The University Of Sheffield.

How can plasticity of lateral interactions affect cortical representation?

Author:

Giacomo SPIGLER

Supervisor:

Dr. Stuart WILSON

Dr. Renee TIMMERS

Prof. Tony PRESCOTT

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Psychology

February 2018

The University of Sheffield

Abstract

Faculty of Science

Department of Psychology

Doctor of Philosophy

How can plasticity of lateral interactions affect cortical representation?

by Giacomo SPIGLER

Lateral connectivity within cortical areas is pervasive in the mammalian neocortex. The lateral interaction between cortical minicolumns mediated by such connections has been shown to play a critical role in cortical function and cognition, and has been used to explain the emergence of large-scale patterns such as cortical maps. Further evidence suggests that aspects of cortical representation of learnt sensory stimuli may be encoded in the synaptic strengths of lateral connections.

This thesis builds upon a program of existing computational neuroscience research, which has identified plasticity in lateral interactions as the key component of cortical functional organisation, to ask whether a neurobiologically plausible computational model of cortical self-organisation can be used to investigate how synaptic plasticity and adaptation in lateral cortical interactions modifies the structure of pre-existing cortical representations and how it affects their decoding.

The *inhibitory sharpening* theory is proposed, based on computer simulations, that shows how repetition suppression is compatible with an increase in the strength of the inhibitory interactions between cortical units co-active during the presentation of the same adapter stimulus due to Hebbian learning. A key prediction of the theory is then derived, that stimuli that produce overlapping patterns of cortical activity, that is that activate a common sub-set of neurons, may produce mutual interference that should be reflected both in changes to the neural signal and in higher level cognition.

The predictions of the theory are tested with two approaches, a neuroimaging experiment to measure the magnitude of repetition suppression in a protocol compatible with that used in the simulations, and a behavioural experiment.

Acknowledgements

I would like to express my special thanks to my supervisor Dr. Stuart Wilson and to my second and third supervisors Dr. Renee Timmers and Prof. Tony Prescott. I also thank the University of Sheffield for the environment, the facilities and the life experiences that it provided.

I would also like to thank Prof. Iain Wilkinson and all the staff of the University of Sheffield MRI Unit at the Royal Hallamshire Hospital, without whose support it would have not been possible to perform the neuroimaging experiment presented in this thesis.

A special thanks is for my family. Words cannot express how grateful I am to my parents for all of the sacrifices that they have made and for their continued support.

Finally, I would like to thank all of my friends who supported me during the ups and downs of a PhD, and helped me to strive towards my goal. For this, I specially thank Valerio and Daniele.

Sheffield,
February 2018

Contents

Abstract **iv**

Acknowledgements **v**

1 Introduction **1**

1.1 Lateral cortical interactions 2

1.2 Organisation of the thesis 5

2 A theory of cortical self-organisation **9**

2.1 Cortical maps 9

2.2 Models of cortical self-organisation 12

2.3 The L-model 15

2.4 Distributed cortical representations in the L-model 18

2.5 How does plasticity in lateral interactions influence cortical dynamics? 21

2.6 Plasticity in lateral interactions affects cortical representation . . . 25

3 Inhibitory sharpening: simulations **29**

3.1 Introduction 29

3.2 Methods 32

3.3 Results 33

3.3.1 Plastic intracortical connectivity is sufficient to explain repetition suppression 33

3.3.2 Interfering representations disrupt repetition suppression 34

3.3.3 Role of plasticity in the afferent and inhibitory interactions 38

3.4 Discussion 42

4 Perceptually similar stimuli disrupt recognition performance 55

4.1 Introduction 55

4.2 Methods 56

4.2.1 Participants 56

4.2.2 Materials 57

4.2.3 Experimental Design 58

4.2.4 Experimental Procedure 60

4.3 Results 61

4.3.1 Experiment 61

4.4 Discussion 64

5 Neuroimaging investigation of the inhibitory sharpening theory 67

5.1 Introduction 67

5.2 Materials and Methods 70

5.2.1 Participants 70

5.2.2 Stimuli 71

5.2.3 Experimental Design 71

5.2.4 Experimental Procedure 73

5.2.5 Scanning Parameters 75

5.2.6 Region of Interest Analysis 77

5.2.7 Data Analysis 77

5.3 Results 78

5.4 Discussion 83

6 Discussion and conclusion 87

6.1 General discussion 87

6.2 Parts-based population coding 90

6.3 Bridging mechanistic models of cortical dynamics with high-level cognition 92

6.4 Effect of cortical overlap in cognition 93
6.5 How can plasticity of lateral interactions affect cortical representation? 95
6.6 Future Work 95

