based only on data from individual channels. The performance was measured as described in §2.5 for each session, and then averaged over all sessions. [perhaps a histogram would show how these are distributed better?]

Fig. 2.8b indicates how the performance of the decoder improves as more channels are added. We can see that the decoder performance saturates after around 10 channels are included in the training data, but at a performance which is still far from the ideal. After this, including additional channels yields only a slight increase in performance, suggesting nearly all the information contained in the remaining channels is redundant, as it has already been given in the first 10 channels. [would be better to look at how the decoding performance is distributed across all possible sets of  $n$  channels, rather than just the best set of  $n$  channels] [this should be compared with shuffled data to see what the impact is and whether performance does not saturate when correlations are removed as per some previous papers [cite]]

NB: the order in which channels were added in Fig. 2.8b is not the same ordering as they are shown in Fig. 2.8a.

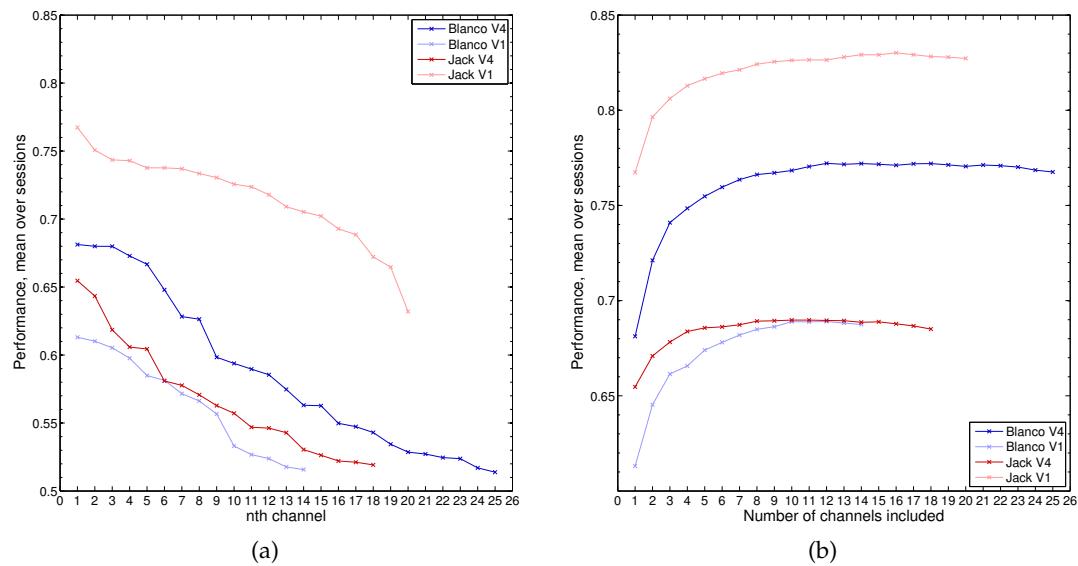


Figure 2.8: (a): Distribution of decoder performance based on spike rate for individual channels, sorted by performance. (b): Decoding performance versus number of channels included. Channels were added one at a time, chosen so they maximise the decoder performance for that number of channels whilst keeping all the channels which had come before.

For each of the datasets, the performance decreases as the last 3 channels are added (Fig. 2.8b). I speculate that this is because these channels only contain redundant information, and the increase in dimensionality decreases the quality of the classifier selected from the finite training data available.

For M2 V4, the trend in increase for behavioural performance is well matched by the change in performance of the decoder (Fig. 2.9b, blue and red lines). In contrast to this, for M1 V4 we see the decoding performance is steady despite training, although

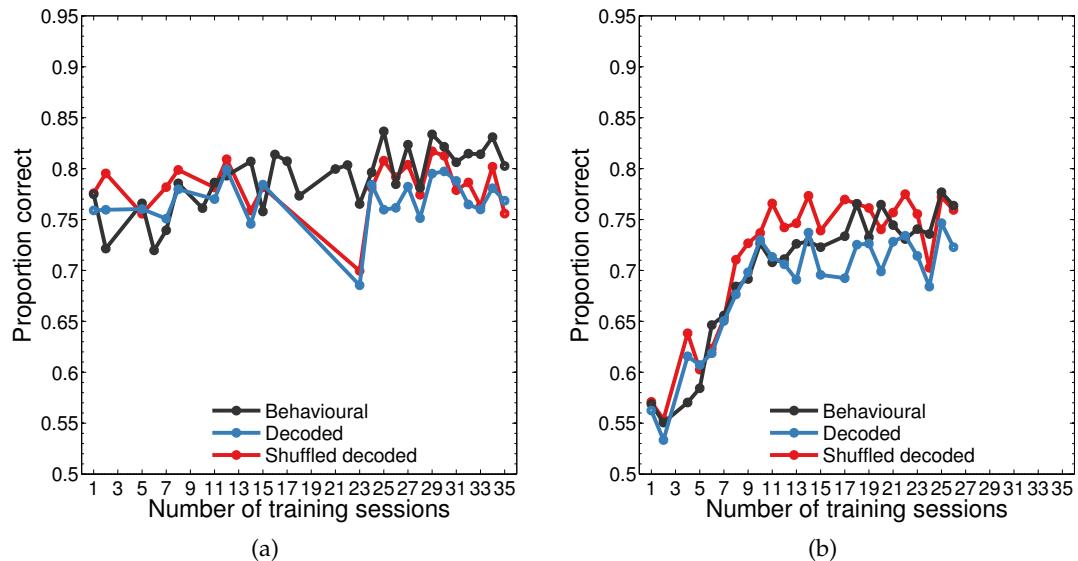


Figure 2.9: Decoding analysis for V4. Performance of behavioural and decoding predictors by session, averaged across all conditions. Left panels: M1. Right: M2. Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis, ‘proportion correct’ is the proportion of trials for which the response is the same as the target. This is presented for behavioural performance (black), decoder performance (blue), and decoder performance when trials are shuffled, destroying noise correlations (red; see text).

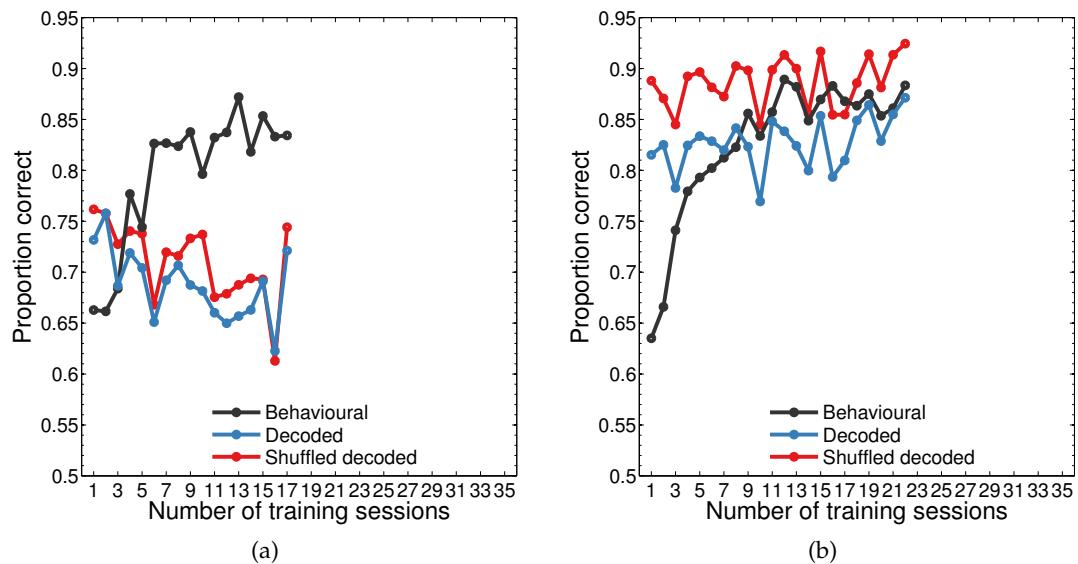


Figure 2.10: Decoding analysis for **V1**. Performance of behavioural and decoding predictors by session, averaged across all conditions. Left panels: **M1**. Right: **M2**. Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis, ‘proportion correct’ is the proportion of trials for which the response is the same as the target. This is presented for behavioural performance (black), decoder performance (blue), and decoder performance when trials are shuffled, destroying noise correlations (red; see text).

it should be noted that the increase in behavioural performance was not so large for this animal.

For **M<sub>2</sub> V<sub>1</sub>**, there is a small gradual increase in decoder performance throughout training, though the increase does not match the timescales and rate of the increase in behavioural performance which is a much sharper transition. Results for **M<sub>2</sub> V<sub>1</sub>** could be interpreted as indicating that no further improvements can be obtained from changes in **V<sub>1</sub>**, and the behavioural performance is limited by the performance of the decoder. For **M<sub>1</sub> V<sub>1</sub>**, there is a decrease in decoder performance despite the increase in behavioural performance (Fig. 2.10a, blue and red lines respectively).

Shuffling the trials to destroy any noise correlations provides an improvement in performance for the data from **M<sub>2</sub>**, indicating that noise correlations are detrimental to the successful decoding of stimuli. However, there is no significant difference in decoder performance before and after shuffling for **M<sub>1</sub>** datasets, which suggests noise correlations have no impact on decoding the contrast of stimuli using firing rate data from a population of neurons. For **M<sub>2</sub> V<sub>1</sub>**, the benefit from shuffling seems to be reduced with training; the shuffled decoding performance is reasonably steady throughout training whilst the decoder with noise correlations included increases in performance, closing the gap.

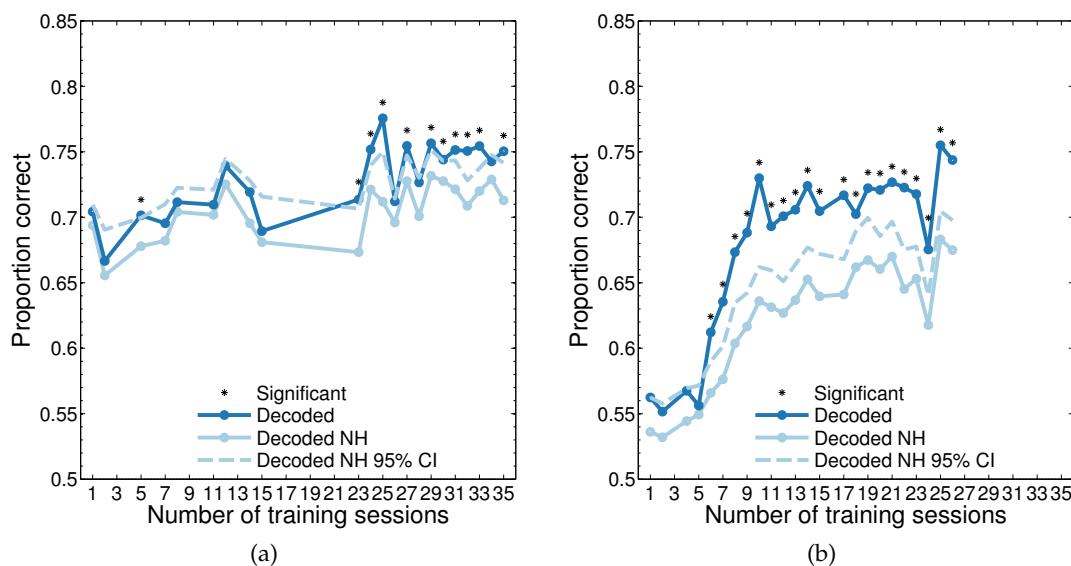


Figure 2.11: Decoding analysis for **V<sub>4</sub>**. Trial-to-trial agreement between behavioural and decoding predictors. Left panels: **M<sub>1</sub>**. Right: **M<sub>2</sub>**. Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis is the proportion of trials for which the response is the same as the behavioural response. The agreement between behaviour and decoding (unshuffled only) is presented alongside the null hypothesis of completely independent binomial distributions. The dashed line indicates the 95% confidence interval of the null hypothesis.

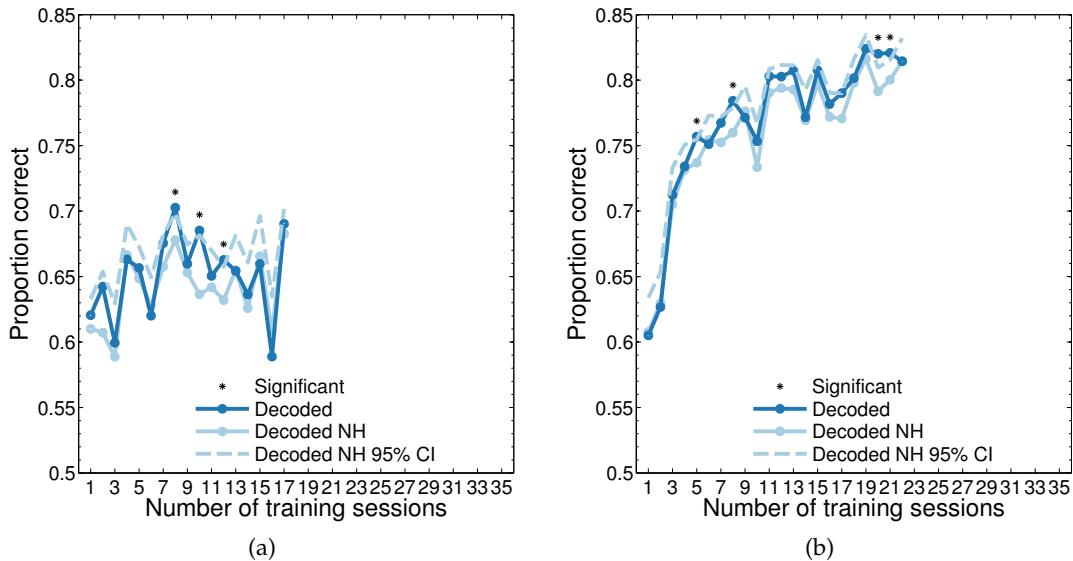


Figure 2.12: Decoding analysis for  $V_1$ . Trial-to-trial agreement between behavioural and decoding predictors. Left panels:  $M_1$ . Right:  $M_2$ . Along the x-axis, ‘Session’ is the animal’s unique session ID, which increments by one for every day of training. On the y-axis is the proportion of trials for which the response is the same as the behavioural response. The agreement between behaviour and decoding (unshuffled only) is presented alongside the null hypothesis of completely independent binomial distributions. The dashed line indicates the 95% confidence interval of the null hypothesis.

With regards to the behavioural and decoding response agreement, we find there is no consistent significant deviation from the null hypothesis for  $V_1$  data. There are a couple of sessions where unshuffled agreement is above the boundary for significance for each of the animals, but this is not a consistent effect. The agreement between shuffled decoding and behaviour does not deviate from the corresponding null hypothesis. This shows that the shuffled is higher than the unshuffled decoding agreement only because the shuffled decoder is more accurate at matching the target response.

For  $V_4$ , we see that the agreement does not deviate significantly from the null hypothesis for earlier sessions, but after a cut off point all later sessions do ( $M_1$ : all sessions before 321 are not, after 329 are significant;  $M_2$ : Before 28 are not, after 28 are significant). The effect is stronger for  $M_2$ , but present for both animals. This indicates that the behavioural responses and the decoded responses based on the neural data were not notably correlated, but have become so with training.

Because the shuffled decoded responses are more accurate (with respect to the target response), they are predicted under the null hypothesis to be in better agreement with the behavioural responses. However we observe these are in worse agreement than the unshuffled decoding, and do not deviate from the corresponding NH line.

A more detailed breakdown of these results with subsets of the 14 conditions is given in Figs. ?? and ?? . Comparing Figs. ?? and ?? with Figs. ?? and ?? , we can see that the decoding responses for the easier conditions fit the null hypothesis model, whilst the more challenging conditions do not and have a statistically significant agreement with the animal behaviour.

## 2.6 INFORMATION THEORY-BASED METHODS

### 2.6.1 Session-based analysis

The information was computed for each day of training (hereafter referred to as a session), for each of the channels from which recordings were available. To gauge the typical behaviour of a neuron, the mean was taken over all the channels. The mutual information between the spiking activity during test presentation and the identity of the test stimulus was computed using a spike-timing based code. This computation was done using the `information` function from the freely available Information Breakdown Toolbox (Magri et al., 2009) for MATLAB.<sup>2</sup>

A moving window of 20 ms was taken and subdivided into 5 sequential bins, each of 4 ms. The number of spikes within each of the 5 bins was totalled up to provide a code letter for the bin. The combination of the 5 sequential counts forms a codeword.

This is performed across all the trials in a session with the 20 ms window placed with the same offset relative to the test stimulus onset. By assembling the trials according to their test contrast, the mutual information between the contrast and the spiking activity in this window can be computed.

Following the methods employed by the research group, only trials in which the monkey responded correctly were used for this analysis.

The results of this initial analysis are shown in Fig. 2.13. No trend in information content with learning can be discerned because the measured mutual information content varies wildly from session-to-session.

Fig. 2.14 shows that the changes in the measured mutual information are dominated by the number of trials in the session, not by increases with perceptual learning (corresponding to increments in the session number). The only exception to this rule seems to be for M2 V1, shown in Fig. 2.14c, where the information increases with learning despite an increase in the number of trials per session over this period.