A Parameters of the simulations **97**

B Supplementary Results for the neuroimaging investigation **99**

C Preliminary experiments with Non-Negative Least Squares (NNLS) **103**

Bibliography **107**

List of Figures

2.1 Cortical maps in mammals 10
2.2 Von der Malsburg 1973 model 14
2.3 Schematics of the L-model 17
2.4 Distributed vs localist representations 20
2.5 Distributed representation of objects 22
2.6 Specific lateral connectivity in the primary visual cortex 24
2.7 Tilt Aftereffect and McCollough Effect 27

3.1 Simulations showing repetition suppression 35
3.2 Effect of afferent and inhibitory plasticity 36
3.3 Influence of intervening stimuli on the degree of cortical overlap 39
3.4 Dynamics for the non-overlap versus overlap simulations 40
3.5 Non-overlap versus overlap 41
3.6 Overlap vs Non-Overlap in Inhibitory-Only vs Afferent-Only . . 43
3.7 Changes in lateral connectivity underlying repetition suppression 49
3.8 Activation and effective inhibition in the three phases of the protocol 50
3.9 Timecourse and tuning curves of sample units 51
3.10 Longer term dynamics of repetition suppression 54

4.1 Stimuli and protocol of the behavioural experiment 59
4.2 Results of the experiment 62

5.1 Predicted magnitude of RS in the different conditions 74

5.2 fMRI experimental protocol 76
5.3 Repetition suppression: single-voxel analysis and ROI PSTHs . . 79
5.4 Repetition suppression in the different conditions 82

B.1 Single voxel analysis: face localizer 99
B.2 Beta values for each first face 100
B.3 Beta values for each regressor and ROI 101

C.1 Results of the NNLS-based Kohonen SOM simulation 106

List of Tables

5.1 **Masks used for the region of interest (ROI) analysis.** 77

Chapter 1

Introduction

One of the fundamental questions in neuroscience is understanding how the brain builds and maintains representations of the world. A useful approach to the problem is to study how the representation of sensory stimuli is encoded in the patterns of neural activity in the nervous system and in its connectivity (Pennartz, 2015).

The neocortex in particular is a good target for the study of the high-level representation of stimuli as it correlates well with the cognitive capabilities of the different mammalian species, and because the increase in human cognitive capabilities seems to have resulted from a disproportionate increase in the size of the cortical sheet during evolution.

Neuroscience has generated a large wealth of data (imaging, electrophysiological, neuroanatomical) that describes cortical dynamics at different levels of abstraction, and at different spatial and temporal scales (from perceptual to developmental timescales). One way to understand how these data fit together is through computational modelling. This approach forces us to make explicit all our assumptions about how the system works, and to make sensible assumptions to fill the gaps in our understanding. Simulation experiments then allow to investigate questions at different levels of abstractions, and at different spatial and temporal scales.

For example, computational modelling of cortical dynamics has been particularly successful at the level of cortical maps (Wilson and Bednar, 2015; Bednar

and Wilson, 2015), leading to a mature theory of cortical development that can be neatly summarised in a handful of equations, the L-model. As we will see in the review presented in Chapter 2, these models capture the key biological constraints on cortical development on timescales relevant for postnatal development (days, weeks, months).

However, understanding how these theories, which have been calibrated to developmental data, can be used on timescales relevant for perception and cognition has only been addressed in a handful of recent studies. The results of these studies have been successful at describing low-level perceptual processing like visual after-effects (e.g. applied to the tilt aftereffect Bednar and Miikkulainen, 2000) and illusions (e.g., applied to the McCollough effect Spigler, 2014).

To work towards a complete theory of how humans form and maintain representations of the world, we need to use these biologically grounded models of cortical development, and derive predictions that can be tested at the level of perception and cognition, and finally test those predictions directly. This is the principal aim of of this thesis.

The critical feature of the models of cortical development used in this thesis is plastic lateral interactions, that is interactions between units within the same model cortical area mediated by connections that can be modified via synaptic plasticity. The role of lateral interactions in the context of this thesis will be outlined in the next section, before presenting a general overview of the thesis.

1.1 Lateral cortical interactions

Significant progress in modelling cortical dynamics has been achieved by studying the role of recurrent connectivity within individual cortical areas and how the strength of those connections changes over time due to cortical plasticity and

learning (Sirosh, 1996; Miikkulainen et al., 2006; Stevens et al., 2013; Wilson et al., 2010; Spigler, 2014; Fischer, 2014).

Lateral connections between neurons in the same cortical areas are pervasive in the mammalian neocortex. It is also known that such lateral connections are predominantly excitatory in the long-range, and inhibitory in the short-range, and that such patterns of connectivity are organised in local patches (e.g., primary visual cortex (V1) in cats and primates (Fisken, Garey, and Powell, 1975; Gilbert and Wiesel, 1989; Schwark and Jones, 1989; Bosking et al., 1997)). Long-range inhibition may however be implemented via long-range excitation of local inhibitory neurons, as is thought to be the case in e.g., primary visual cortex (V1) for high-contrast visual inputs (Hirsch and Gilbert, 1991; Weliky et al., 1995; Ren et al., 2007; Martin, 2002; Somers et al., 1998; Silberberg and Markram, 2007), and primary somatosensory cortex (S1) for strong tactile inputs (Helmstaedter, Sakmann, and Feldmeyer, 2009; Moore, Nelson, and Sur, 1999). Indeed, in many cases the actual effect of lateral interactions observed between cortical neurons is thought to be long-range net-inhibitory and short-range net-excitatory (Wilson et al., 2010; Stevens et al., 2013).

Lateral interactions in the mammalian neocortex have been shown to play a critical role in cortical processing. For example, they are thought to be critical for neural information processing operations such as feature learning and decorrelation (Barlow and Foldiak, 1989; Dong, 1996), normalization and sharpening of activity (Somers et al., 1996; Stemmler, Usher, and Niebur, 1995; Edelman, 1996; Sabatini, 1996), associative encoding of features (Dong, 1996), illusory contours and perceptual grouping (Choe, 2001) (for a complete overview, please refer to Miikkulainen et al., 2006). Lateral interactions have been also used to explain perceptual phenomena like the Hermann grid illusion (Hermann, 1870). In particular, learning decorrelated features, associative encoding of features and perceptual grouping seem to require synaptic plasticity in the lateral interactions. A computational study further evidenced the potential role of lateral plasticity in

encoding category-specific information, aiding in the perceptual segmentation and grouping of objects in a crowded scene (Evans and Stringer, 2015). Moreover, previous modeling studies have shown that plasticity and adaptation in lateral inhibitory interactions can explain changes in perception like in visual illusions such as the tilt aftereffect (Bednar and Miikkulainen, 2000) and the McCollough effect (Spigler, 2014). For example, in the tilt-aftereffect the strength of the lateral inhibitory interactions between neurons selective to the orientation of an adapter grating increase due to their co-activation (Hebbian plasticity). The increase in the strength of the inhibitory interactions then results in a decrease in the activation of those neurons on subsequent presentations of gratings at similar orientations and a change in the pattern of activation that would otherwise characterise them, finally producing a population-decoded perception that is shifted away from the orientation of the target grating. The synaptic strengths of lateral interactions thus seem to have an effect on the cortical representation of sensory stimuli.

Lateral interactions between cortical neurons or local groups of neurons (e.g., cortical units such as cortical minicolumns) have been further used to explain the development of broad patterns of neuronal feature selectivity observed in a variety of cortical areas. Specifically, they have been shown to play a critical role in models based on dynamics of self-organisation, that are capable of spontaneously producing ordered patterns from an initially disordered system. A variety of models in this class have been successfully used to explain and predict experimental data (Malsburg, 1973; Miikkulainen et al., 2006; Stevens et al., 2013).

In this thesis I build upon a program of existing computational neuroscience research, which has identified plasticity in lateral interactions as the key component of cortical functional organisation, to ask whether a neurobiologically plausible computational model of cortical self-organisation can be used to investigate how synaptic plasticity and adaptation in lateral cortical interactions

modifies the structure of pre-existing cortical representations and how it affects their decoding. The thesis is composed as a mixture of modeling and experimental work. The key component of the model used in this thesis is the assumption that lateral interactions between cortical neurons are net inhibitory at long range, and that those interactions adapt by Hebbian plasticity.

As in the previous works, this thesis will use computational modeling as a way to formalise a theory of cortical function and to generate new predictions to guide future experiments. The experiments presented to test the predictions of the theory are then further used to bridge between low-level mechanisms of cortical self-organisation and high-level computation such as cognition and perception.

In particular, this thesis will explore how changes in the cortical representation of sensory stimuli due to plasticity and adaptation in lateral inhibitory connections relates to both short-term dynamics, on the timescale of perceptual illusions, and long-term effects, on the timescale of learning and memory.

Finally, this thesis will explore how the encoding of cortical representations in the synaptic weights of lateral interactions may lead to interference between distributed representations that share a large subset of neurons in the same cortical areas, and in particular how plastic and adaptive changes due to the presentation of one stimulus may affect the perception of subsequent stimuli.

1.2 Organisation of the thesis

This thesis is organised as a sequence of theoretical and experimental works building on top of each other. We start by reviewing a body of literature which is beginning to converge on a model of cortical self-organisation that produces patterns of distributed activity representing sensory stimuli. This model is then extended to investigate the phenomenon of repetition suppression, resulting in

the 'inhibitory sharpening' theory, and in novel, testable predictions. The predictions are finally tested by means of a neuroimaging study and a behavioural experiment.