In particular, the measured value for the mutual information is strongly correlated with the reciprocal of the number of trials in the session (Fig. 2.15). The correlation is still very strong even if we use one of the two bias correction techniques. As discussed

---

<sup>2</sup> Available at <http://www.ibtb.org>

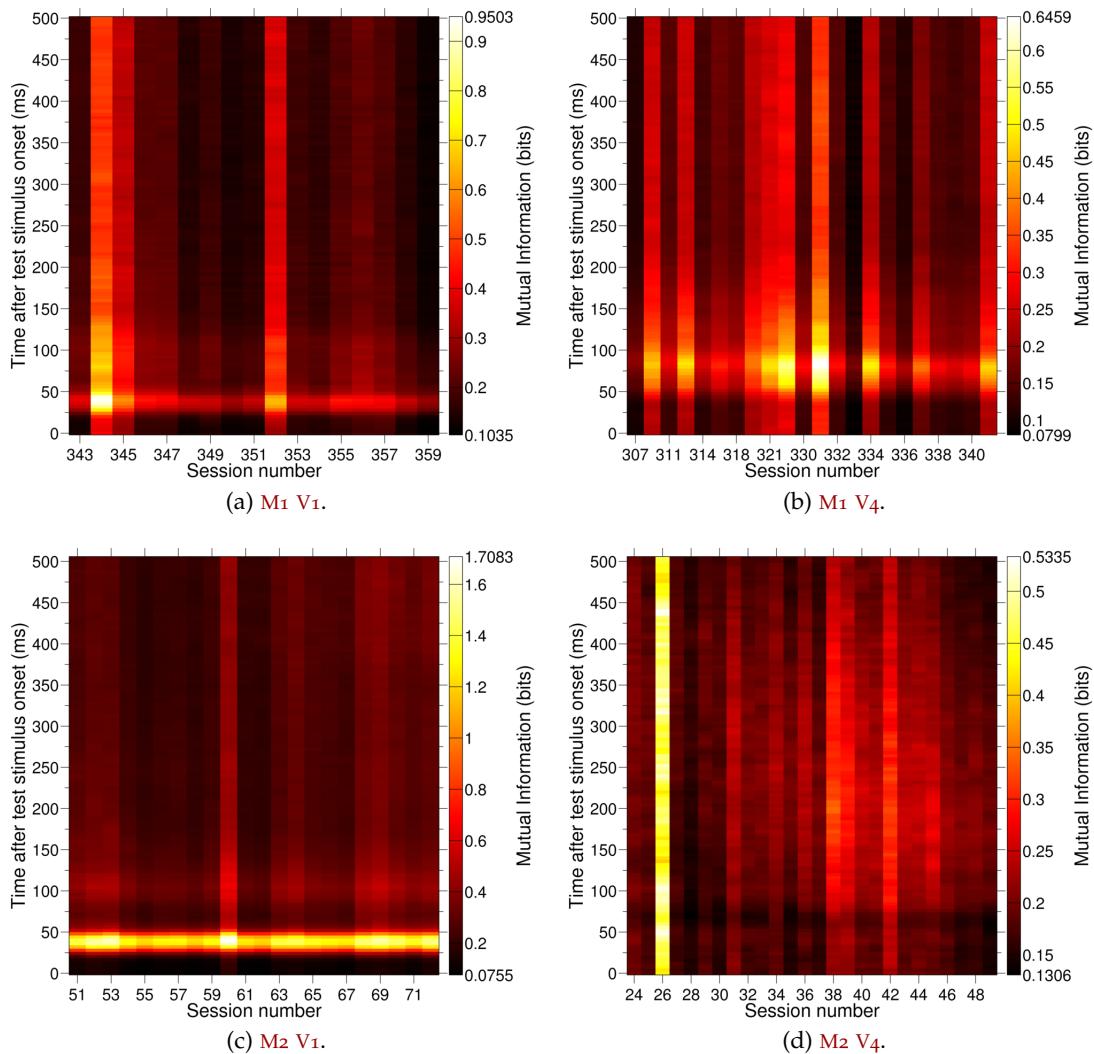


Figure 2.13: Mutual information between the test stimulus and the neural activity during test presentation. The mutual information with the test stimulus is taken for a spike timing based code for a 20 ms window of spiking activity, sampled with the start of the window offset ( $y$ -axis) from 0 ms up to 500 ms after test stimulus onset (which is slightly more than 20 ms before test stimulus offset). The sampling is in intervals of 5 ms, so any 4 adjacent squares within each session are highly correlated. The recording session number for the data is given along the  $x$ -axis, and the number of days the animal has been trained for increases from left to right. Average mutual information across all the channels is denoted by the pseudo-colour of each of the rectangular patches, centred around the  $(x, y)$  co-ordinate to which the measurement relates. In each case the average is taken across all available channel data: 2.13a 23 channels, 2.13b 30 channels, 2.13c 25 channels, 2.13d 20 channels. No bias correction method was used.

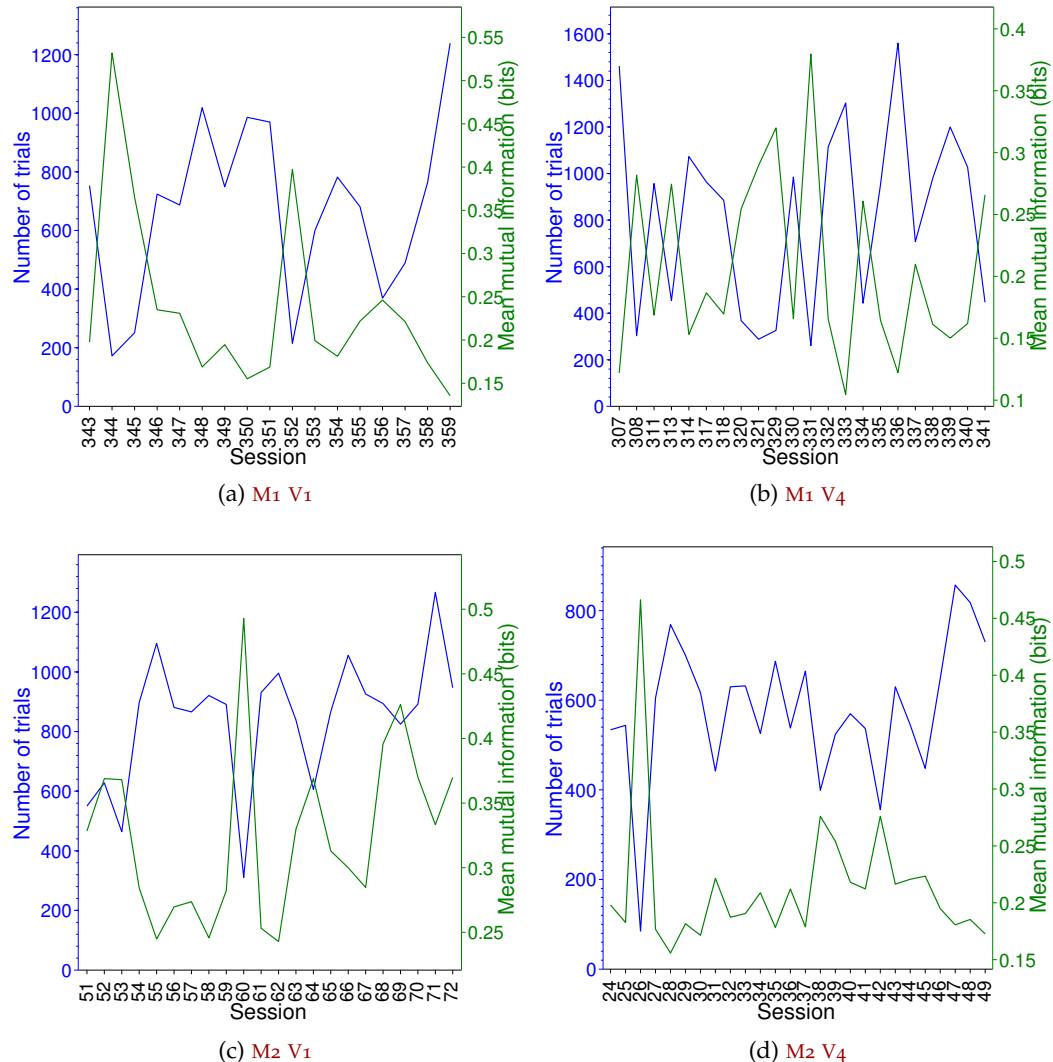


Figure 2.14: Mutual information seems to be anti-correlated with the number of trials in the session. The number of correctly responded trials in each session is plotted on top of the mean information in individual sessions. The mutual information was not corrected for bias, and averaged across the whole period of test stimulus presentation.

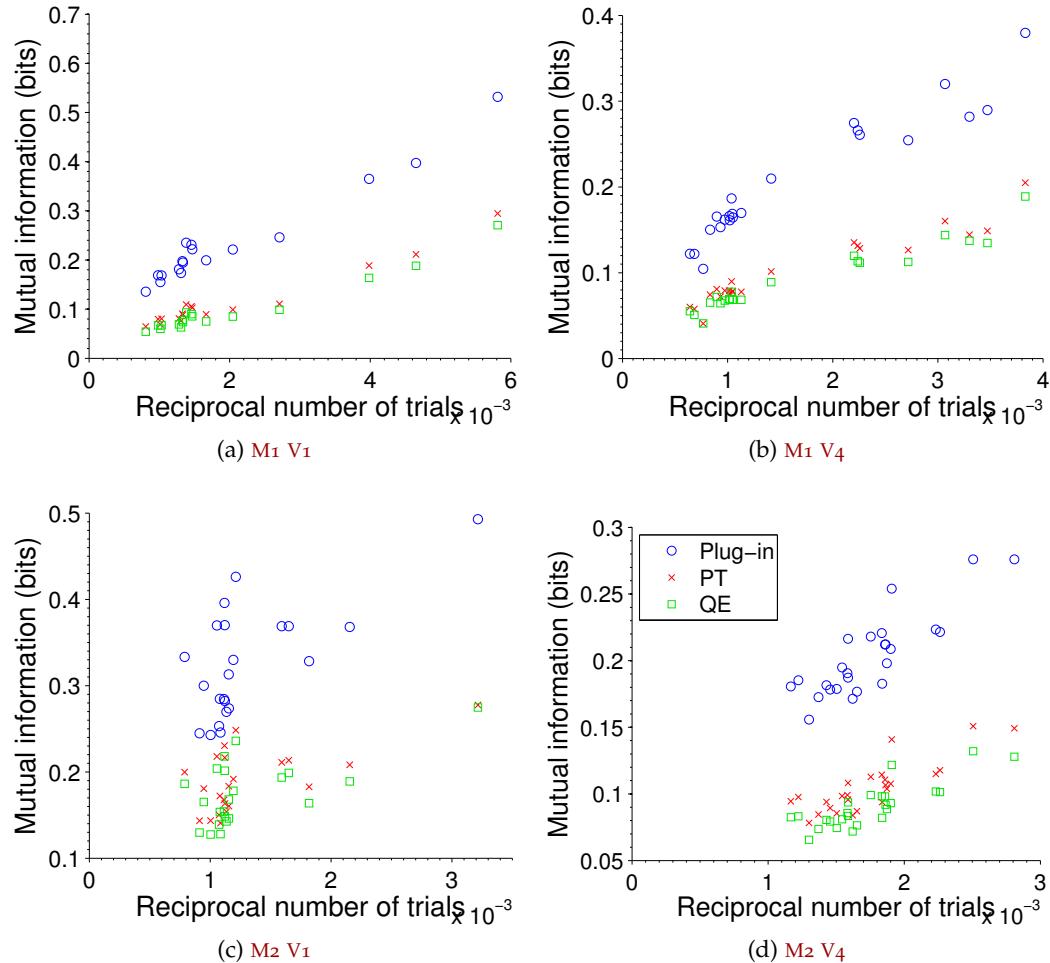


Figure 2.15: Mutual information is inversely correlated with the number of trials in the session. The mutual information for a single session is averaged over all channels, and averaged over all times during stimulus presentation, and plotted against the reciprocal of the number of trials in the session ( $1/N$ ). Plug-in: without bias correction. PT: using the Panzeri-Treves method to correct for bias (Panzeri and Treves, 1996). QE: using quadratic extrapolation for bias correction (Strong et al., 1998).

above, for **M2 V1** the number of trials per session and the measured information are both correlated with time, which reduces the correlation observed in Fig. 2.15c.

The problem is that although PT and QE both do a reasonable attempt at removing the bias, they are not perfect and some bias will remain. As the bias is dependent on the number of trials, it is unsurprising the remaining bias follows the same dependency. In particular, in the asymptotic regime the leading term in the expansion of  $I_{\text{measured}}$  is known to be proportional (Treves and Panzeri, 1995) to  $1/N$ ; the same relationship observed here.

In particular, the difficult faced is changes due to perceptual learning which are of interest may only be slight, but the the number of trials per session varies five-fold: from 250 up to 1250. Consequently it is unsurprising that the observed differences in information day-to-day are dominated by the number of trials.

## 2.6.2 Trial-based analysis

To counter the correlation between measured information and number of trials, an obvious solution is to use the same number of trials for every computation. We now consider how many trials should be taken at once to obtain a reasonably reliable estimate.

Some rules-of-thumb for the number of trials are offered by (Panzeri et al., 2007). Let  $\bar{R}$  denote the number of possible response codes. Then, for the bias-uncorrected estimate of mutual information, we need to have at least  $N_S \geq 2^{32} \bar{R}$  trials per stimulus, whilst if the PT or QE bias correction methods are applied, we only need  $N_S > 2 \bar{R}$ .

The probability of having two spikes within 4 ms from a non-bursting neuron is very low. For example, a neuron with a high firing rate might fire at 100 Hz, which means inter-spike intervals are typically around 10 ms (and 100 Hz is a high firing rate for the channels in our dataset). Consequently any 4 ms bin will realistically contain either 0 or 1 spikes, and the number of possible response codes in our spike timing code analysis with 5 bins each of 4 ms is  $\bar{R} = 2^5 = 32$ . In comparison for a spike count code, if we assume spikes cannot be closer together than 3 ms, there are between 0 and 7 spikes in any 20 ms interval. Consequently there are  $\bar{R} = 8$  possible response codes.

As we wish our analysis to work for both spike timing and spike count codes, using the above rules we need to use at least 64 trials per stimulus to get a reasonable estimate of the information with one of the bias correction methods in place. Since there are 14 different stimuli, this means at least 896 trials in total are needed. This presents a dilemma, since most of the sessions are shorter than this, even when both

correctly and incorrectly responded trials are included. Excluding sessions with fewer than 896 trials would severely limit the size of the dataset.

The solution found was to concatenate the sessions together and analyse groups of  $N$  trials taken from multiple consecutive sessions. The naïve justification for this approach is a “first-order approximation” to perceptual learning would be that the monkey gets better at recognising the stimuli every-time they perceive it, and so the most important quantifier for the amount of perceptual learning which has taken place is the total number of trials the monkey has performed to date.

This rather basic assumption neglects several factors which influence the animal’s performance, such as their mood during the particular training day; and factors which influence the animal’s willingness to work (in turn influencing performance), which depends on their recent level of access to water (the reward used in the study). For instance, the animal performs less well on Mondays, which follows on from readily available water during the weekend. Furthermore, the consolidation which occurs during sleep is commonly believed to be important to perceptual learning, and (as this only occurs between training sessions) this effect is completely ignored by this approach.

However, the session-concatenation approach was attempted regardless, and a value of  $N_S = 100$  trials per stimulus was chosen. Though there are enough trials available to use more and reduce the bias further, it is undesirable data from so many sessions at once.

As mentioned in Chap. 2.2, a delayed repeat is used for any trials to which the animal does not correctly respond. Consequently, more difficult test stimuli (with contrasts close to the sample contrast) are presented significantly more frequently than easier stimuli. For instance, when presented with the most difficult test stimuli, the animal will have a success rate only just above 50 % on the first day of the experiment, rising to around 70 % by the last day. For the easiest stimuli, the success rate will be nearly 100 % throughout.

Say we arrange all the trials into groups based on their stimulus, then take 100 subsequent trials from each of the groups to perform the analysis on. Due to the different number of trials per stimulus, it will not be long before the trials selected for each stimulus are from very different points in time and from different sessions.<sup>3</sup> To ensure all the trials are from when the animal has had the same amount of training, we can instead take a group of  $S \cdot N_S = 14 \cdot 100 = 1400$  consecutive trials regardless of

---

<sup>3</sup> Even if we only analyse the correctly responded trials, there is still a difference in the number of trials per stimulus due to a limit on the number of repeats of any test condition. The difference was negligible for M<sub>1</sub>, but accumulated to a 100 trial difference over all the sessions between the most and least correctly responded conditions for M<sub>2</sub>.

the stimulus presented and analysed. The differences in number of trials per stimulus are not particularly important, so long as there are always enough.

All trials where the monkey completed the trial and gave a response (either correct or incorrect) were included in the analysis. Trials where the monkey did not complete the task by fixating and then providing a response as required were excluded. Using both the correct and incorrectly responded trials means our distribution for  $P(s)$  is the same as that presented to the monkey, and there are notably more trials available to perform the analysis on.

### 2.6.3 General Information plots

Starting with  $\text{V}_1$ , the first thing which is noticed in Figs. 2.16 and 2.17 2.16a and 2.16c is the large peak in information over the 20 ms window starting at around 40 ms after stimulus onset. This is due to the onset transient response, where there is a larger amount of neural activity, and this is also less variable than usual (Müller et al., 2001). A second, smaller peak from the “rebound” of the transient also occurs around 100 ms after stimulus onset. Looking at the scale bars, we can see there is about 10 times as much information on average from the neurons in  $\text{M}_2$  than  $\text{M}_1$ . This is probably due to differences in data quality between the two animals.

Comparing the information found using the spike timing code 2.16c with the spike count code 2.16a, it seems that there is significantly more information when the binned spike times are considered: there is three times as much information for the spike timing code for  $\text{M}_1$  and twice as much for  $\text{M}_2$ . However, if we look at information in the spontaneous activity, this reveals we cannot trust this result, as the bias for the information from the spontaneous activity is much higher for the spike timing code than spike count.

For  $\text{M}_1$ , aside from the transient, the information measured with the spike timing code from the spontaneous activity is about the same as the information from the test presentation, suggesting something has gone very wrong!

For the spike timing codes, there is a clear decrease in information with learning in  $\text{M}_1$ , and a clear increase in information for  $\text{M}_2$ . This is, however, present in both the test presentation activity and the spontaneous activity, suggesting it is not a genuine effect. This is believed to be due to a decrease in signal quality in the implants in  $\text{M}_1$ , and possibly an increase in  $\text{M}_2$ . The increase is also seen in the spike count code for  $\text{M}_2$ , but not to any real extent and it is doubtful that the result is significant.

The bias in the information for the spontaneous activity changes with time for the spike timing code, but does not for the spike count code. Since any changes in the dataset are the same for both of these, this suggests there are too few trials for the

spike timing code to give a reliable reading of the information. This would make sense because with an average of  $N_S = 100$  trials per stimulus and a spike timing code with  $\bar{R} = 32$ , there are on average  $N_S/\bar{R} = 3.125$  trials per response per stimulus, whilst for a spike count code with  $\bar{R} = 8$  there are on average  $N_S/\bar{R} = 12.5$  trials per response per stimulus. Similarly if we assume a minimum of  $N_S = 64$  trials per stimulus, there are at least  $N_S/\bar{R} > 2$  trials per response per stimulus for the spike timing code, and at least  $N_S/\bar{R} > 8$  trials per response per stimulus for the spike count code.