- Chapter 2 presents a literature review, which identifies input-driven self-organisation as a theory of how functional organisation emerges in the developing neocortex. This theory is presented formally, with its assumptions stated explicitly in terms of a small number of equations that enable aspects of cortical development to be recreated in computer simulations. The chapter is structured around a mixture of theoretical and experimental developments that converge on a specific self-organising map algorithm, the L-model, which explicitly models the contribution of plastic recurrent interactions to the development of topographic maps in primate primary visual cortex. It is further suggested that this low-level model of cortical dynamics may be used to explain higher-level cognitive functions such as perception and memory. This algorithm provides a starting point for the development of the theoretical work to follow in chapter 3.

- Chapter 3 investigates whether the L-model can account for the phenomenon of repetition suppression, presenting the results in the context of the novel theory of "inhibitory sharpening". Computer simulations are then used to produce novel, testable predictions of the theory based on stimuli that produce overlapping cortical representations, specifically that such stimuli may be designed to interfere with one another in carefully designed experimental protocols. The work presented in this chapter has been published in PLoS ONE (Spigler and Wilson, 2017).

- Chapter 4 explores whether the interference between stimuli that produce overlapping cortical representations predicted by the inhibitory sharpening theory can affect perception and behavior. The results of the experiments presented are used to provide preliminary indirect support to the

theory, without relying on an explicit measure of repetition suppression, and are thus complementary to the neuroimaging investigation of Chapter 5.

- Chapter 5 aims at testing the predictions of the inhibitory sharpening theory by measuring the magnitude of repetition suppression in the relevant cortical areas with the use of functional neuroimaging (fMRI), and to investigate the predicted interference between stimuli that produce overlapping cortical representations using an appropriate experimental protocol. While the results are not conclusive, a trend in the data is observed in agreement with the predictions of the inhibitory sharpening theory, suggesting that further experiments with a larger sample size and similar experimental protocols should manage to draw stronger conclusions.

- Chapter 6 finally presents a summary of the dissertation and a discussion of general points of interest related to the work that was presented.

Chapter 2

A theory of cortical self-organisation

This chapter will first review the concepts of cortical maps and cortical self-organisation. It will then introduce the L-model as a theory of cortical self-organisation and it will set the basis to study the distributed representations produced as its patterns of activity and how they are affected by plasticity in the lateral interactions of the model.

2.1 Cortical maps

A large number of cortical areas, especially primary sensory areas, have been found to be characterised by smooth spatial patterns of tuning properties across the cortical sheet, termed 'cortical maps' (Wilson and Bednar, 2015; Bednar and Wilson, 2015). Figure 2.1 shows a small collection of example cortical maps in cats, primates and rodents. The maps are visualized by labelling the cortical minicolumns in a cortical area with their preferred stimulus, along a specific feature dimension. Cortical maps can be topographic, mapping the structure of sensory surfaces like the retina or skin onto the bidimensional cortical surface, or topological, mapping a complex feature manifold.

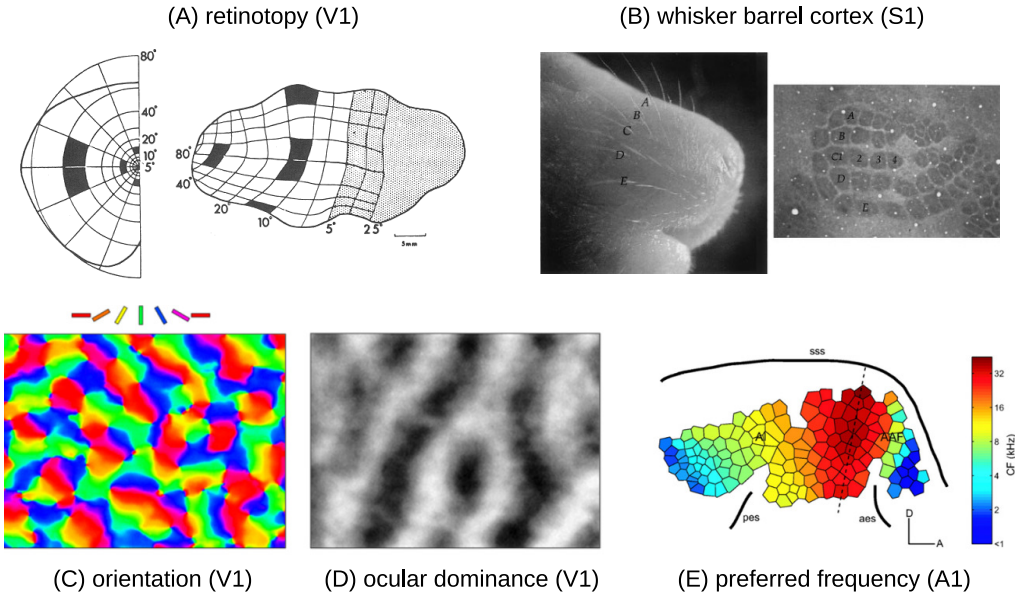


FIGURE 2.1: **Cortical maps in mammals. (A)** retinotopic organisation of the primary visual cortex (V1) of macaque monkeys (adapted from (Connolly and Van Essen, 1984)); **(B)** whisker barrel cortex (S1) in rodents (from (Wilson et al., 2000)); **C** orientation selectivity and **(D)** ocular dominance in the monkey striate cortex (adapted from (Blasdel, 1992; Miikkulainen et al., 2006)), and **E** selectivity to tones of different frequencies in the cat auditory cortex (from (Imaizumi et al., 2004)). Cortical maps are visualized by labelling the columns in a cortical area with their preferred stimulus, along a specific feature dimensions.

Cortical maps have been measured in a variety of cortical areas. The organisation of the somatosensory cortex is known to be somatotopic, with neighboring columns mapping nearby areas on the skin and overall representing the whole body as "homunculus" (Penfield and Boldrey, 1937; Woolsey et al., 1951). The primary somatosensory cortex of rodents is further characterised by an ordered map of the individual whiskers of the animal, the whisker barrel cortex (Wilson et al., 2000). Visual areas are perhaps the most intensively studied examples of cortical maps. Early visual areas are characterised by a retinotopic organisation that maps points close together on the retina to selective activation in neighboring cortical columns (Tootell et al., 1982; Connolly and Van Essen, 1984; Blasdel, Salama, et al., 1986). The primary visual cortex represents all the possible orientations of lines at every position in the retina in a locally smooth arrangement (Hubel and Wiesel, 1974; Bosking et al., 1997; Blasdel, 1992) interrupted by discontinuous points, pinwheels, around which the full range of orientations is represented, due to topological constrains (Schwartz and Rojer, 1994). The primary visual cortex also retains a locally smooth map of ocular dominance (Blasdel, 1992), colour tuning (Dow, 2002), spatial frequency (Nauhaus et al., 2012) and motion direction (Ohki et al., 2005).

A smooth organisation of the tuning properties of neurons is also present in the primary auditory cortex as a tonotopic map (Imaizumi et al., 2004) and in the inferotemporal cortex of primates as a smooth variation in selectivity to complex objects and features (Tsunoda et al., 2001; Tanaka, 2003). The functional organisation of the primate motor and pre-motor areas have also been described in terms of a locally continuous mapping of the behavioural repertoire of the animal and of target muscle groups controlled by the individual neurons (Graziano and Aflalo, 2007; Graziano, 2008).

While cortical maps have been measured in a large variety of different species

and cortical regions, it is still debated whether map patterns reflect particular *functional* organisation, or whether they are just a by-product of other phenomena, like achieving minimal length of connectivity in a cortical area (e.g., Koulakov and Chklovskii, 2001).

2.2 Models of cortical self-organisation

Self-organising systems are characterised by the spontaneous emergence of order from an initially disordered system. Self-organisation is considered an emergent property of a system as large-scale ordered patterns can develop by simple, local, recurrent interactions between composing parts.

A classical example of self-organising systems are reaction-diffusion processes (Turing, 1952), that have been shown to produce complex patterns by local interactions between the components of the system. Reaction-diffusion processes have been used to model the development of cortical maps in the primary visual cortex (Wolf, 2005) and in the primary somatosensory cortex (Ermentrout, Simons, and Land, 2009). A detailed overview is presented in Wilson and Bednar (2015).

In general, self-organisation has been suggested as a theory for the formation of the broad patterns characterising cortical maps, such that they have been suggested to develop from a process that maps high-dimensional vectors onto the bidimensional cortical surface while trying to preserve local continuity (Schwartz and Rojer, 1994). Such theory has been developed extensively by means of computational modeling. A popular branch of models relies on "input-driven" self-organisation, for which cortical maps arise by passive exposure to a set of sensory stimuli. The earliest model was presented by Von der Malsburg in 1973 (Malsburg, 1973) (shown in Figure 2.2), and comprises a sheet of input units

that are connected to a sheet of cortical units, separated into excitatory and inhibitory populations. The input units are connected to the cortical units by *afferent* weighted connections, and the cortical units are connected to each other by *lateral* weighted connections that are excitatory over short distances and inhibitory over larger distances. The inputs elicit an initial response in each cortical unit, computed as a weighted sum of its inputs via the afferent connections, which is squashed using a non-linear (e.g., sigmoidal) output function. The initial cortical activation then propagates through the lateral connections, and the net effect of the short-range excitation and long-range inhibition is a dynamic that clusters an initial distributed cortical activation into a pattern of localised 'activity blobs'. Hebbian plasticity in the afferent connections consolidates these dynamics by increasing the strength of the connections of units that are consistently co-active, such that a similar pattern of input will cause a similar pattern of blobs to emerge in the future. The afferent weights for each cortical unit are normalized by dividing them by the sum of the afferent weights. If the network is presented with many patterns from a set with some underlying statistical structure then the consolidation of the recurrent dynamics through Hebbian plasticity gives rise to a topological map pattern, such that adjacent units develop similar receptive fields (i.e., similar patterns of afferent connectivity) and thus respond selectively to similar patterns. For example, inputs describing a range of image orientations yield orientation preference maps resembling those measured in primate V1.