The information in the spontaneous activity was subsequently computed with an average of  $N_S = 200$  and with  $N_S = 400$  trials per stimulus. This means there is now  $N_S/\bar{R} > 8$  for the spike timing code, however precisely the same relationship was observed. From this, we can conclude the problem is due to the inconsistencies between sessions. If the firing rate changes between two sessions, the probability distribution of the responses generated will change for each condition. This will not matter so much for the spike count code because there are only 8 possible responses, and even if there are only 250 trials in the session there should be at least 16 trials per condition, meeting the minimal requirements for the PT method to function.

The small bump in the information in the spontaneous activity around 50 ms in Figs. 2.16c and 2.17c will be due to an increase in activity from a transient response. This data is normalised so 0 ms is when the animal begins fixating on the fixation target, and there will be a transient response after the animal saccades to the target. Although the increase in activity does not relate to the conditions used on the subsequent test presentation, the extra spikes will increase the variability of the response, and thus  $H(\mathbf{R})$ , which will not be cancelled out by an increase in  $H(\mathbf{R}|\mathbf{S})$  for the same reasons the bias appears in the first place.

In short, the data for M<sub>1</sub> V<sub>1</sub> does not seem to be of good enough quality whilst M<sub>2</sub> V<sub>1</sub> does, and the information in the spike timing code cannot be trusted for either monkey.

Turning our attention to the V<sub>4</sub> results in Figs. ?? and 2.18, we can see the effect of the transient is present in M<sub>1</sub>'s data (at the later start time of 75 ms), but not in M<sub>2</sub>'s. This is surprising because, looking at the rasters, in both animals there are some channels which exhibit a transient response and some which do not.

Similar to V<sub>1</sub>, it seems as if there is four times as much information in the time-binned code compared with the count code. However, there is much more information measured for the spontaneous activity data again. This is not reduced by increasing the number of trials either.

For the spike count code in M<sub>2</sub>, the spontaneous information is nearly distributed around 0, suggesting the bias has been all but removed and the data is of very high

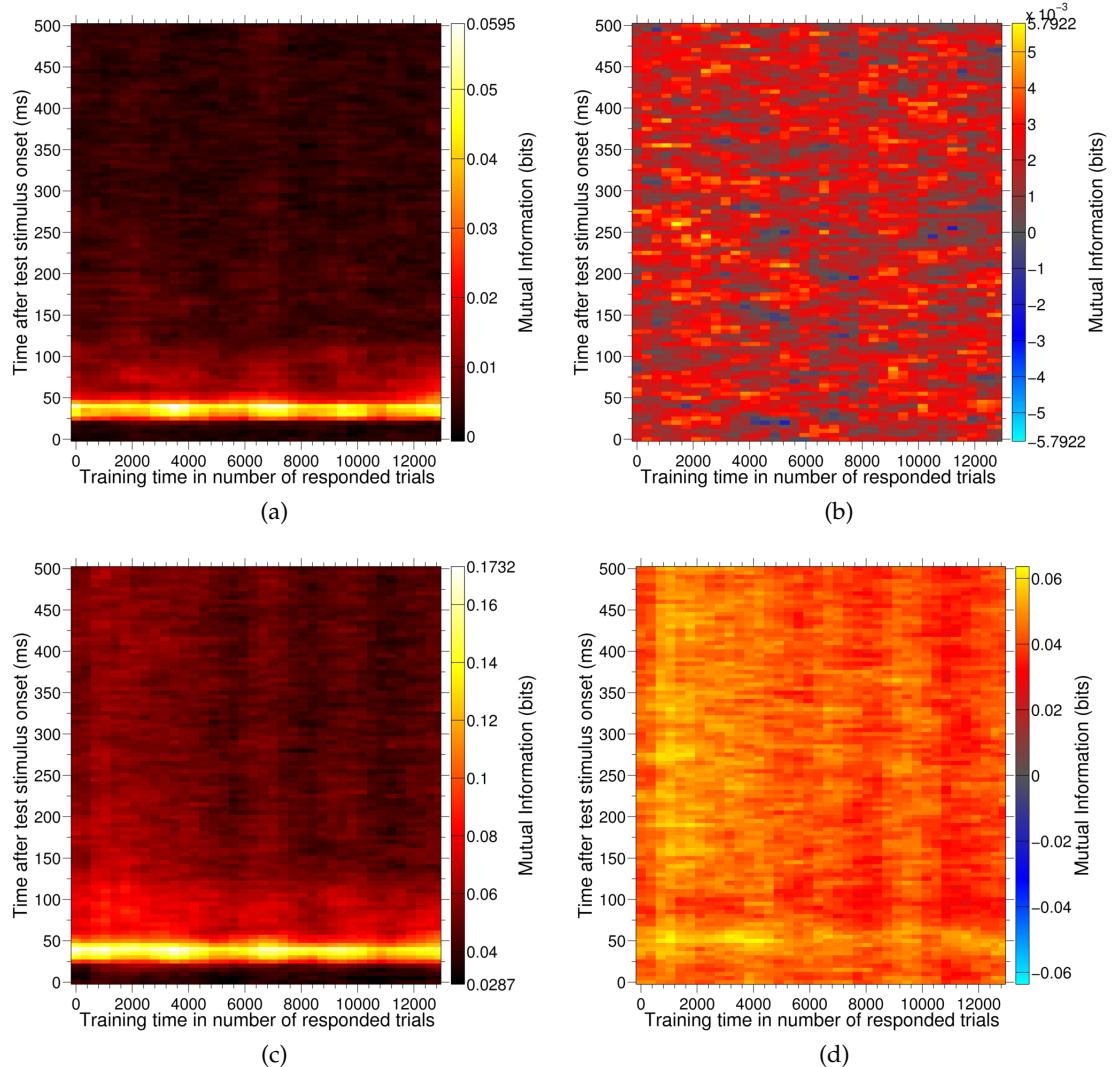


Figure 2.16: M<sub>1</sub> V<sub>1</sub>: Mutual information between the test stimulus and 20 ms of spiking activity, averaged across 23 channels. The PT bias correction method was used in all estimates of the information, and the modification to the spiking data to remove the artifact as described in Sec. 2.3.1 was performed. The neural code used in 2.16a, 2.16b is a spike count code, whilst in 2.16c and 2.16d it is a spike timing code where the 20 ms window was subdivided into 5 bins each of 4 ms. In 2.16a and 2.16c the spike-train is taken from the test presentation part of the trial; for 2.16b and 2.16d the spike-train is taken from spontaneous pre-stimulus activity. The data is sampled in intervals of 350 trials in the  $x$ -direction and 5 ms in the  $y$ -direction, so there is significant correlation between any pair of pixels in the image with less than 4 pixels between them in either cartesian direction.

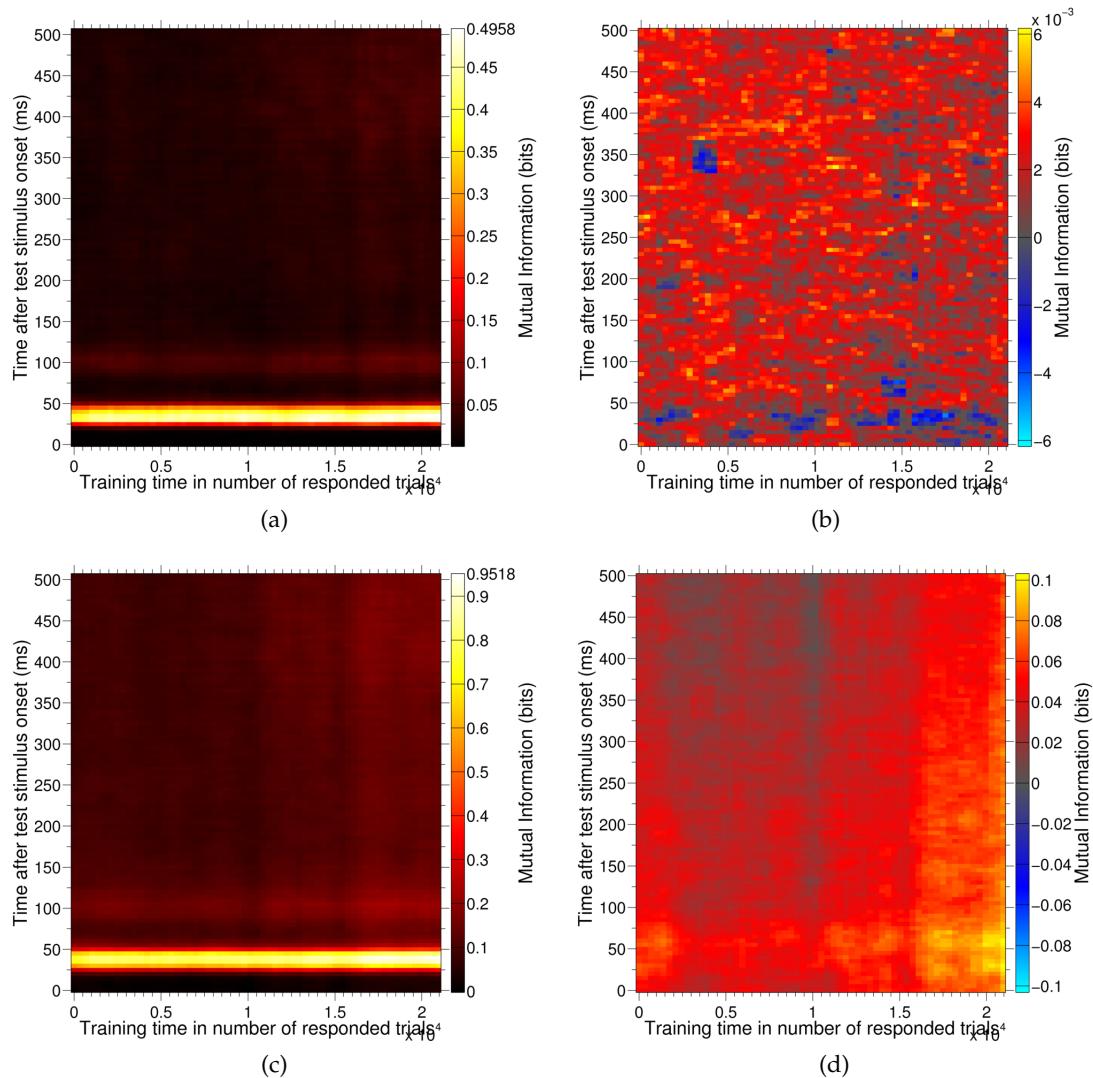


Figure 2.17: M2 V1: Mutual information between the test stimulus and 20 ms of spiking activity, averaged across 25 channels. The PT bias correction method was used in all estimates of the information. Panels 2.17a–2.17d are the same as for Fig. 2.16.

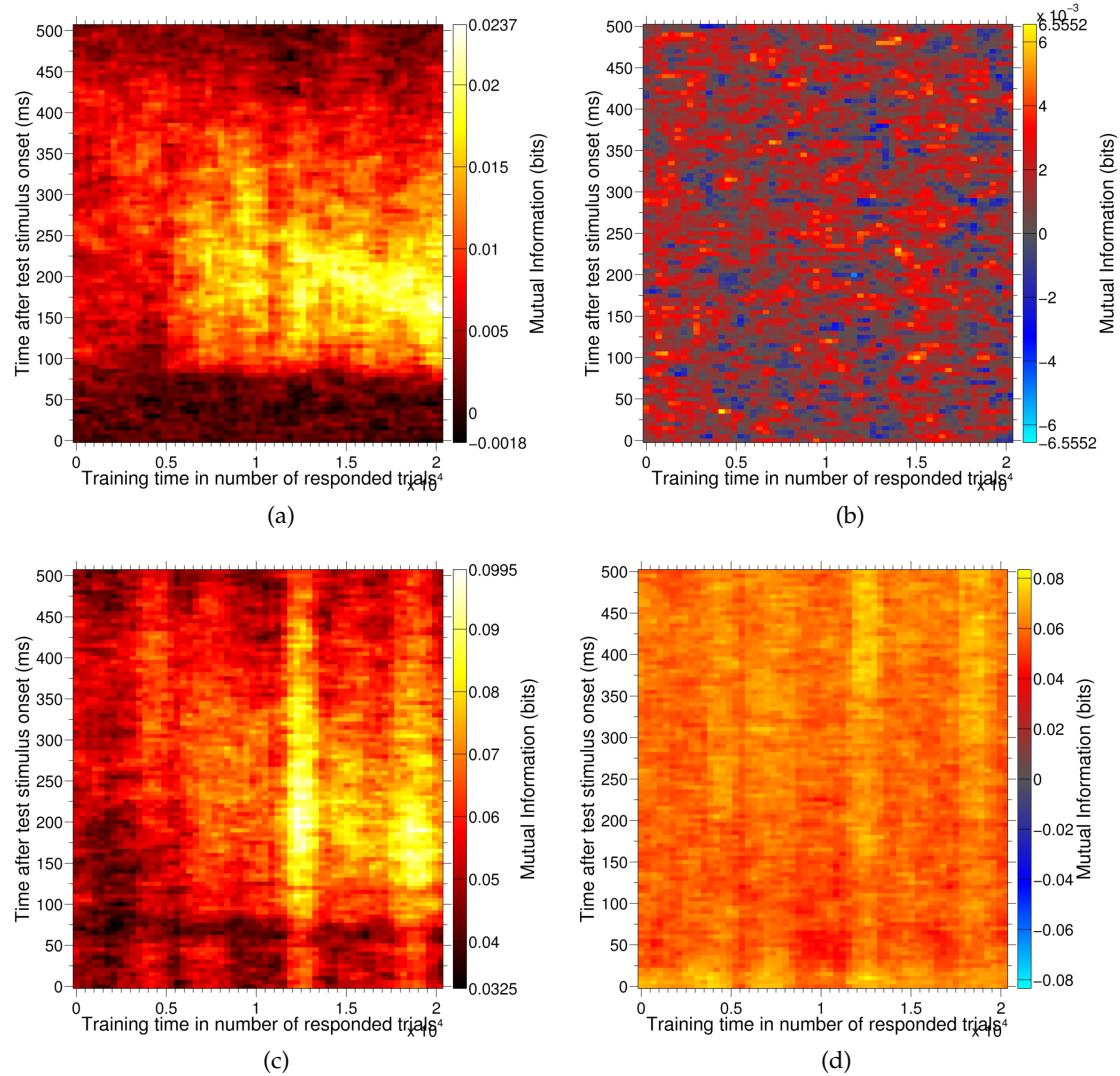


Figure 2.18: M2 V4: Mutual information between the test stimulus and 20 ms of spiking activity, averaged across 20 channels. The PT bias correction method was used in all estimates of the information. Panels 2.18a–2.18d are the same as for Fig. 2.16.

quality. For this animal, we can see a distinct increase in the information content with time, for both the spike count and timing codes. Simultaneously, there is a movement of the peak information to earlier times closer to the stimulus onset.

In  $M_1$ , there is a small increase in the information content with time which may or may not be significant. However, it is reassuring to see that this is not due to an improvement in the data with time, as the trend in the spontaneous activity information bias (Fig. 2.18d) is a decrease with time.

#### 2.6.4 *Information in fine vs. coarse distinctions*

The difference between information about fine differences in contrast can be studied by only considering trials where the contrast presented is one of the middle 6 contrasts: 22 %, 25 %, 28 %, 32 %, 35 % and 40 % for  $V_1$  and 27 %, 28 %, 29 %, 31 %, 32 % and 33 % for  $V_4$ . To keep the dimensionality of the stimulus the same, this has been compared to the information across some more separated contrasts: 5 %, 15 %, 22 %, 40 %, 50 % and 90 % for  $V_1$  and 10 %, 15 %, 20 %, 40 %, 50 % and 60 % for  $V_4$ . For  $V_4$ , the group of coarsely differentiated contrasts is the outer most six contrasts, but for  $V_1$  the coarsely differentiated contrasts are alternate

##### 2.6.4.1 *Results*

Comparing Figs. 2.19a and 2.19b where the outer 6 contrasts are included with Figs. 2.16a and 2.16b where all contrasts are included, it seems as if the amount of information has increased, which should not be possible. However, the difference will be due to the difference in trials contained in each of the analyses. In each case, an average of 100 trials per stimulus is used, but since the easier test conditions are presented less frequently, they are under-represented in Figs. 2.16a and 2.16b (about 75 trials per stimulus). Obviously these are more discriminable, so the under-representation comparably reduces the information.

Looking at  $V_1$  (Fig. 2.19), we observe there is much more information for  $M_2$  than  $M_1$ , as we found before. The quality of the data seems to have severely hampered the analysis for  $M_1$ , destroying the the fine differences in the data needed to evaluate the information contained about fine contrast differences (Fig. 2.19c).

Unsurprisingly, there is more information when considering the coarsely distinct contrasts than the finer differences, as the neural activity is bound to be more discriminable for these. For  $M_2$ , there is a small upward trend again for both coarse and fine contrast differences, which may or may not be genuine.