Another example of cortical maps development due to optimization of local continuity was given by Obermayer (Obermayer, Ritter, and Schulten, 1990; Obermayer, Blasdel, and Schulten, 1992), who used a variant of the Kohonen Self-Organizing Map (SOM) (Kohonen, 1982) to show that a continuous mapping of retinal positions and orientations of lines on two dimensions could produce patterns of receptive field preferences similar to those observed in the mammalian primary visual cortex.

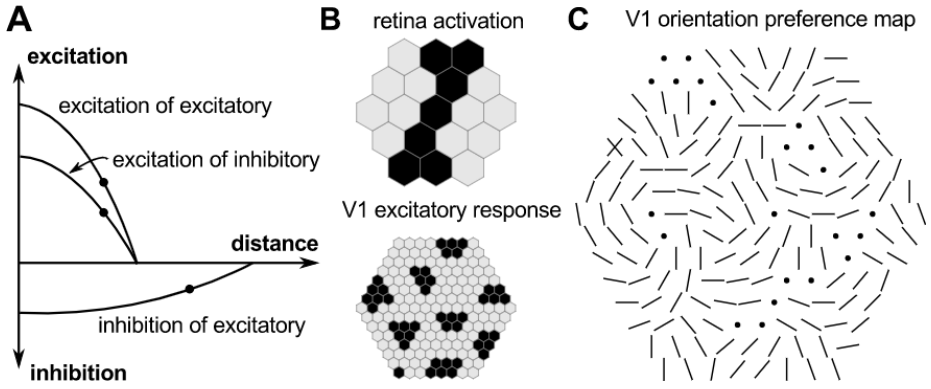


FIGURE 2.2: **Von der Malsburg 1973 model. (A)** Profile of the strength of lateral connectivity in the Von der Malsburg 1973 model of cortical self-organisation. **(B)** Spatial layout of the input (retina) and model (V1) units with an example activation due to an oriented line. **(C)** learnt orientation preference map featuring locally continuous change in preferred tuning and pinwheel points. Reproduced, with permission, from (Bednar and Wilson, 2015; Malsburg, 1973).

2.3 The L-model

The L-model (LISSOM, Laterally Interconnected Synergetically Self-Organizing Map) (Miikkulainen et al., 2006; Stevens et al., 2013; Wilson et al., 2010; Spigler, 2014; Fischer, 2014), which can be considered an extension of the first model of map self-organisation proposed by von der Malsburg (Malsburg, 1973), relies on Hebbian plasticity and short-range excitatory and long-range inhibitory recurrent interactions intrinsic to the cortical area to explain the emergence of cortical maps (Bednar and Miikkulainen, 2000; Wilson et al., 2010; Spigler, 2014; Fischer, 2014). The units in the model are arranged as a bidimensional lattice and, for computational efficiency, they are defined to be a micro-column rather than a neuron, which allows simulation of a single population of cortical units that are each able to excite or inhibit one another to support map self-organisation. The activation $\eta_j(t)$ of cortical unit j at time t is given by,

$$\eta_j(t) = \sigma\left(\alpha_A \sum_a A_{ja} x_a + \alpha_E \sum_e E_{je}\eta_e(t-\delta t) - \alpha_I \sum_i I_{ji}\eta_i(t-\delta t)\right) \quad (2.1)$$

where A is its set of afferent connection weights, E is its set of excitatory weights, I is its set of inhibitory weights, and values of α are interaction strengths of each connection, afferent, lateral excitatory and lateral inhibitory. σ is a piecewise-linear output function. x_{ja} is the input to unit j from the afferent input unit a (within the afferent receptive field of unit j).

The L-model is sometimes used in conjunction with homeostatic adaptation mechanisms (Adaptive L-model, AL) to distribute activity evenly across the network. This mechanism is implemented as a dynamic threshold (θ) of the piecewise-linear output function σ

$$\sigma(x) = \max(0, x - \theta)$$

At the end of each iteration, the dynamic threshold is updated as

$$\theta(t) = \theta(t-1) + \lambda(\overline{\eta_j}(t) - \mu)$$

where λ is the homeostatic learning rate, and μ is the target average activity. $\overline{\eta_j}(t)$ is a smoothed exponential average of the settled activity of each model unit

$$\overline{\eta_j}(t) = (1-\beta)\eta_j(t) + \beta\overline{\eta_j}(t-1)$$

with β being a smoothing parameter.