For  $V_4$ , we find there is no information about fine contrast differences in either animal (Figs. 2.20c and 2.20d). The information about the coarse differences is higher

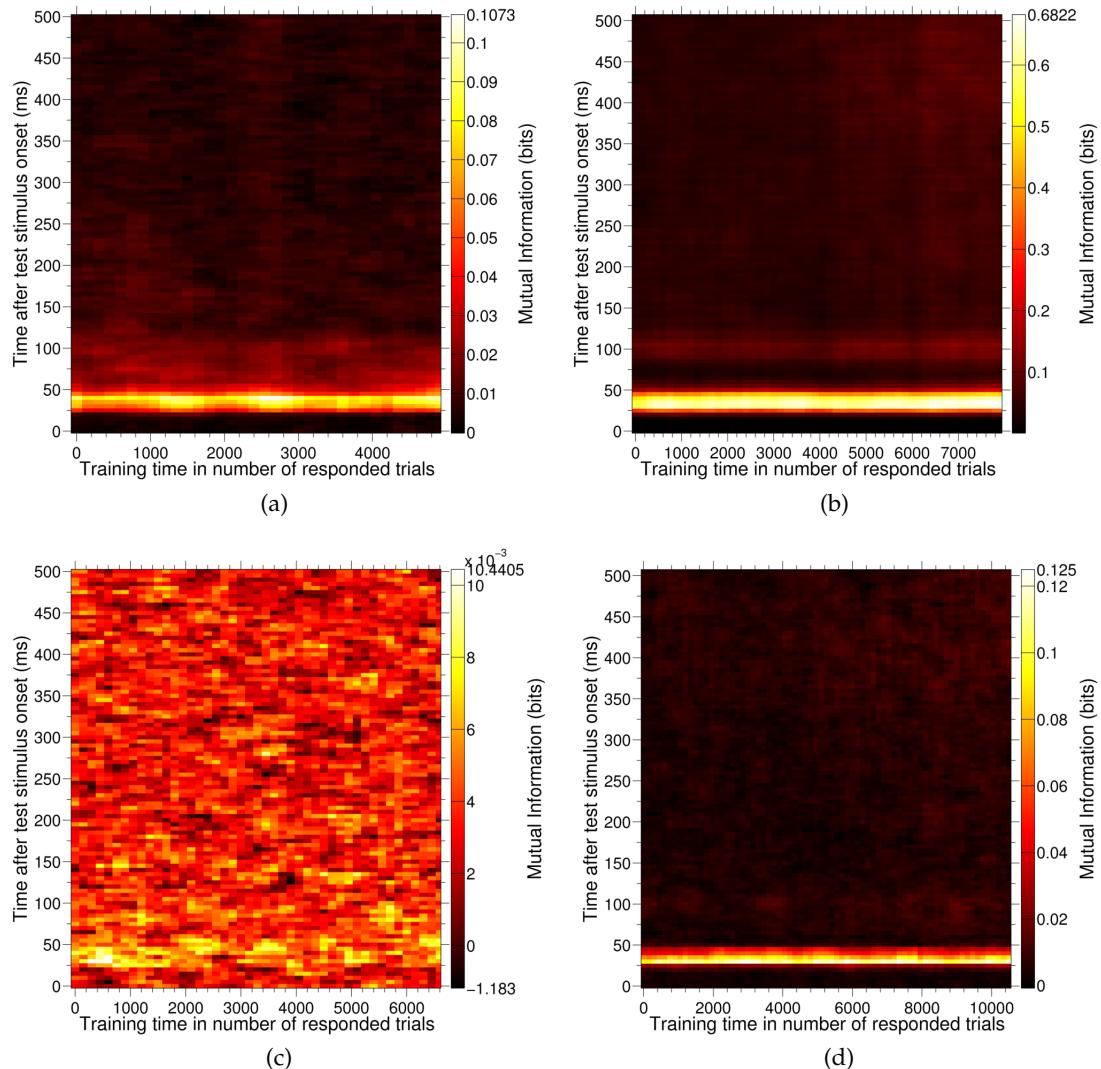


Figure 2.19: **V<sub>1</sub>**: Fine vs. coarse contrast differences. In the top panels, the six contrasts included are {5, 15, 22, 40, 50, 90}%; bottom panels {22, 25, 28, 32, 35, 40}%. An average of 100 trials per stimulus is used in each of these. Left panels are for **M<sub>1</sub>**, right are **M<sub>2</sub>**. In each case, mutual information between the six test stimuli and 20 ms of spiking activity was measured using a spike count code, and bias corrected using the **PT** method.

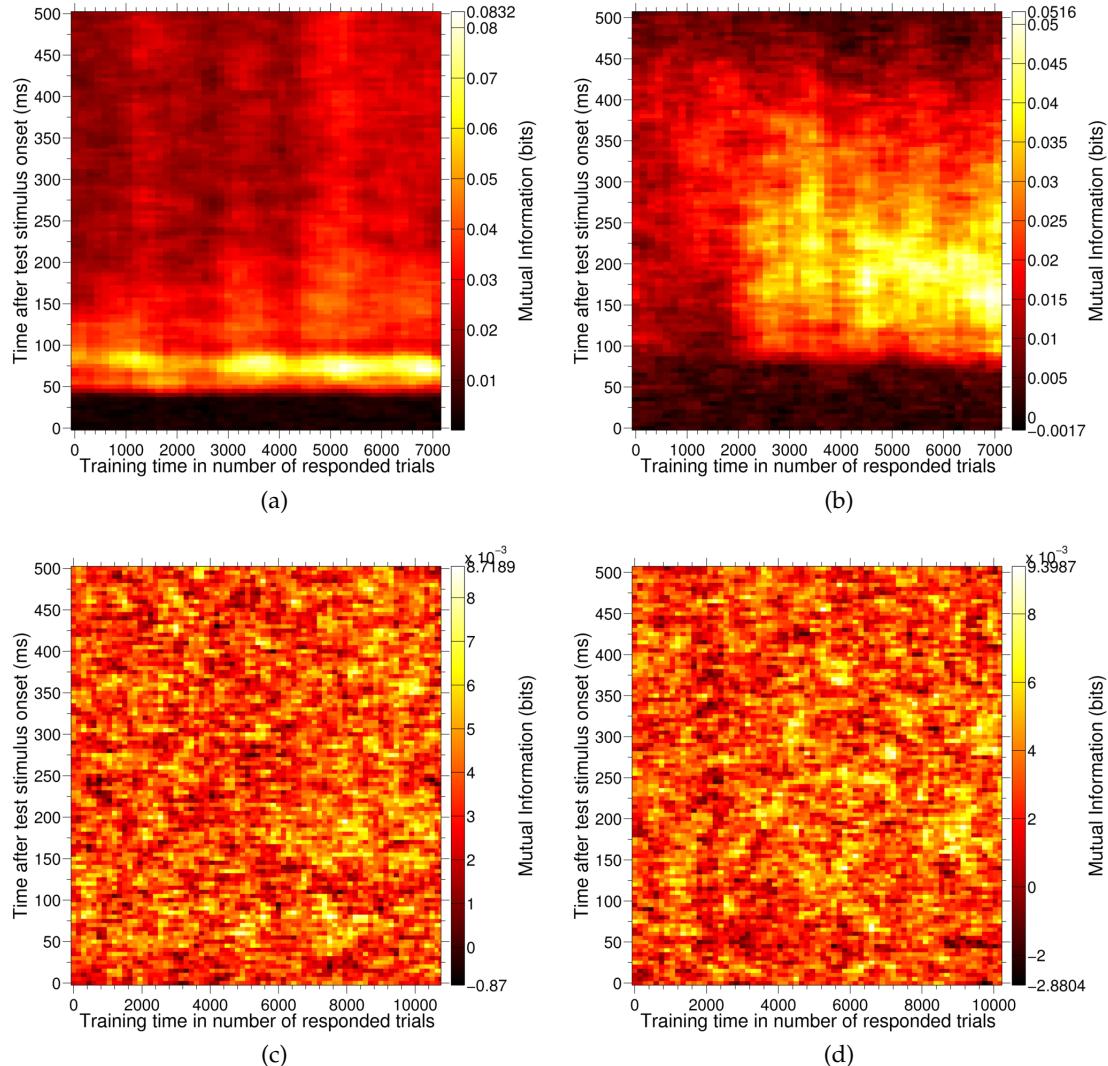


Figure 2.20: **V4**: Fine vs coarse contrast differences. In the top panels, the six contrasts included are {10, 15, 20, 40, 50, 60}%; bottom panels {27, 28, 29, 31, 32, 33}%. An average of 100 trials per stimulus is used in each of these. Left panels are for **M1**, right are **M2**. In each case, mutual information between the six test stimuli and 20 ms of spiking activity was measured using a spike count code, and bias corrected using the **PT** method.

than when all conditions are considered, for reasons discussed above, and these show the same trends as when we analysed all the conditions, in Figs. ?? and 2.18a.

#### 2.6.4.2 Discussion

We did not find that learning was focused on the more difficult contrasts with the fine differences between them. If there is an increase in the information in V<sub>1</sub>, it is not very large and occurs for all groups of contrasts to the same degree. Curiously, for V<sub>4</sub>, it seems that the neurons improve their discriminability for the coarsely differentiated contrasts, but not the fine differences, which is the opposite to what was anticipated. It could be that fine stimuli differences cannot be accurately resolved using the signals from only individual neurons, and concurrent signals from a population of neurons are needed for this to be finer discrimination to be possible.

#### 2.6.5 Information in spike timing code vs. spike count code

Because the possible responses in the spike timing and count codes have different dimensionality, they have different biases (Panzeri et al., 2007) and it is not possible to compare them directly. Consequently, to investigate how much more information there is in a spike timing based code we must consider a set of responses which have the same dimensionality, but lack the timing information.

To do this, the spike-timing response codes are taken, then shuffled across the 5 bins for every individual trial.<sup>4</sup> Since the 5 bins are now in a random order, any information contained in their order is lost. To make ensure the estimate of the information in the spike-timing code with the timing information destroyed was reasonable, the information was measured for five different shuffles and then the mean was taken.

The information contained in the 4 ms level spike timing can then be found by subtracting the mean shuffled information from the unshuffled information.

#### 2.6.5.1 Results

For M<sub>2</sub> V<sub>1</sub>, there seems to be some information in the millisecond-level timing of the spikes during the transient response, but not afterward this has elapsed (Fig. 2.21, right-hand panels). This band due to the transient is clearly well above the variance of the sampling for the rest of the window offsets. However, the information in the transient is only present for the coarse contrasts and not for the fine contrasts. For the fine contrast discrimination in M<sub>2</sub> V<sub>1</sub>, shown in Fig. 2.21f, (and possibly to a lesser

---

<sup>4</sup> The shuffling of bins was performed using the open source Shuffle.m available from <http://www.mathworks.com/matlabcentral/fileexchange/27076-shuffle>.

degree on a couple of the other figures) there is an unusual effect where there seems to be more information in the shuffled bins than the unshuffled bins.<sup>5</sup>

For **M<sub>1</sub> V<sub>1</sub>**, and also **M<sub>1</sub> V<sub>4</sub>**, there seems to be an increase in the information contained in the spike timing during the transient also. However, these results are not as clear-cut as in **M<sub>2</sub> V<sub>1</sub>**.

For **M<sub>2</sub> V<sub>4</sub>**, there is no information in the spike timing measured on the millisecond timescale: not even during the transient response.

For any of these figures there certainly does not seem to be any change in the information contained in the spike timing alone, so it does not seem to be a trait which can be learned.

### 2.6.5.2 Discussion

The observation about there being more information in the shuffled time-bins than when they are in their genuine order (Fig. 2.21f) should be an impossibility because shuffling the response bins around can only destroy information and cannot generate it as it is a random process, unrelated to the stimuli. This could, however, be explained if spikes are similarly timed in the actual data for many stimuli, which causes their responses to be more similar than one would expect by chance. In this case, there might be reliably more information measured in the shuffled bins due to the bias of the measurement being higher when the responses are less predictable.

The information in the precise spike timing for **V<sub>1</sub>** would fit with previous results which have suggested there is information in the latency of the onset response (Reich et al., 2001; Tovée et al., 1993; Rolls and Treves, 2011). Moreover, (Tovée et al., 1993; Rolls and Treves, 2011) indicates that for the primary visual cortex there is no information in the spike-timing beyond that of the spike code except for the limited amount of information given in the latency. In addition, the work presented here suggests that although the latency of the response in **V<sub>1</sub>** conveys information about the presented contrast, this is not a learned or learnable trait.

It was not found that there is more information contained in the spike timing for finer contrast differences. This is contrary to the initial hypothesis, and contradicts several existing pieces of research (Reich et al., 2001; Arabzadeh et al., 2006), but it corroborates the items of research just mentioned indicating there is no information in spike-timing beyond the response latency (Reich et al., 2001; Tovée et al., 1993; Rolls and Treves, 2011). As the bodies of work suggesting there would be differences for fine and coarse discrimination were based on recordings in the rat somatosensory system, whilst the latter is based in the visual cortex, this is not too surprising. How-

---

<sup>5</sup> When this is analysed for the raw data with the artifact included, this is subtly more prominently on several of the plots.

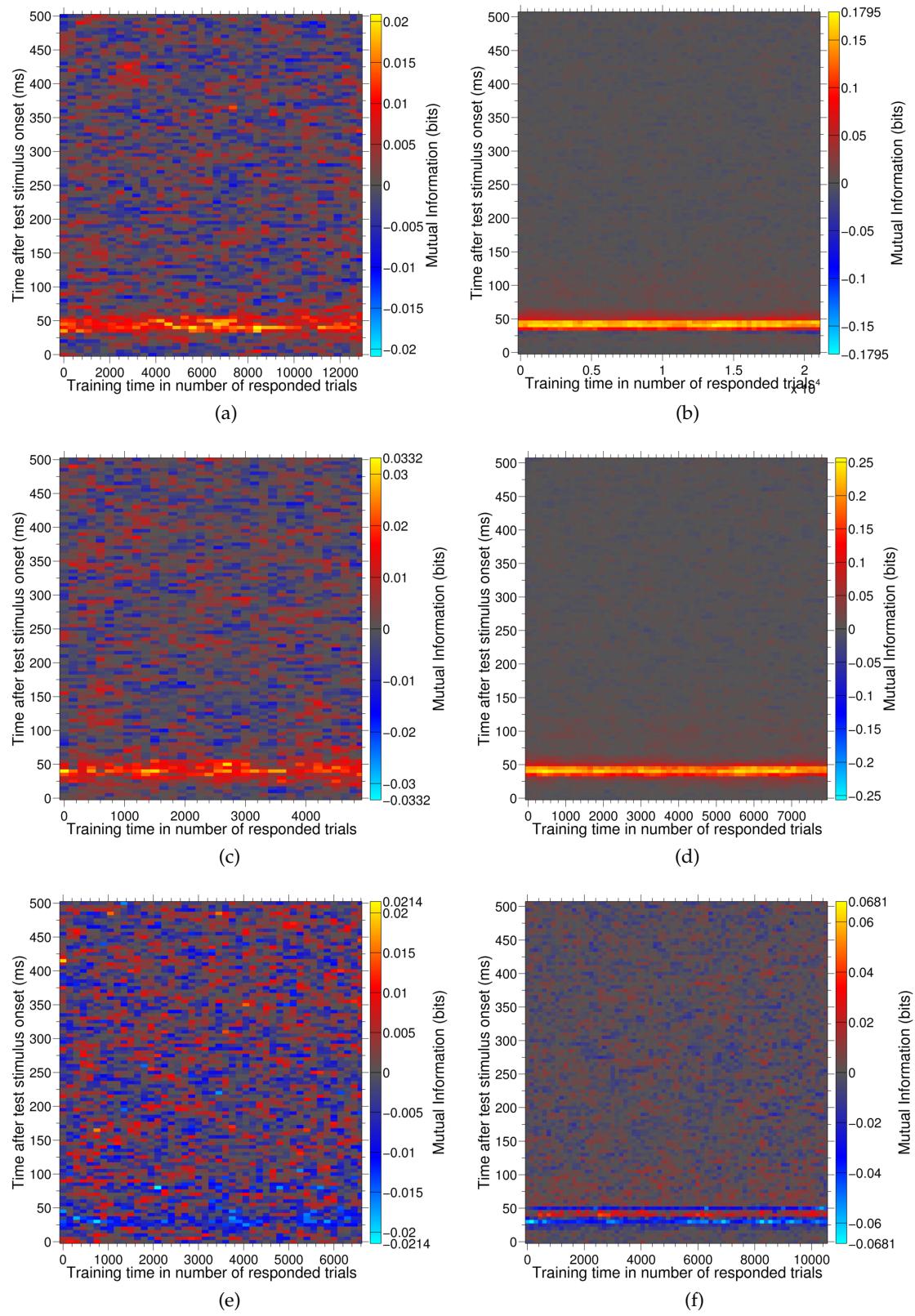


Figure 2.21:  $\text{V}_1$ : Information in millisecond level spike timing. The information with time-wise shuffled bins was subtracted from information in the spike time code with a 20 ms window subdivided into 5 bins. Information was bias corrected using the PT method. Left panels:  $\text{M}_1$ ; Right:  $\text{M}_2$ . Top panels: all contrasts,  $\{10, 15, 20, 25, 27, 28, 29, 31, 32, 33, 35, 40, 50, 60\}\%$ . Centre panels:  $\{5, 15, 22, 40, 50, 90\}\%$ . Bottom panels:  $\{22, 25, 28, 32, 35, 40\}\%$ . An average of 100 trials per stimulus is used in the analysis for each.

ever, as we have already stated, any results from our analysis on spike timing codes are questionable due to the our control of the information in the spontaneous activity.

### 2.6.6 Spontaneous activity

To check whether changes in the data quality between sessions and other differences between sessions could be influencing the results, the information theoretic analysis was also applied the spontaneous activity from the animal. In particular, the spontaneous activity from prior to the sample presentation was used, so the animal had not yet seen the test contrast. Although monkey's brain activity cannot possibly contain any genuine information about the test contrast, we expect to measure a non-zero value for the information content due to the sampling bias.

During the course of the analysis, results were inspected using **PT** and using **QE** for bias correction. The plots were similar in distribution, and each gave a similar magnitude for the information. However, **PT** gave visibly lower variance than **QE**, so only plots using **PT** are presented in this results section.

Using the  $I_{sh}$  approach where the bias is corrected by shuffling bins across trials (Montemurro et al., 2007) was also attempted for the spike timing code, but this increased the variance far too much to be of any practical use. The unexpectedly significant increase in variance is probably due to the nature of the correlations between the bins which are being shuffled.

We also compared some of the results both with the monitor artifact left in the data and with the dataset redacted to have it removed. There was no appreciable difference for any of the plots, though the amount of information went down when the data was redacted due to the loss of some of the spikes.

## 2.7 DISCUSSIONS

We now discuss the findings of the analysis described in the previous chapter, and suggest ways in which this work may proceed in the future.

### 2.7.1 Validity of results

From Figs. 2.16–2.18, we can conclude several things about the reliability of our other results. Firstly, it seems the data for **M<sub>2</sub>** is more trustworthy than that of **M<sub>1</sub>**. Secondly, from the information measured in the spontaneous activity, it seems the results for the spike timing code cannot be trusted, certainly not any changes which seem to

occur with learning. Thirdly, this latter point may call into jeopardy the reliability of the results for the spike count code, since these problems seem to be inherent to the raw data, though the information in the spike count code is more robust due to its fewer possible response vectors.

### 2.7.2 Discussion of results

The analysis has demonstrated there is far more information given by the onset transient response, which is in keeping with previous findings (Müller et al., 2001).

The finding that information increases more rapidly with learning in  $V_4$  than  $V_1$  is in line with our hypothesis made earlier.

The result that the peak in information in  $V_4$  neurons moves to being sooner after stimulus onset (Fig. 2.18a) is an entirely novel finding. Although the values for the information here are very low — only 0.02 bits — this is the average over many channels, only a couple of which exhibit the increase with time. The more responsive channels have information peaking at around 0.65 bits, but the non-responsive channels pull this average down. I have also done some work to identify which channels should be excluded on the basis of their  $d'$ , but this is not presented here.

### 2.7.3 Future Work

Further work will need to be done on this analysis to try and fix the inconsistencies in data between days so they can be studied together more easily. Also, work will need to be done to establish if any of the results presented here are genuinely statistically significant.

Here we describe some follow-up work which could be performed, either by myself or the perceptual learning lab group.

#### 2.7.3.1 Elimination of artifacts from continuous data

Fig. 2.3d shows that the current method of correcting for the monitor artifact is flawed. If the monitor artifact were an artifact similar to previously known artifacts, such as the “reward artifact” generated by the water dispenser, a large potential is induced in the electrode, which can be registered as a spike since it exceeds the threshold. However, this sort of behaviour would not cause a reduction in the number of spikes preceding the artifact incident, as seen here and in many of the sessions for other channels too.

To create the preceding reduction in spikes over a period of around 0.2 ms, the measured potential must be reduced, suppressing the spikes which are being elicited so they are not passing threshold and being detected. A positive electric potential following this could increase the proportion of spikes meeting threshold, and resulting in the problematic sharp peak in the number of spikes. Lending credence to this theory is the way the spikes detected as the monitor artifact have the same waveform as the usual spike, but have an amplitude towards the lower end of what would be expected.

If this is the effect we are witnessing, this can be fixed in a manner similar to the way the monitor artifacts were removed in this paper. Instead of considering the spiking data, we will need to look at the continuous data. We can take the modulo of the times with respect to the monitor refresh rate again, and bin the data points together with a bin width the reciprocal of the sampling frequency, but instead of taking the number of data points in each bin (which will be the same since there is always a continuous voltage value even if there isn't a spike), we take the mean of the voltages. The amount of modification made on each trial to the voltages should then appear as a fluctuation in the mean voltage recorded and can consequently be subtracted from the original data.

### *2.7.3.2 Normalisation of firing rates*

Combining trials from different sessions in one analysis has problems because the recording quality varies from day to day. Can possibly be normalised by adjusting the spike detection threshold so that there is always the same spontaneous activity exceeding the threshold. This is justifiable because even if neuronal firing rates change greatly due to perceptual learning, homeostasis should keep the spontaneous activity rates relatively consistent.

This will hopefully make the measured information content of the spontaneous activity be consistent throughout training. If this is the case, the result will be much more trustworthy results. This will be especially important in improving the potential of the results for M<sub>1</sub>, and for the spike-timing code, potentially yielding some more interesting results.

### *2.7.3.3 Comparison with psychometrics*

An interesting comparison would be to see how the information correlates with the behaviour of the monkey. In particular, it is suspected that there might be a correlation between the magnitude of the maximum information available to the animal in V<sub>4</sub> and the performance of the monkey in the task. Also, there is reason to suspect

there might be a correlation between the latency of the information in V4 and the response time of the monkey, and I tentatively hypothesise this will be the case.

#### 2.7.3.4 Statistical analysis of significance of results

It will be very important to clarify the statistical significance, if any, of the results tentatively presented here. The statistical significance of mutual information can be estimated by using bootstrapping, which is where the stimuli and response vectors are shuffled and paired together at random (Ince et al., 2011). Because some combinations will happen to provide more discriminability than others, the distribution of the mutual information fits a Gaussian distribution with a certain mean and standard deviation. By taking the distribution of information with bootstrapping, the mean and variance can be computed, and the actual value of the information from the data can be compared with this to see if its distance above the mean is statistically significant.

Using this method, we can also pick out a value of information which is statistically significant and see how long after test presentation it takes for the information about the test stimulus to become significant. This will provide a metric for information latency.

#### 2.7.3.5 Further directions

There are several different routes down which the project could be more broadly extended.

- Analysis of information about the test contrast contained in the multi-unit activity (MUA) and LFP signals contained in the raw recordings from which the spiking data was extracted.
- Analysis of roving task. This could be done by considering the activity during test presentation identifying the trial condition as being dependent on both test and sample contrasts. To see how important the sample stimulus is to the brain activity during the test presentation, we would subtract from this an estimate of the information where condition is changed by keeping the test contrasts the same, but shuffling the sample contrasts.
- Information rate in bits per spike (Rolls and Treves, 2011).
- Population wide information, using a decoding approach. There are too many channels to use a direct information theoretic approach (Quiroga and Panzeri, 2009). If we were to use information theory directly, even with binary bins for every channel, there are  $2^{20}$  possible responses and we would need in excess of 2 million trials per condition to have a reasonable estimate of the information

content. The dataset is not large enough for this, and even if it were we would have missed the changes due to perceptual learning as they only occur in the first 20,000 trials.

- Examining correlations between neurons.
- Investigate whether firing rate for different contrasts becomes more discriminable due to the means becoming more distant or due to the variability in rate for each contrast being reduced.

#### 2.7.4 *Summary*

The project has expanded on existing literature and demonstrated that the millisecond-level timing of spikes from individual neurons, both in  $V_1$  and  $V_4$ , is not important for contrast discrimination tasks, but the firing rate is important. It has also been demonstrated that little information regarding the discrimination between similar contrasts can be gained from the spiking activity of individual neurons, even during/after perceptual learning has occurred. In  $V_1$ , there is no discernible change in contrast information in individual neurons at all, but in  $V_4$  there is an increase with perceptual learning in the magnitude of the peak information after stimulus onset, and the peak occurs sooner after stimulus onset as well.

However, the analysis indicates the quality of the results is good for one monkey ( $M_2$ ), but not so good for the other ( $M_1$ ). The discrepancy between the two datasets needs further investigation, and the conclusions made here need to be validated against other studies.

## 2.8 OTHER DISCUSSION

Based on data from M<sub>2</sub> V<sub>4</sub>, one might conclude our results to offer some corroboration with those of Gu et al. (2011), since we find a decrease in noise correlations and an increase in decoder performance. However, comparing the performance of the decoder to a decoder based on shuffled data without the noise correlations present suggests that the improvement in decoder performance is not due to the relatively small reduction in noise correlation observed, but is from other sources.

The data from M<sub>2</sub> V<sub>1</sub> does not support a “reduction in noise correlation” hypothesis either, and indicates that neural spike rates across the V<sub>1</sub> population recorded from are no more informative after training than they were before. Together, results from M<sub>2</sub> V<sub>1</sub> and V<sub>4</sub> suggest that information in V<sub>1</sub> is consistent throughout training, but the ability of V<sub>4</sub> to read out the information in V<sub>1</sub> improves over the same period, leading to an improvement in behavioural performance.

However, these findings are not supported by the data from M<sub>1</sub>.

For both animals, from the data from V<sub>4</sub> we find there is a statistically significant level of agreement between decoded and behavioural trial-to-trial responses after training but not before, which shows the neural activity in V<sub>4</sub> and the behavioural responses are correlated. This implies that the monkey becomes dependent on the activity of the population of neurons for which the recordings are representative. That the agreement is better without shuffling could be taken to mean the neural correlations are important in the determination of the animal’s response, however shuffling does by its very nature destroy the correspondence between the trials given to the decoder and those experienced by the animal.

## CORTICAL OSCILLATIONS IN CSD ACROSS V<sub>1</sub> LAMINA

---

### 3.1 INTRODUCTION

The cortical column is widely regarded as the fundamental processing unit of the neocortex (Mountcastle, 1957). Under this hypothesis, there is a common microcircuit spanning the depth of the cortex which is repeated across the cortical plane. As the circuitry of the microcolumn is expected to have structural and functional similarities across the sensory modalities, therefore understanding this generic circuitry of the columnar computation will have far reaching impacts. However, despite some progress towards understanding knowing the different types of neurons in present in different layers [CITATION], and the distribution of their most prominent inter-connections [CITATION], and a structural wiring diagram for the cortical microcircuit [CITATION], is still unknown. Furthermore, the functional structure and computation of the microcircuit, which is to say the purpose of the processing in each cortical layer, is also still unknown. In this paper, we aim to elucidate the functional structure of the cortical layers in V<sub>1</sub> by examining the information contained in population activity at the various layers using LFPs.

LFPs are thought to reflect an integration of the membrane depolarisation in the neurons surrounding the electrode location. The LFP captures changes within the dendritic trees of neighbouring neurons as well as the soma. The low frequency LFP (XXXX Hz) captures slower changes in the population activity, and reflects more of the dendritic level of processing, integrated over a larger region than the high frequency LFP (Leski, Linden, Tetzlaff, Pettersen, & Einevoll, 2013). Since early EEG studies, it is has been hypothesized that different frequencies convey complementary information [CITATION] originates with EEG studies, but and this has subsequently been extended to the LFP [CITATION NEEDED].

We previously found that in the macaque V<sub>1</sub> there are two LFP frequency bands, 1Hz to 8Hz and 60Hz to 100Hz, which encode independent information in the macaque V<sub>1</sub> about natural stimuli (Belitski et al., 2008). In this study we expand the previous study by studying information as a function of cortical depth, and identify one aspect of natural scenes which is encoded differently by the two cortical frequency bands. We hypothesised the two bands of information are generated through different cortical processes and originate at different locations in the cortex. In this

study we expand the previous study by studying information as a function of cortical depth, and identify one aspect of natural scenes which is encoded differently by the two cortical frequency bands.

Recent work has shown stimulation in  $V_1$  induces gamma activity in  $V_4$  (feed-forward), whilst stimulation in  $V_4$  induces alpha oscillations in  $V_1$  (feedback) (van Keroerle et al., 2014).

## 3.2 EXPERIMENTAL METHODS

### 3.2.1 Ethics Statement

Data was collected from  $V_1$  in four healthy rhesus monkeys (*Macaca mulatta*; four males 8 kg to 11 kg; 10 years to 12 years). All the experimental procedures were approved by the local authorities (Regierungspräsidium, Baden-Württemberg, Tübingen, Germany; Project Number KY4/09) and were in full compliance with the guidelines of the European Community (EUVD 86/609/EEC) and were in concordance with the recommendation of the Weatherall report for the care and use of non-human primates (Weatherall, 2006). The animals were group-housed in an enriched environment, under daily veterinarian care. Weight, food and water intake were carefully monitored on a daily basis.

### 3.2.2 Anesthesia for Neurophysiology Experiments

The anesthesia protocol for all the experimental procedures have been described previously (Logothetis, Guggenberger, Peled, & Pauls, 1999; Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001). Briefly, glycopyrrolate ( $0.01 \text{ mg kg}^{-1}$ ) and ketamine ( $15 \text{ mg kg}^{-1}$ ), were used previous to general anesthesia. Induction with fentanyl ( $3 \text{ mg kg}^{-1}$ ), thiopental ( $5 \text{ mg kg}^{-1}$ ) and succinylcholine chloride ( $3 \text{ mg kg}^{-1}$ ), animals were intubated and ventilated using a Servo Ventilator 900C (Siemens, Germany) maintaining an end-tidal CO<sub>2</sub> of 33 mm Hg to 35 mm Hg and oxygen saturation above 95 %. The anesthesia was maintained with remifentanil ( $0.5 \text{ } \mu\text{g kg}^{-1} \text{ min}$  to  $2 \text{ } \mu\text{g kg}^{-1} \text{ min}$ ) and mivacurium chloride ( $2 \text{ mg kg}^{-1} \text{ h}$  to  $6 \text{ mg kg}^{-1} \text{ h}$ ) which ensured no eye movement during electrophysiological recordings. The anesthetics dosage were established by measuring stress hormones and were selected to ensure unaffected physiological response at normal catecholamine concentrations (Logothetis et al., 1999). In addition, it has been shown that using remifentanil has no significant effect on the neurovascular and neural activity of brain areas that do not belong to the pain matrix (Goense & Logothetis, 2008; Zappe, Pfeuffer, Merkle, Logothetis, &

Goense, 2008). In particular, visual cortex does not bind remifentanil. We monitored the physiological state of the monkey continuously and kept within normal limits. Body temperature was tightly maintained at 38 °C to 39 °C. Throughout the experiment lactate Ringer's (Jonosteril, Fresenius Kabi, Germany) with 2.5 % glucose was continuously infused at a rate of 10 ml kg<sup>-1</sup> h<sup>-1</sup> in order to maintain an adequate acid-base balance and intravascular volume and blood pressure were maintained by the administration of hydroxyethyl starch as needed (Volulyte, Fresenius Kabi, Germany).

We used anesthetized animals because the preparation allows longer data acquisition times and to associate particular neural events to specific stimulus features without the strong effects of animal cognitive state, including effects of attention and arousal that would introduce additional complication in the interpretation of signals.

### 3.2.3 *Visual Stimulation*

A few drops of 1 % cyclopentolate hydrochloride were used in each eye to achieve mydriasis. Animals were wearing hard contact lenses (Wöhlk-Contact-Linsen, Schönkirchen, Germany) to focus the eyes on the stimulus plane. The visual stimulation in all experimental sessions was presented in the eye with stronger ocular preference of recording sites. The stimulus was presented using either an in-house custom-built projector (SVGA fibre-optic system with a resolution of 800x600 pixels, a frame rate of 30 Hz), or a CRT monitor (Iiyama MA203DT Vision Master Pro 513, frame rate 118 Hz) placed at eye level, 50 cm in front of the eye. We found the same results with both display devices, except that monitor refresh with the 30 Hz stimulus induced cortical oscillations at 30 Hz. Since this is the result of using an artificial stimulus with a low refresh rate (a well-known issue at this stimulus frequency), we removed this from the data (see Artefact Removal) and pooled the results across all sessions. The visual stimulus consisted of high contrast (100 %), gamma corrected, fast-moving, colourful movie clips (no soundtrack) from commercially available movies. Stimulus timings were controlled by a computer running a real-time OS (QNX, Ottawa, Canada). Stimulus-on periods of 120 s (5 sessions; 1 session: 40 s) were interleaved with stimulus-off periods (isoluminant grey screen) of 30 s.

### 3.2.4 *Neurophysiology Data Collection and Analysis*

The electrophysiological recordings were performed by doing a small skull trepanation, after which the dura was visualized with a microscope (Zeiss Opmi MDU/S5, Germany) and carefully dissected. The electrodes were slowly advanced into the

visual areas under visual and auditory guidance using manual micromanipulator (Narashige Group, Japan). Electrodes consisted of laminar probes (NeuroNexus Technologies, Ann Arbor, USA). These electrodes contained 16 contacts on a single shank 3 mm long and 150  $\mu\text{m}$  thick. The electrode sites were spaced at 150  $\mu\text{m}$  apart, with a recording area of 413  $\mu\text{m}^2$  each. We used a flattened silver ( $\text{Ag}$ ) wire, which was positioned under the skin, as reference electrode (Murayama et al., 2010). The recording access was filled with a mixture of 0.6 % agar dissolved in sodium chloride ( $\text{NaCl}$ ) 0.9 %, pH 7.4 solution which guaranteed good electrical connection between the ground contact and the animal (Oeltermann, Augath, & Logothetis, 2007). The impedance of the contact points was always measured during the experiments and ranged from 480  $\text{k}\Omega$  to 800  $\text{k}\Omega$ . The signals were amplified and filtered into a broadband of 1 Hz to 8000 Hz (Alpha-Omega Engineering, Nazareth, Israel) and then digitized at 20.833 kHz with 16 bit resolution (PCI-6052E; National Instruments, Austin, TX).

### 3.2.5 Luminosity Function

In order to best approximate the luminosity perceived by macaques, we relied on analogies with the human visual system. Research with humans suggest the luminosity function is linearly related to the long (L) and medium (M) cone activation, and independent of the short (S) cone activation (Stockman, Jagle, Pirzer, & Sharpe, 2008). Furthermore, the weighting of L and M activations towards perceived luminance is believed to be similar to the L:M ratio in the individual (Stockman et al., 2008). Old world monkeys such as macaques have an L:M ratio which is approximately 1:1 (Dobkins, Thiele, & Albright, 2000), so we assumed a luminosity function equally weighed between the L and M cone activations,  $Y = L + M$ . The 10° cone fundamentals of Stockman & Sharpe (2000) were used since the cone fundamentals of old world monkeys are known to be very similar to humans (Dobkins et al., 2000). By taking the product of the emission spectra for pure red, green and blue with the luminosity function, integrating over wavelength and normalising, we obtained the following equations for relative luminance in terms of pixel intensity for the two devices used in the experiment .

$$Y_{\text{projector}} = 0.2171 \cdot R + 0.6531 \cdot G + 0.1298 \cdot B$$

$$Y_{\text{CRT}} = 0.1487 \cdot R + 0.6822 \cdot G + 0.1691 \cdot B$$

### 3.2.6 Artefact Removal

An artefact removal procedure was performed to reduce the effects of line noise (one session) and monitor refresh (the three sessions with 30 Hz stimulus). Artefact frequencies were identified by large, localised peaks in the power spectral density, which was computed with the periodogram method (see Supplementary Table 1). In each case, the average artefact waveform was found and subtracted from the recorded signal. To correct for phase shifts of the artefact, the averaging and subsequent subtraction were performed in blocks of 50 artefact periods with a phase chosen to maximise the cross-covariance of the signal with the artefact waveform.

### 3.2.7 Current Source Density

The current source density (**CSD**) was computed using the inverse **CSD** method (Petersen, Devor, Ulbert, Dale, & Einevoll, 2006). To compute this, we used a  $\delta$ -source model of local field generation with a diameter of 500  $\mu\text{m}$ , chosen to correspond to the effective size of columnar activity (Horton & Adams, 2005; Lund, Angelucci, & Bressloff, 2003). Since this method requires an even spacing between voltage measurements, gaps caused by faulty recording contacts in the electrode were filled in with a local average (Wójcik & Leski, 2010). A homogeneous cortical conductivity of  $0.4 \text{ S m}^{-1}$  was assumed (Logothetis, Kayser, & Oeltermann, 2007). The agar solution placed on top of the recording access point had an **NaCl** concentration of  $9 \text{ mg mL}^{-1}$ , and the conductivity of this was estimated to be  $2.2 \text{ S m}^{-1}$  (Kandadai, Raymond, & Shaw, 2012). The **CSD** was spatially smoothed with a three-point Hamming filter (add ref that justifies this smoothing).