The L-model then extends the 1973 model of von der Malsburg by allowing the recurrent weights between cortical units to change according to the same Hebbian rule as for the afferent weights,

$$w_{jk}(t) = \frac{w_{jk}(t-1) + \epsilon_p \eta_j \eta_k}{\sum_p w_{jp}(t-1) + \epsilon_p \eta_j \eta_p} \tag{2.2}$$

where w_{jk} may be the weight of an afferent connection (i.e., by setting $w_{jk} = A_{jk}$ and $\eta_k = x_k$), an excitatory connection (i.e., $w_{jk} = E_{jk}$), or an inhibitory connection (i.e., $w_{jk} = I_{jk}$), ϵ is the learning rate, and p is an index over the units for which there are corresponding weights in the set A, E, or I. The schematics of the L-model are shown in Fig. 2.3, together with an example of how the equations of the L-model lead to the emergence of complex patterns of activity.

An *iteration* of the L-model algorithm occurs at integer timesteps ($t = 1$, $t = 2$ etc.), and each iteration involves defining an input pattern, applying Equation 2.1 to all cortical units τ times to allow the dynamics to settle ($\delta t = 1/\tau$), applying Equation 2.2 to modify the weights, and then resetting all activity in the network to zero before the next iteration.

It is important to emphasize that the L-model does not assume that long-range inhibitory interactions are implemented via long-range inhibitory connections in the cortex. Long-range inhibitory interactions may be implemented via

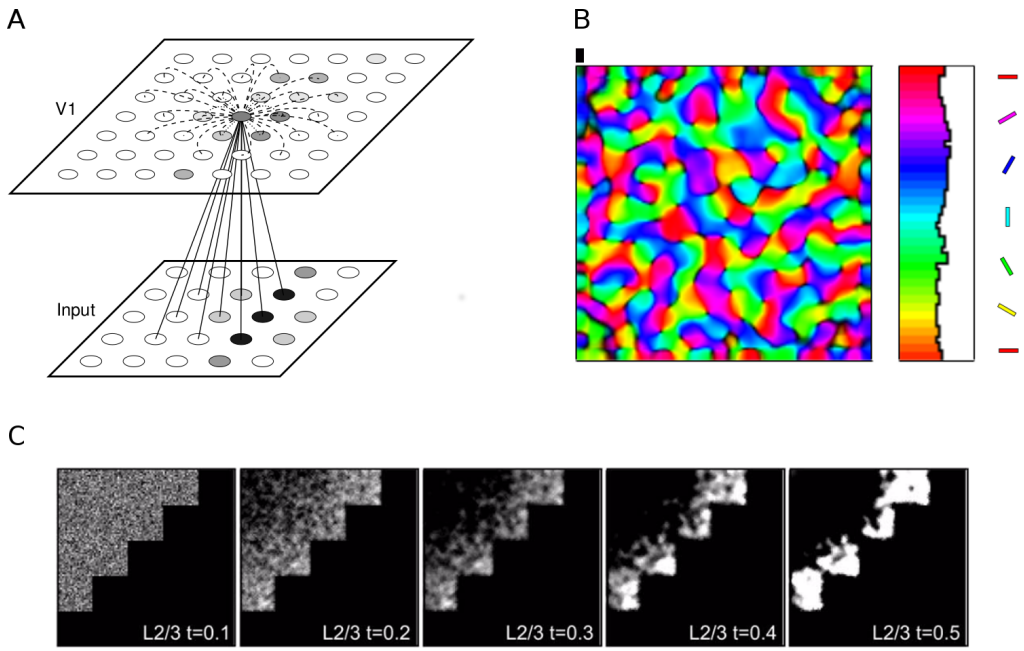


FIGURE 2.3: **Schematics of the L-model.** **A**. Schematics of the L-model. Model units are arranged as a bi-dimensional lattice, and are connected via short-range excitatory and long-range inhibitory connections. **B**. Orientation map simulated using the L-model with a histogram of the orientation preferences of the model units. Compare with Figure 2.1C. Adapted from (Miikkulainen et al., 2006). **C**. Example of how the equations of the L-model lead to the emergence of complex patterns of activity in a model of the barrel cortex, adapted from Figure 2 of (Wilson et al., 2010).

long-range excitation of local inhibitory neurons, as is thought to be the case in e.g., V1 for high-contrast visual inputs (Hirsch and Gilbert, 1991; Weliky et al., 1995) (also see refs. (Ren et al., 2007; Martin, 2002; Somers et al., 1998; Silberberg and Markram, 2007)), and in S1 for strong tactile inputs (Helmstaedter, Sakmann, and Feldmeyer, 2009; Moore, Nelson, and Sur, 1999). The architecture of the L-model is deliberately simplified and does not reflect the detailed anatomy of cortical connectivity. Related models with more complex architectures have demonstrated how the more elaborate circuitry in animal cortices could yield similar results (Law, 2009; Bednar, 2012), but these require many more parameters and more complicated analysis methods. Whether long-range inhibition is implemented by monosynaptic or disynaptic *connections* is not important for the present modelling results, only that *interactions* be net inhibitory at long distances. See Wilson et al. (2010) and Stevens et al. (2013) for further discussion.