### 3.2.8 Multiunit Activity

Fig. 3.1: **MUA** was calculated by band-passing the voltage recording between 900 and 3000 Hz with a zero-phase sixth-order Butterworth filter, taking the absolute value, applying a 300 Hz low pass third-order Butterworth filter, and then downsampling. This yields a smoothed spike rate, analogous to a population firing rate.

Fig. 3.3: **MUA** was calculated by downsampling by a factor of 3, band-passing the voltage recording between 900 Hz to 3000 Hz with a zero-phase sixth-order Butterworth filter, taking the absolute value, downsampling by a further factor of 12.

### 3.2.9 Receptive Field Locations

The spatial receptive fields (RFs) were found by reverse correlating the MUA and the pixel-by-pixel Z-scored frame-by-frame difference in luminance with a fixed lag of 66.7 ms. The rate of change in luminance was used because it is known to correlate well with thalamic drive. For each session, the RF centre was located using the average of the reverse correlation across all cortical channels.

### 3.2.10 Identification of Cortical Laminae

Depth calibration of the electrode was performed by considering the spikes and CSD induced by both the onset of the movie and by 100 ms full-screen flashes (6 s flash interval). From the measured potentials, we determined the CSD and spike densities. Spikes were detected by high pass filtering the raw signal above 500 Hz with a zero-phase eighth-order Butterworth filter, and classifying any points more than 3.5 standard deviations above the mean signal during pre- and post-stimulus periods as a spike, with a minimum inter-spike-interval of 1 ms. The majority of thalamic afferents in V<sub>1</sub> stimulate layer 4 of V<sub>1</sub> (L<sub>4</sub>) (indirectly: see Hansen et al. (2012)), with the first cortical response manifesting at layer 4C $\alpha$  (E. M. Callaway, 1998), resulting in an initial current sink and first burst of spiking activity located here. For each recording session, we found the contact exhibiting the first spiking and CSD response, the centre of the most responsive region, and the centre of the first CSD sink for both the movie onset and flash evoked activity. We took the average of these 8 locations and identified the closest electrode contact as the location of layer 4C $\alpha$ . We estimated the laminar location of the rest of the recording depths by cross-referencing literature describing average thickness of cortical laminae in *Macaca mulatta*, area 17, (Lund, 1973). From this, the cortical depth was divided into 3 broad regions: supragranular (SG; layers 1–3), granular (G; layer 4), and infragranular (IG; layers 5–6). High-resolution magnetic resonance imaging (MRI) scans of two of the animals were used to determine their cortical thickness at the recording location (see Supplementary Materials).

### 3.2.11 Identification of cortical laminae

To identify the depth of each contact, we measured the potential evoked in response to the onset of the movie clip, and in response to full-screen maximum-luminance 100 ms flash stimuli with a 6 s interval. From the measured potentials, we identified the boundary between the granular (G; layer 4) and infragranular (IG; layers 5–6)

regions as the source/sink reversal in the evoked current source density (**CSD**; see Experimental Procedures and Supplementary Materials). We estimated the location of the boundary between the granular and supragranular (SG; layers 1–3) regions by cross-referencing literature describing average thickness of cortical laminae in *Macaca mulatta*, area 17 (Lund, 1973).

### 3.2.12 *Power as a function of depth and frequency*

To compute power and information as a function of temporal frequency, the cortical data (**LFP** and **CSD**) were filtered in a series of bands each with a fractional bandwidth of 50 %, because cortical power falls off rapidly with frequency in a  $1/f$  relationship. Each successive band begins and ends with frequencies 1.291 times higher than the last, so that each band has 0 % overlap with bands further away than its immediate neighbours and a 44 % and 56 % overlap with its preceding and succeeding bands respectively. The data was filtered with a zero-phase sixth-order Butterworth filter, after which the instantaneous power was estimated by taking the squared absolute value of the Hilbert transform. The power in each band was integrated over a series of 50 ms windows, centred at the time of each frame change in the movie (once every 33 ms, leading to a 50 % overlap of neighbouring windows). The power in the 4 Hz to 16 Hz and 60 Hz to 170 Hz bands was computed similarly. Fig. 3.3A–B are plotted with power values averaged over all windows and trials, then expressed in decibels relative to the average power 1.5 Hz to 248 Hz (estimated by summing the power in alternate bands). Throughout Figures 3 and S3, datapoints are shown at the band centres.

### 3.2.13 *Information as a function of depth and frequency*

Power in each band was computed as above, then for each frequency band and depth we took a 10-bin histogram of the power across all the 50ms windows for all repetitions. The bin edges were chosen such that 10 % of the distribution fell into each bin, and the identity of which bin the window was allocated into was taken to be its “stimulus”. We found the mutual information between the response and which frame was on screen at that time — the “stimulus” — by computing the Shannon information using the information breakdown toolbox (Magri, Whittingstall, Singh, Logothetis, & Panzeri, 2009). Bias due to undersampling was corrected for using the **PT** method (Treves & Panzeri, 1995). Each information calculation was also bootstrapped 20 times with a randomly shuffled mapping of stimulus to response (also bias-corrected). To ensure the amount of information was statistically significant, we checked each in-

formation estimate exceeded the bootstrap mean by more than 3 standard deviations of the bootstrap values. The bootstrap mean was then subtracted from the estimated information, to counter any residual bias.

### 3.2.14 *Cortical Distribution of Power*

For each session, the distribution of power across the cortical depth (2A–B right-hand insets) was determined by normalising the power at each depth by the summed power across all cortical depths for that band. We then took an average across sessions, weighted by the number of cortical recording sites in each session to prevent faulty (omitted) electrode contact sites from distorting the result.

### 3.2.15 *Information Redundancy*

Information redundancy was computed with the same stimuli windows as used in the information calculations. Let  $S$  denote the set of stimuli, and let  $X$  and  $Y$  each be the set of powers during each stimulus in one of the frequency bands at a particular depth. The information in each  $I(X; S)$  and  $I(Y; S)$  was computed in the same manner as above. The information in the joint distribution  $I(\{X, Y\}; S)$  was computed by considering each combination of the binned  $X$  and  $Y$  as a different response, yielding a total of 100 different responses for  $\{X, Y\}$ .

The relative redundancy is then defined as

$$\text{Redundancy} = \frac{I(X; S) + I(Y; S) - I(\{X, Y\}; S)}{I(\{X, Y\}; S)}$$

and was computed using the information breakdown toolbox (Magri et al., 2009).

### 3.2.16 *Information about Spatial Components*

The method to find the change in luminance in each spatial frequency band is illustrated in Fig. 3.4. First, we took the 2D fast-Fourier transform of a 224px square from the movie. A fourth-order Butterworth filter with a width of one octave was applied using a mask in the Fourier domain, and the result was projected back to the spatial domain. We then took the pixel-wise difference between each spatially filtered pair of consecutive frames. To provide a measure of the amount of change in luminance at this spatial resolution, we took the absolute amount of change in each pixel

and summed this within a  $2^\circ$  diameter circular window centred at the receptive field location.

Applying this to the entire movie provided a temporal sequence of luminance changes in each spatial range. Similar to before, we took a 10-bin histogram and took the identity the bin in which each luminance change fell to be the “stimulus”. The mutual information between this stimulus and the neural response — the power within 4 Hz to 16 Hz and 60 Hz to 170 Hz frequency bands — was computed with a 67ms lag between stimulus and response.

### *3.2.17 Information about Fine and Coarse Luminance Changes*

Coarse and fine luminance changes in the stimulus were isolated in the same manner as the spatial components above, but using a low-pass ( $<0.3$  cpd) and high-pass ( $>1$  cpd) fourth-order Butterworth filter respectively. For both the 4 Hz to 16 Hz and 6 Hz to 170 Hz CSD powers, we computed the correlation and mutual information with the coarse and fine luminance changes, and averaged across sessions.

### *3.2.18 Information lag between granular and infragranular regions*

The information about fine and coarse stimuli contained in 4 Hz to 16 Hz and 60 Hz to 170 Hz neural frequency bands was computed as a function of the lag between stimulus and response, in steps of 1.73ms. For each cortical recording depth, we found response lag at which the information was at its maximum. For each session, the response peak-lag was averaged across the three electrode contacts in infragranular (IG) and also averaged across the three electrodes in supragranular (G). A paired t-test was performed across all 6 sessions to test whether the G information peaked with a shorter lag than the IG information.

### *3.2.19 Cross-Frequency Phase-Amplitude Coupling*

Strength of cross-frequency coupling was measured using the Modulation Index (Tort, Komorowski, Eichenbaum, & Kopell, 2010). CSD data was filtered for two bands, 4 Hz to 16 Hz and 60 Hz to 170 Hz, using a zero-phase sixth-order Butterworth filter, and the instantaneous phase of 4 Hz to 16 Hz and envelope amplitude of 60 Hz to 170 Hz were each estimated using a Hilbert transform. We took a histogram of the 4 Hz to 16 Hz phase datapoints with 16 bins each of width  $\pi/8$  radians, and took the average of the 60 Hz to 170 Hz amplitudes simultaneous with the phase datapoints

in each bin. This provides a distribution of amplitude at one depth as a function of phase at another. The Modulation Index is then the normalised Kullback-Leibler divergence of this distribution from a uniform distribution.

### 3.3 RESULTS

Four anesthetized monkeys (*Macaca mulatta*) were presented with a Hollywood movie clip, repeated 40–150 times, whilst we recorded neural activity in the **V1** with a multicontact laminar electrode. Each electrode housed 16 equally spaced (150  $\mu\text{m}$ ) contacts spanning a total depth of 2250  $\mu\text{m}$ , and was inserted perpendicular to the cortical surface (Fig. 3.1A). We identified the depth of each probe using the **CSD**, and identified **G**, supragranular (**SG**), and **IG** cortical regions based on literature (see Experimental Methods for details). We performed reverse correlation between the rate of change of luminance for each pixel in the movie and the **MUA** at each contact site to find the spatial-**RF** (Fig. 3.1B). The **RF** locations did not vary with depth, indicating that all electrode contacts were recording from the same cortical column.

#### 3.3.1 *Distribution of information across depth and frequency*

Fig. 3.1C shows, at three cortical depths, **CSD** traces from eight example trials during a portion of the stimulus. The traces have been filtered within three frequency bands: 4 Hz to 16 Hz, 28–44 Hz and 60 Hz to 170 Hz. One can observe that the low-frequency activity repeats across trials for the **G** and **IG** depths. Activity in the 28–44 Hz range is inconsistent at all depths, and does not seem to be stimulus modulated. The envelope amplitude of the 60 Hz to 170 Hz band is also consistent across trials, most clearly for the **SG** depth. The mutual information We quantified these observations by computing the amount of information about the movie contained in the neural activity.

We computed information about which frame is currently on screen in various frequency components of the **LFP** and **CSD** (see Experimental Methods). Excluding boundary effects at the top and bottom of the cortex where white matter contaminates estimates, power is fairly smooth across depth and decays as frequency increases (Fig. 3.2A-B). However, the information contained in the power does not have such a smooth distribution and differs from this in both space and frequency domains. Instead, information is contained in specific frequencies at specific depths, in a similar manner for **LFP** and **CSD** (Fig. 3.2C–D), with prominent maxima in the 4 Hz to 16 Hz range at the top of the **G** region, and the 60 Hz to 250 Hz range near the top of the **SG** region. Additionally, there are local maxima in **IG** for both the 4 Hz to 16 Hz and 60–250 Hz ranges. These results are consistent across sessions (Sup Fig. 3.2). As the

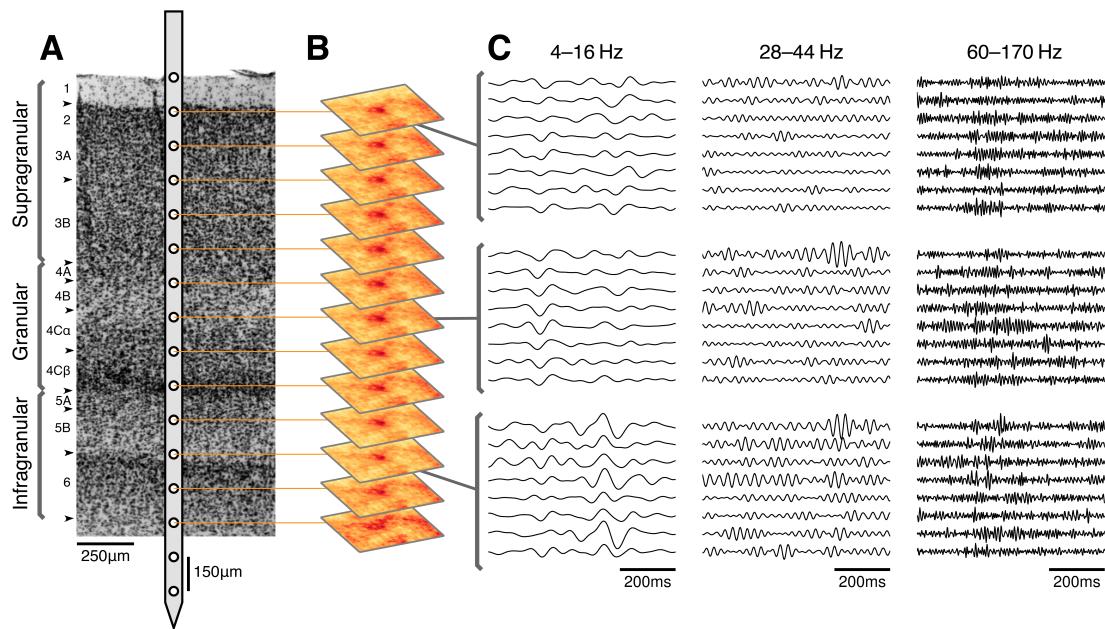


Figure 3.1: *Overview of data collection and example data.* A: Illustration of experimental recording setup, showing approximate locations of electrode contacts in relation to a Nissl-stained section of macaque V1 cortex. Boundaries between cortical laminae are indicated with arrowheads. (Stain reproduced from Tyler et al. 1998, with permission.) (Note: Electrode width is not to scale.) B: Receptive field locations were consistent across the cortical depth. Location of receptive field for each cortical recording site was identified by reverse correlating the MUA with the luminance changes of each pixel in the movie (session E07nm1). C: Example CSD traces from simultaneous recordings at three cortical depths for eight repetitions of a movie fragment (session H05nm7). The data are split into three temporal frequency bands (4 Hz to 16 Hz, 28–44Hz and 6 Hz to 170 Hz, see Methods).

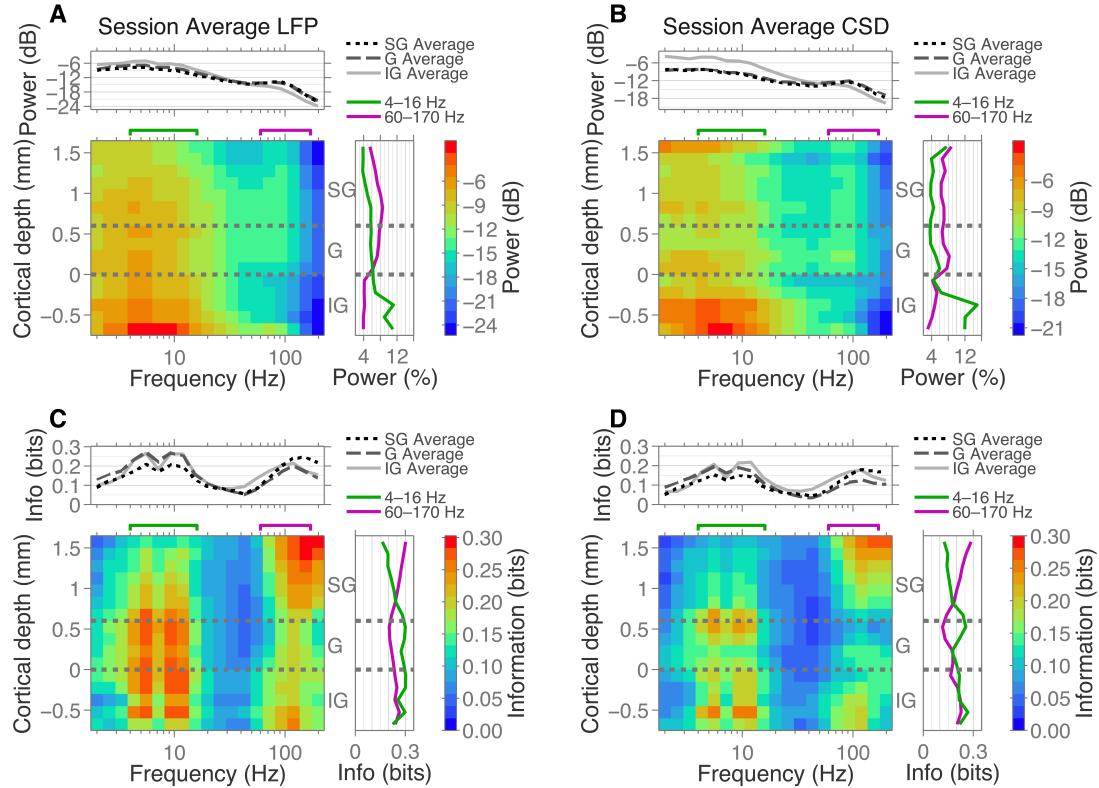


Figure 3.2: *Distribution of visual stimulus information across both cortical depth and frequency.* A: Distribution of LFP power during stimulus presentation. Plot shows the geometric mean power over 6 sessions. Above, mean power within SG, G and IG regions. Right, laminar distribution of LFP power in 4Hz to 16Hz and 6Hz to 170Hz frequency bands. B: Same as A, but distribution of CSD power instead of LFP power. C: Distribution of information about the stimulus contained in LFP power. Plot shows the mean information over 6 sessions. Above, mean information within SG, G and IG regions. Right, cortical distribution of information in the power in 4Hz to 16Hz and 6Hz to 170Hz frequency bands. D: Same as C, but for information in CSD power instead of LFP power. Note that the information (C+D) is distributed very differently from the LFP and CSD power.

information in the **CSD** has better spatial localisation than the **LFP** (Einevoll, Kayser, Logothetis, & Panzeri, 2013; Kajikawa & Schroeder, 2011), for the remainder of the paper we study only the **CSD**.

These findings suggest that there are different information channels in a single cortical column.

### 3.3.2 *Information redundancy between frequencies*

Having identified the most informative regions in depth and frequency, there are two possibilities: either these regions contain the same information about the stimulus, through transcoding of one frequency range to another across the cortex [CITE COMPUTATIONAL STUDY ON THIS], or the regions contain different information about the stimulus. We investigated how similar the information was by computing the redundancy of information contained in pairs of frequencies (see Experimental Methods). We found there are two frequency domains within which information is redundant: 4 Hz to 40 Hz and  $>40$  Hz (Fig. 3.3A). Furthermore, the information contained in neural frequencies  $<40$  Hz is different to the information contained in frequencies  $>40$  Hz, since these measured to be independent (redundancy  $\leq 0\%$ ). The same  $<40$  Hz and  $>40$  Hz division is observed for the signal correlation (Sup Fig. 3.3), and our results corroborate earlier findings (Belitski et al., 2008). Based on these and the above results, we extracted two bands (4 Hz to 16 Hz and 60 Hz to 170 Hz) that contain the most information and independently encode information about the stimulus.

### 3.3.3 *Information redundancy across depth*

Having established the independence of these bands, we investigated the redundancy between the power of oscillations at different cortical depths (Fig. 3.3B). For the 4 Hz to 16 Hz frequency range, we found there is some redundancy across the entire cortical depth, but there are two distinct cortical regions (above and below the **CSD** reversal, marked as **omm** depth) within which information is more redundant. These findings are in agreement with (Maier, Adams, Aura, & Leopold, 2010), who found a transition corresponding to the **G/IG** boundary which isolated two cortical regions with high coherence  $<100$  Hz.

Gamma oscillations (6 Hz to 170 Hz) code, with substantial redundancy across the cortical depth, with some compartmentalisation of **SG** and **IG** activities. In addition we also included the **MUA** signal, which corresponds to the local population firing rate. There is less redundancy of information across cortical depths for **MUA** than for

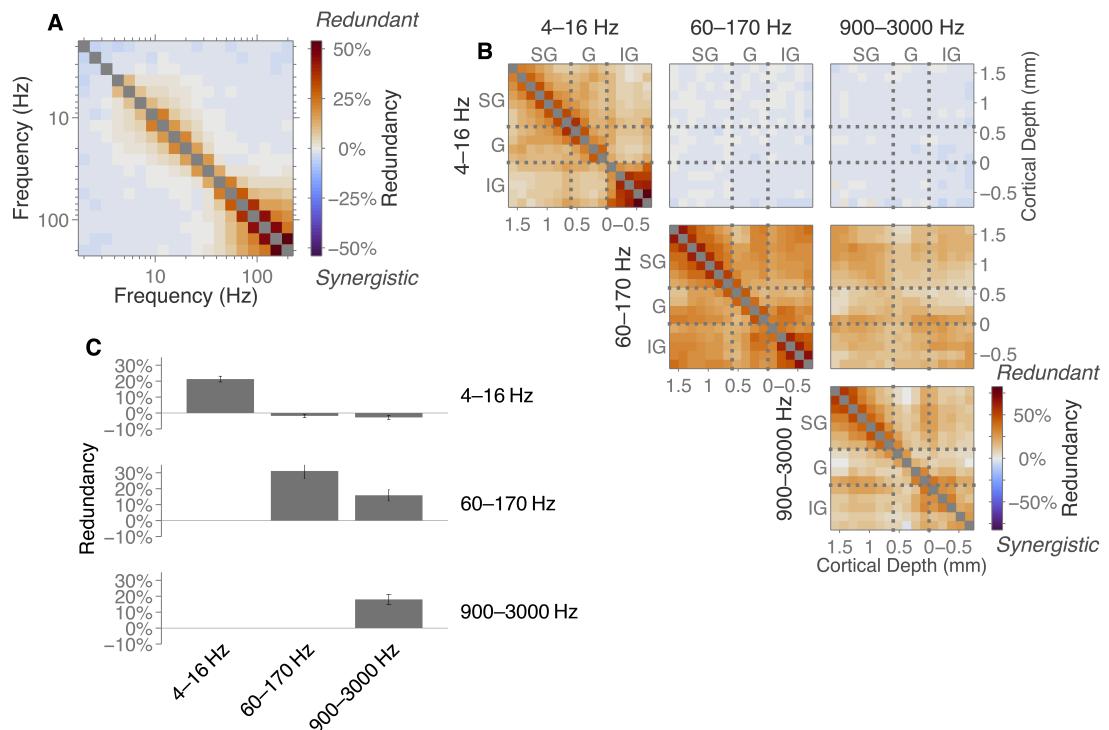


Figure 3.3: *CSD information redundancy across frequency bands and laminae*. A: Median redundancy between pairs of frequencies over the 12 recording sites, averaged over 6 sessions. B: Redundancy between pairs of recording sites of the information in three frequency bands. Mean of 6 sessions. C: Average of cross-channel redundancy shown in B. Note, that while there is substantial redundancy within bands and between the 60 Hz to 170 Hz and 900 Hz to 3000 Hz bands, there is little redundancy between the 4 Hz to 16 Hz and 6 Hz to 170 Hz band, indicating independent coding.

gamma; this observation is due to spiking activity being more localised than gamma oscillations. In agreement with previous findings (Belitski et al., 2008), we find that information contained in the gamma range and information in the MUA are redundant with each other. This is to be expected, since MUA is known to be correlated with the gamma cycle.(due to peaks/troughs in gamma relating to peaks/troughs in firing rate)

Overall redundancy is summarised in Fig. 3.3C, which shows the average across all cortical depths for each pair of frequency bands.

Importantly, we find the information in the 4 Hz to 16 Hz range is independent of the information contained in both gamma and MUA frequency ranges across all cortical depths. In particular, this means the two localised high information regions in depth-frequency space from Fig. 3.2D contain independent information to one-another. Importantly, this mean this argues against a situation where SG contains the same information as G/IG activity transcoded from low-frequency to high-gamma oscillations; at least some of the information is unique to each.

### 3.3.4 *Information about spatial frequency components of visual stimulus*

In the above, we have seen there are two frequency bands in V<sub>1</sub> which, across all the cortical depth, contain independent information to each other. Next we investigate what aspects of the visual scene these two independent components contain. Since neurons in the primary visual cortex are known to respond strongly to moving sinusoidal gratings with specific spatial frequencies, we considered how much information the frequency bands contained about changes in luminance as a function of spatial frequency. Hereto, we decomposed the series of frames in the movie into set of spatial frequency components by finding the rate of change of luminance within a given set of spatial frequency bands (see Fig. 3.4; Experimental Methods), and then computed the amount of information about this series contained in the neural activity.

We found the low frequency CSD bands (<40 Hz) contained more information about low, coarse spatial frequencies (0.1 cpd to 0.6 cpd), whereas the higher frequencies (>40 Hz) contained more information about high, fine spatial frequencies (0.6 cpd to 5.0 cpd) (Fig. 3.5B). This was not a continuous transition; instead we observe an abrupt change at 40 Hz, with lower and higher neural oscillation frequencies tuned to stimulus features with different spatial frequencies. This was true across the entire cortical depth (Fig. 3.5C–D), where the two frequency bands (4 Hz to 16 Hz and 60 Hz to 170 Hz) contained information about opposing spatial frequencies. The distribution of information across the cortical depth corresponds to that found in

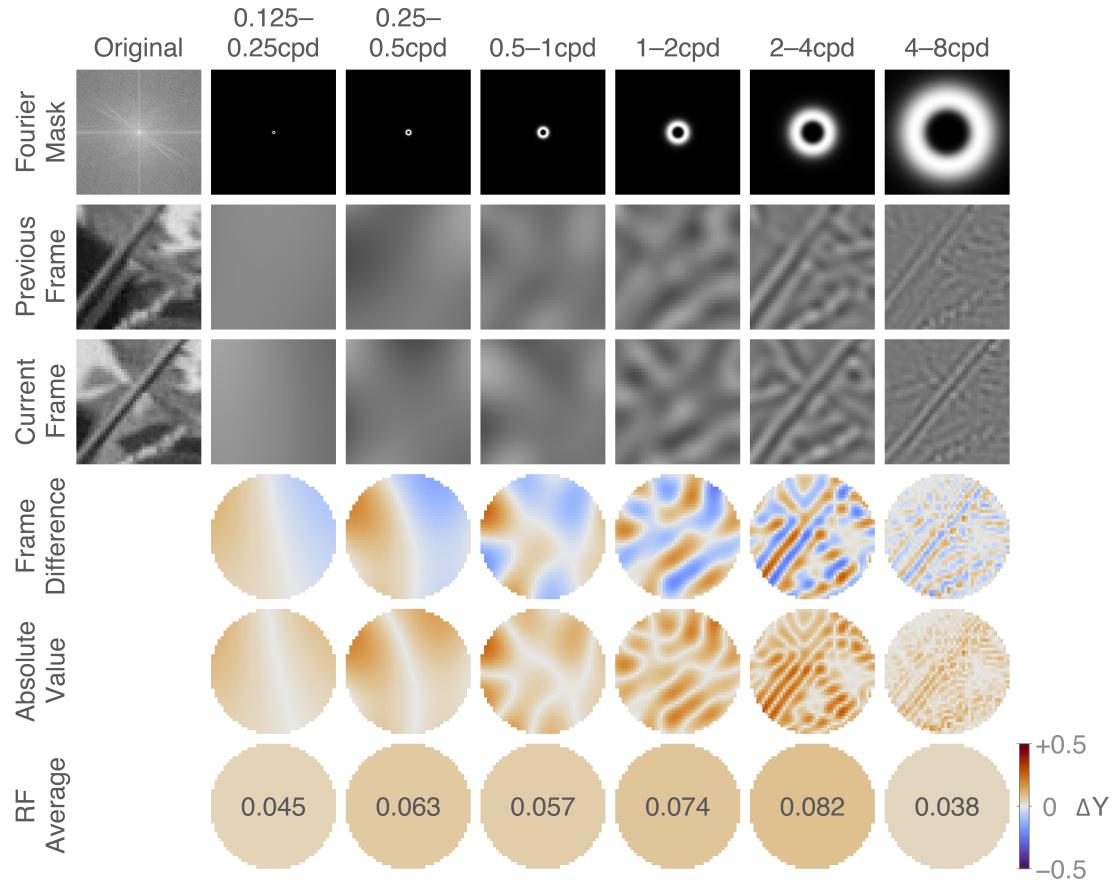


Figure 3.4: *Extraction of spatially filtered luminance components.* The luminance of the original video (left) is fast-Fourier transformed in a  $224 \times 224$  px square for each frame (top-left: FFT of “current frame”). The mask isolates bands of spatial frequencies that are one octave wide (Row 1), yielding the spatially filtered frames (Rows 2 and 3). The stimulus magnitude at each spatial frequency band was obtained by taking the luminance difference of successive frames (Row 4), taking its absolute value (Row 5), and averaging this within the receptive field (Row 6).

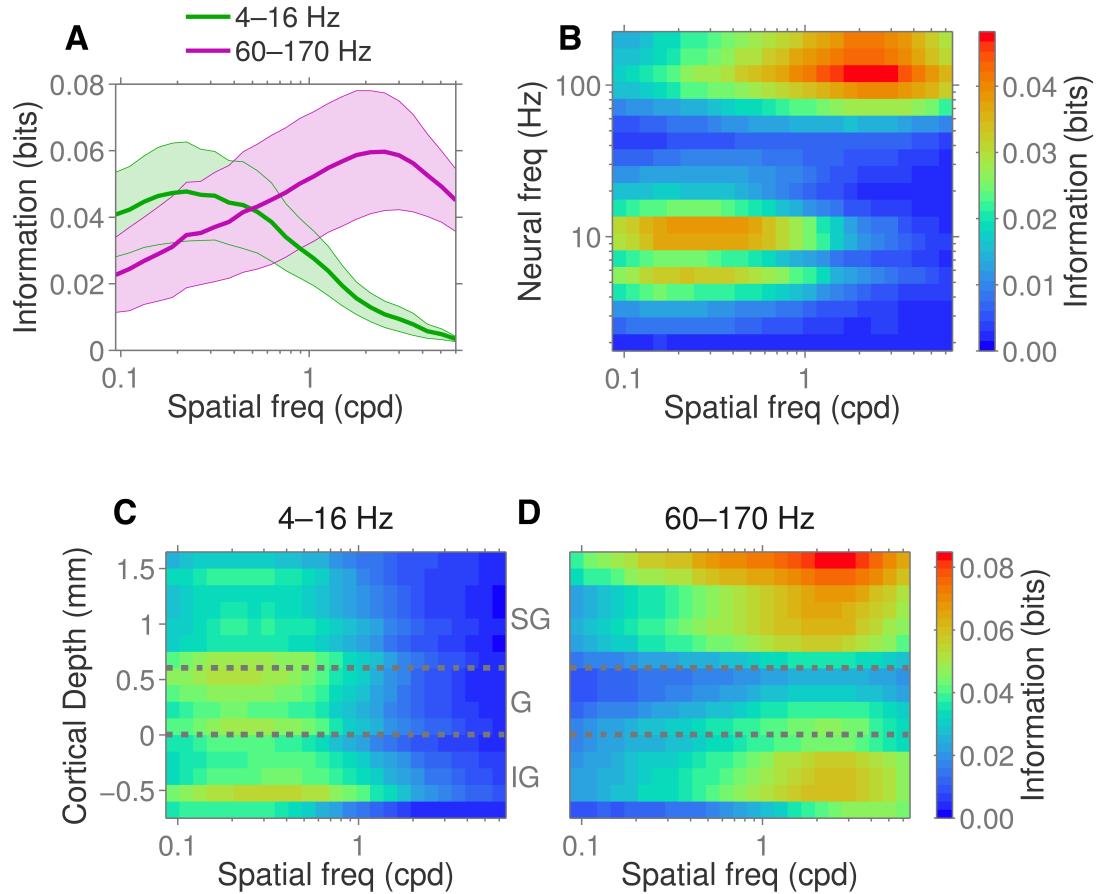


Figure 3.5: *Information about different spatial components across laminae and frequency bands.* A: Information about spatial components of the stimulus contained in low frequency CSD power (4 Hz to 16 Hz, average of information within **G** region; green) and high frequency CSD power (6 Hz to 170 Hz, average of information within **SG** region; purple). Shaded region: standard error across 6 sessions. B: Information about visual spatial components contained in a range of CSD frequencies, median over 12 recording sites. C,D: Information in low (4 Hz to 16 Hz) and high (6 Hz to 170 Hz) CSD frequency bands across cortical laminae. Plots A-D are mean of 6 sessions.

Fig. 3.2D. These results are summarised in Fig. 3.5A, which shows the average across the cortical depth. Information reaches its maxima around 0.2 cpd for the 4 Hz to 16 Hz frequency range and 2.5 cpd for the 60 Hz to 170 Hz frequency range.

In Fig. 3.6, we summarise the previous results by extracting two spatial frequency bands: coarse (<0.3 cpd, low-pass spatial filter) and fine (>1 cpd, high-pass spatial filter), example traces for which are shown above Fig. 3.6A. These spatial components have low correlation between them (Fig. 3.6B;  $r = 0.18$ ). Example CSD traces are also shown for two electrode contacts over same time period (left side). Note that peaks and troughs in the coarse luminance signal are coincident with peaks and troughs with the alpha power, and similarly for the fine luminance and gamma power, indicating the positive relationship between stimulus and cortical response. This observation is quantified by the correlation and mutual information between these components (Fig. 3.6A).

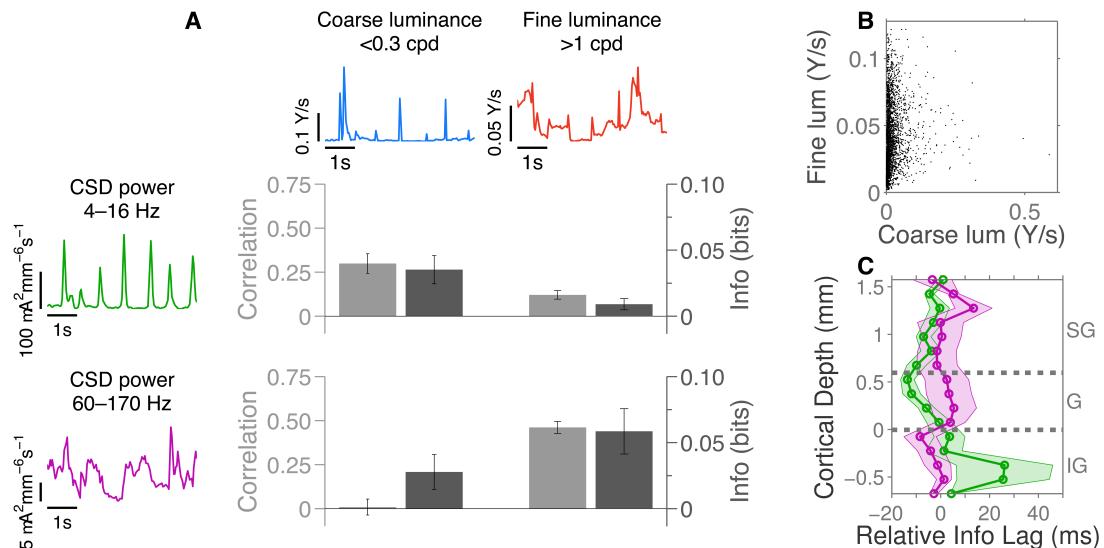


Figure 3.6: Overview of information components. A: Relationship between Coarse/Fine changes in luminance and Low/High frequency neural activity. Left: Instantaneous power in 4Hz to 16Hz band (averaged over trials and SG layers) and 6Hz to 170Hz band (averaged over trials and G layers) for an example session (Ho5nm7). Above: Coarse (<0.3 cpd) and fine (>1 cpd) rate of change in luminance over the same time period. The barchart shows, for each pair of stimulus and response, Pearson's correlation coefficient (pale grey; left-hand axis) and mutual information (dark grey; right-hand axis). B: Fine versus coarse change in luminance for each frame change in the stimulus. C: Lag between stimulus and response yielding maximal information (green: 4Hz to 16Hz and coarse luminance; purple: 6Hz to 170Hz and fine luminance).

### 3.3.5 Layer 1 6 Hz to 170 Hz amplitude is coupled to L5 4 Hz to 16 Hz phase

In previous section “Information redundancy across depth”, we showed that high and low **LFP** frequencies contain independent information to one-another. To further investigate the relationship between these two bands, we computed the cross-frequency coupling between the low frequency phase and high frequency oscillation amplitude. In agreement with previous work (Spaak, Bonnefond, Maier, Leopold, & Jensen, 2012), we found there is significant coupling between the 4 Hz to 16 Hz phase in lower-**G** and mid-**IG** with the amplitude of 6 Hz to 170 Hz oscillations in upper **SG** (Fig. 3.7). There is also localised coupling between the 4 Hz to 16 Hz phase with 6 Hz to 170 Hz amplitude in **G** and **IG**. These findings were all true of both the stimulus driven and spontaneous recordings.

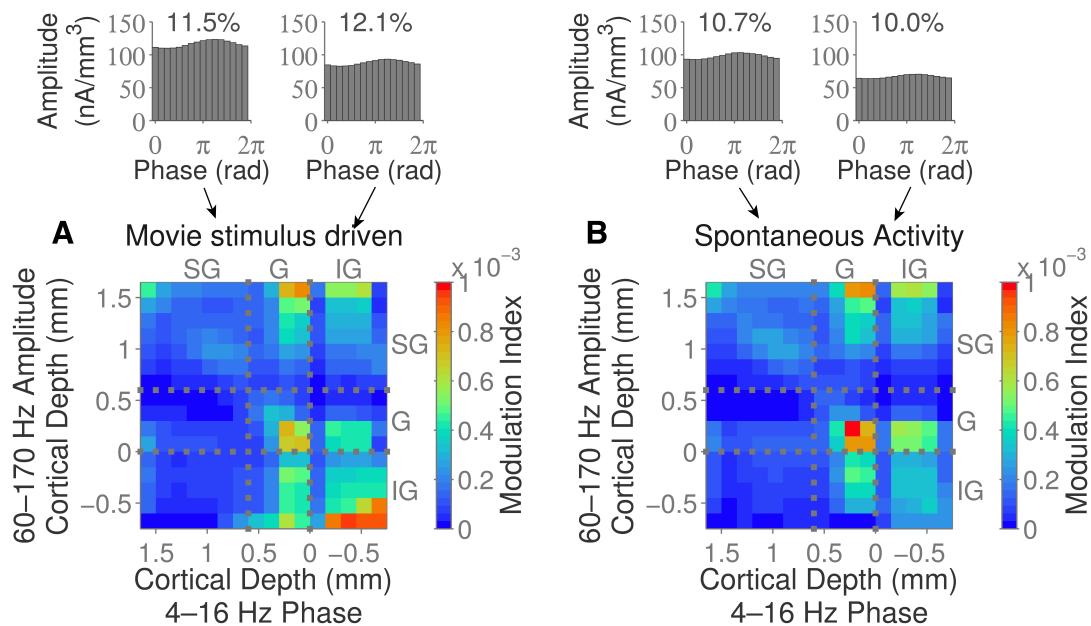


Figure 3.7: *Cross-frequency phase-amplitude coupling* Phase-amplitude modulation index between low frequency (4 Hz to 16 Hz) phase and high frequency (6 Hz to 170 Hz) amplitude (A: movie driven activity; B: spontaneous activity). Mean of 5 sessions. Above, amplitude as a function of binned phase for an example session (Ho5391). Left inset: **IG/IG** coupling; Right inset: **IG/SG** coupling.

## 3.4 DISCUSSION

In summary, we find while **LFP** power is smooth and its depth profile is close to flat (Fig. 3.2a,b) the information that the **LFP** encodes reveals much more structure. We found there are two cortical regions at which oscillations in these frequency ranges are much more informative. Namely 4 Hz to 16 Hz at upper granular and

mid-infragranular, and 6 Hz to 170 Hz at upper supragranular and mid-infragranular regions. Previous work (Belitski et al., 2008) has shown that in the macaque primary visual cortex information is coded in two frequency bands ( $<40$  Hz and  $>40$  Hz) containing independent information about natural visual scenes. Our analysis extended across the cortical depth, and we found there are two cortical regions at which oscillations in these frequency ranges are much more informative (4 Hz to 16 Hz at upper-**G** and mid-**IG**; 6 Hz to 170 Hz at upper-**SG** and mid-**IG** regions).

We also examined whether changes in luminance at different spatial frequencies induced differential changes in the cortex as a function of neural frequency and depth. Namely, high spatial frequencies are encoded in oscillations faster than 40 Hz and low spatial frequencies are encoded in oscillations slower than 40 Hz. We found that frequencies below and above 40 Hz contain information about different spatial frequencies.

There are multiple possible interpretations of these findings, of which one, many, or even none may be correct. Firstly, it is conceivable that the coding of different aspects of the stimulus into different frequency bands is a computational strategy of the cortex. Our results suggest there is multiplexing in the cortex, with low frequency and high frequency oscillations of the same population activity simultaneously encoding low and high spatial frequency components of the stimulus respectively. The idea of different frequency bands conveying different spatial frequency components of the stimulus has been proposed before from the results of an **EEG** study (Smith, Gosselin, & Schyns, 2006).

One would expect that if certain oscillation frequencies in the visual cortex contain information about specific aspects of the stimulus, this is likely to be because the brain has encoded this information into oscillations in the activity of the local population. This would only make sense if the information is utilized by the brain in order to interpret its stimuli. Consequently, our results indicate there is multiplexing in the cortex, with low frequency and high frequency oscillations of the same population activity simultaneously encoding low and high spatial frequency components of the stimulus respectively. Intuitively, information contained in the two frequency bands can be combined by downstream visual cortical regions to regain the original stimulus as necessary. The idea of different frequency bands conveying different spatial frequency components of the stimulus has been proposed before from the results of an **EEG** study (Smith, Gosselin, & Schyns, 2006).

Additionally, we can speculate about why separating the visual scene into low frequency (coarse) and high frequency (fine) components in **V<sub>1</sub>** is useful. One possibility is that low frequency oscillations are output from **V<sub>1</sub>** along the dorsal visual stream, whereas high frequency oscillations travel propagate through the ventral stream [IS

THERE EVIDENCE FOR THIS?]. Another possibility is that broad, coarse changes in the stimulus are useful for making rapid responses in the motor cortex to sudden changes, such as approaching threats.

Separation into low and high frequency domains with different properties seems to be a common property of the cortex. In motor cortex, activity at  $<13\text{ Hz}$  and  $>60\text{ Hz}$  relates to behaviour but there is a separating band  $\approx 30\text{ Hz}$  which does not (Rickert et al., 2005). In the hippocampus, there is a gating effect between  $30\text{ Hz}$  and  $40\text{ Hz}$ , with lower but not higher frequencies able to propagate to the cortex (Moreno, Morris, & Canals, 2015). [also some unpublished research by Julian Hoffman into independent oscillations in the barrel cortex]. This suggests this encoding scheme is common across the cortex. Some studies have suggested that the coupling of oscillations between two cortical regions facilitates the transmission between them [CITE SOME EXPERIMENTAL & COMPUTATIONAL WORKS].

A separation of visual stimuli into coarse and fine channels is known to occur before the stimuli arrive in the cortex. The outputs from different types of retinal ganglion cells (**RGCs**) travel to the cortex through different regions of a lateral geniculiate nucleus (**LGN**). The M-pathway arises from **RGCs** with large, achromatic receptive fields, and projects mainly onto layer  $4C\alpha$  in **V<sub>1</sub>**. The P-pathway originates with **RGCs** with smaller, chromatic **RFs** providing higher spatial resolution but lower temporal resolution; this pathway projects onto layer  $4C\beta$  (E M Callaway, 1998). It is possible that the two frequency channels in **V<sub>1</sub>** relate to the two pathways providing its inputs.. [Cite Nathaniel J Killian from AREADNE on information in low-frequencies of **LGN**. Nothing available to actually cite?]

Since **L<sub>4</sub>** is generally regarded as the primary layer of **V<sub>1</sub>** which receives afferent inputs from a **LGN**, some readers might wonder how information in the gamma band has “arisen” in **SG** layers without passing through **G**. However, our results do not necessitate this. Fine-resolution information about the visual stimulus can arrive from the **LGN** into **L<sub>4</sub>** of **V<sub>1</sub>**, with the information encoded into which neurons the afferent connections target. This information is not detectable from the population level activity.

As many readers will be aware, it has long been known that neurons in the primary visual cortex have a response curve tuned to a preferred spatial frequency. Work demonstrating the spatial frequency preference of single neurons typically involves the presentation of moving sinusoidal grating with a particular spatial frequency. [NOT SURE WHERE I WAS GOING WITH THIS...]

The higher frequency band contains information about higher spatial frequencies changes in the stimulus. This corresponds to the detection of edges and texture, which are properties that single neurons in **V<sub>1</sub>** are known to be selective for.

We observed that each frequency has a similar amount of power across the cortical depth, but oscillations at these frequency ranges contain much more information at particular cortical depths. This is curious as it indicates that, for any given frequency band, oscillations are present in all cortical depths, but most of the oscillations exhibited are not stimulus encoding. This seems wasteful. [important point]

We observed qualitatively that information-carrying events in  $L_4$  were large, temporary deflections with a long duration (low frequency), whereas  $L_5/6$  contained sustained oscillations (see Fig. 3.1A for examples). The deflections in  $L_4$  were usually coincident with scene cuts or rapid changes in the stimulus. This could be interpreted as an error signal, since sudden, large changes in the stimulus would result in any predictive model of the stimulus making large errors. However, a more simple interpretation is these deflections correspond to changes in the afferent input to  $V_1$  from  $LGN$ . In support of this, we note that the spatial scale of the information in the low frequency band (0.25 cpd) approximately corresponds to the size of receptive fields for regions of the  $V_1$  corresponding to the parafovea ( $2^\circ$ ).

The sustained oscillations in  $L_5/6$  also contain information about coarse changes in the stimuli. These cortical layers are known to have connections to the motor cortex, feedback to a  $LGN$  and receiving feedback from higher cortical regions.

Recent work has indicated that alpha and gamma bands are important for feedback and feedforward activity respectively (van Keroerle et al., 2014). This study (van Keroerle et al., 2014) found that gamma waves are initiated at  $L_4$  and propagate outwards to the top of  $SG$  and bottom of  $IG$ , with alpha waves propagating in the opposite direction. Our study finds the most information in gamma bands at the very top (and very bottom) of the cortex, and the most information in alpha bands at the top of  $L_4$  (and layer 6 of  $V_1$  ( $L_6$ )). Reconciling these results together, we find that there is most information in the power of the alpha and gamma oscillations at the cortical depths where they terminate, and the least where they originate. This suggests that the oscillations are generated at one cortical depth without much stimulus dependency, but as the oscillations propagate up and down the cortex they are either amplified or suppressed in a stimulus dependent manner.

In agreement with previous work (Spaak et al., 2012), we found there was cross-frequency coupling between the stimulus-encoding power of gamma oscillations in layer 1 of  $V_1$  ( $L_1$ ) and the phase of alpha oscillations in lower  $L_4$ . Anatomically, we believe this is related to the pyramidal cell bodies in layer 5A of  $V_1$  ( $L_{5A}$ ), which have apical dendritic tufts in  $L_1$  (Hill, Jia, Sakmann, & Konnerth, 2013; Zhu & Zhu, 2004). This cross-frequency coupling could be one mechanism through which the  $L_1$  gamma wave containing high levels of information about the stimulus is converted into an alpha oscillation for feedback into the hierarchically lower cortical region. Neurons

in layer 5 of V<sub>1</sub> (L<sub>5</sub>) are known to be related to long-range cortical output (Hill et al., 2013).

## GENERAL DISCUSSION

---

The end

## BIBLIOGRAPHY

---

- Adini, Y., Sagi, D., and Tsodyks, M. (2002). Context-enabled learning in the human visual system. *Nature*, 415(6873):790–3.
- Ahissar, M. and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–64.
- Arabzadeh, E., Panzeri, S., and Diamond, M. E. (2006). Deciphering the spike train of a sensory neuron: counts and temporal patterns in the rat whisker pathway. *Journal of Neuroscience*, 26(36):9216–26.
- Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews. Neuroscience*, 7(5):358–66.
- Ball, K. and Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research*, 27(6):953–965.
- Berardi, N., Bisti, S., and Maffei, L. (1987). The transfer of visual information across the corpus callosum: spatial and temporal properties in the cat. *The Journal of physiology*, 384:619–32.
- Brasselet, R., Panzeri, S., Logothetis, N. K., and Kayser, C. (2012). Neurons with stereotyped and rapid responses provide a reference frame for relative temporal coding in primate auditory cortex. *The Journal of Neuroscience*, 32(9):2998–3008.
- Chen, X., Sanayei, M., and Thiele, A. (2013). Perceptual learning of contrast discrimination in macaca mulatta. *Journal of Vision*, 12(12):1–15.
- Chen, X., Sanayei, M., and Thiele, A. (2014). Stimulus roving and flankers affect perceptual learning of contrast discrimination in Macaca mulatta. *PLoS ONE*, 9(10):13–15.
- Cohen, M. R. and Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–600.
- Dinse, H. R., Ragert, P., Pleger, B., Schwenkreis, P., and Tegenthoff, M. (2003). Pharmacological modulation of perceptual learning and associated cortical reorganization. *Science*, 301(5629):91–4.
- Ecker, A. S., Berens, P., Tolias, A. S., and Bethge, M. (2011). The effect of noise correlations in populations of diversely tuned neurons. *Journal of Neuroscience*, 31(40):14272–83.
- Fahle, M. (2005). Perceptual learning: specificity versus generalization. *Current opinion in neurobiology*, 15(2):154–60.
- Fiorentini, A. and Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency.
- Fiorentini, A. and Berardi, N. (1981). Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research*, 21(7):1149–1158.
- Ghose, G. M., Yang, T., and Maunsell, J. H. R. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *Journal of neurophysiology*, 87(4):1867–88.

- Gilbert, C. (1994). Early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, 91(February):1195–1197.
- Gilbert, C. D., Sigman, M., and Crist, R. E. (2001). The Neural Basis of Perceptual Learning. *Neuron*, 31:681–697.
- Gu, Y., Liu, S., Fetsch, C. R., Yang, Y., Fok, S., Sunkara, A., DeAngelis, G. C., and Angelaki, D. E. (2011). Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron*, 71(4):750–61.
- Hochstein, S. and Ahissar, M. (2002). View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*, 36(3):791–804.
- Ince, R. A. A., Mazzoni, A., Bartels, A., Logothetis, N. K., and Panzeri, S. (2011). A novel test to determine the significance of neural selectivity to single and multiple potentially correlated stimulus features. *Journal of neuroscience methods*.
- Karni, A. and Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, 88(11):4966–4970.
- Magri, C., Whittingstall, K., Singh, V., Logothetis, N. K., and Panzeri, S. (2009). A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC neuroscience*, 10(81).
- Miller, G. A. (1955). Note on the bias of information estimates.
- Montemurro, M. a., Senatore, R., and Panzeri, S. (2007). Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural computation*, 19(11):2913–57.
- Müller, J. R., Metha, A. B., Krauskopf, J., and Lennie, P. (2001). Information conveyed by onset transients in responses of striate cortical neurons. *The Journal of Neuroscience*, 21(17):6978–90.
- Nemenman, I., Bialek, W., and de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: progress on the sampling problem. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69(5 Pt 2):056111.
- Optican, L. M. and Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *Journal of Neurophysiology*, 57(1):162–178.
- Oram, M. W., Földiák, P., Perrett, D. I., and Sengpiel, F. (1998). The ‘Ideal Homunculus’: decoding neural population signals. *Trends in neurosciences*, 21(6):259–65.
- Panzeri, S., Senatore, R., Montemurro, M. A., and Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, 98(3):1064–72.
- Panzeri, S. and Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107.
- Pleger, B., Foerster, A. F., Ragert, P., Dinse, H. R., Schwenkreis, P., Malin, J. P., Nicolas, V., and Tegenthoff, M. (2003). Functional imaging of perceptual learning in human primary and secondary somatosensory cortex. *Neuron*, 40(3):643–53.
- Poggio, T., Fahle, M., and Edelman, S. (1991). Fast Perceptual Learning in Visual Hyperacuity. *Information Systems*.

- Poggio, T., Fahle, M., and Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science (New York, N.Y.)*, 256(5059):1018–21.
- Polley, D. B., Steinberg, E. E., and Merzenich, M. M. (2006). Perceptual learning directs auditory cortical map reorganization through top-down influences. *The Journal of Neuroscience*, 26(18):4970–82.
- Quiroga, R. Q. and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature reviews. Neuroscience*, 10(3):173–85.
- Raiguel, S., Vogels, R., Mysore, S. G., and Orban, G. a. (2006). Learning to see the difference specifically alters the most informative V4 neurons. *The Journal of Neuroscience*, 26(24):6589–602.
- Reich, D., Mechler, F., and Victor, J. (2001). Temporal coding of contrast in primary visual cortex: when, what, and why. *Journal of Neurophysiology*, 85:1039–1050.
- Rolls, E. T. and Treves, A. (2011). The neuronal encoding of information in the brain. *Progress in neurobiology*, 95(3):448–90.
- Schoups, A., Vogels, R., and Qian, N. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412(August):549–553.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(July 1948):379–423.
- Strong, S., Koberle, R., de Ruyter van Steveninck, R., and Bialek, W. (1998). Entropy and Information in Neural Spike Trains. *Physical Review Letters*, 80(1):197–200.
- Tovée, M. J., Rolls, E. T., Treves, a., and Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *Journal of neurophysiology*, 70(2):640–54.
- Treves, A. and Panzeri, S. (1995). The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Computation*, 7(2):399–407.
- Yang, T. and Maunsell, J. H. R. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *The Journal of Neuroscience*, 24(7):1617–26.
- Yu, C., Klein, S., and Levi, D. (2004). Perceptual learning in contrast discrimination and the (minimal) role of context. *Journal of Vision*, 4(3):169–182.