Further elaborations of the algorithm to include biologically plausible mechanisms of homeostatic adaptation (e.g., a dynamic threshold in the output function σ) yield maps that match all available data on the patterning, stability, and robustness of (non-rodent) mammalian maps (Stevens et al., 2013). The ability of Hebbian-modifiable lateral inhibition to explain these data motivates the L-model as a strong theory of cortical self-organisation (Bednar and Wilson, 2015; Wilson and Bednar, 2015).

2.4 Distributed cortical representations in the L-model

A crucial question in neuroscience is how activity in the neocortex encodes the representation of perceptual stimuli (Pennartz, 2015). Evidence exists in support of both *distributed* representations, that make use of widespread patterns of activations to represent objects by the joint activity of a group of units responsive to a large number of stimuli, and *sparse* ones, that are instead characterised by a small average number of active units (Connor, 2005; Quiroga et al., 2008;

Quiroga, Fried, and Koch, 2013). Figure 2.4 shows an example of representations with different degrees of sparsity. In the extreme case in which complete features or stimuli are represented by individual units or by a small cluster of units, the representation is referred to as *localist*. It is important to observe that distributed coding makes it possible for different objects to produce overlapping cortical representations, where a shared subset of neurons responds to both, while neurons in localist codes are generally selective to more specific objects (Page, 2000).

Distributed representations can subserve population coding, in which a perceptual variable (e.g., orientation of a line or colour of a patch) or the identity of an object may be recovered from the responses in a population of neurons as a linear combination of the features to which those neurons typically respond, weighted by their activation. For example, population activity in cat motor cortex has been found to represent the 3D location of the paw as a vector sum of the paw locations preferred by the individual active neurons (Ethier et al., 2006). A similar type of distributed representation has been found to encode complex shapes in primate visual area V4 (Pasupathy and Connor, 2002), and to encode objects as a combination of simpler features and smaller parts in primate temporal cortex (Wachsmuth, Oram, and Perrett, 1994; Tsunoda et al., 2001). In general, areas representing complex objects like faces might use a population representation, constructed as the joint activity of neurons selective to similar objects or their parts (e.g., in representations of faces, neurons that are selective to specific eyes, mouths etc.).

The L-model in particular can support both distributed and sparse representations, to different degrees depending on the strength of the inhibitory connections between the model units. Indeed, activity in the model settles to an equilibrium that depends on the balance between afferent and excitatory input and lateral inhibition. Previous studies have exploited the model's intrinsic population coding to read out perceptual information such as the perceived orientation

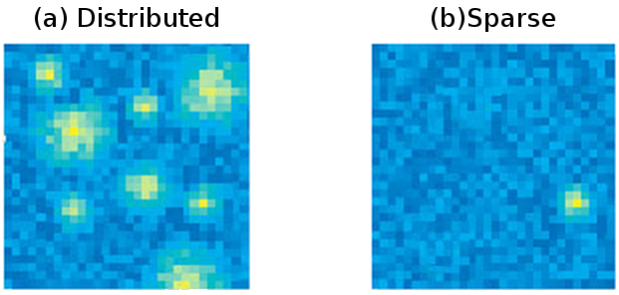


FIGURE 2.4: **Distributed vs sparse representations.** Comparison of an example (a) distributed versus (b) sparse representation coded by the firing rate of neurons in a square patch of cortex (the color of each pixel represents the activity of a neuron). Adapted from (Connor, 2005).

of gratings (e.g., as measured in terms of the tilt-aftereffect Bednar and Miikkulainen, 2000) and perceived colour (e.g., as measured in terms of the McCollough Effect, for example in Figure 2.7 adapted from Spigler, 2014).

The L-model can also produce an approximate parts-based population code as in primate visual area V4 (Pasupathy and Connor, 2002) and inferotemporal cortex (Wachsmuth, Oram, and Perrett, 1994; Tsunoda et al., 2001). Figure 2.5 shows an example of the encoding of the cortical representation of a *whole* object composed by two independent parts (*part 1* and *part 2*). The figure compares an example of data recorded by Tsunoda (Tsunoda et al., 2001) showing distributed and overlapping, parts-based cortical representations (Fig. 2.5A) in the monkey inferotemporal cortex with a similar result from a simulation using the L-model (Fig. 2.5B). In particular, subsets of the cortical units involved in the representation of complex 'whole' objects are selective to the component features. It is interesting to observe that the representations produced by the L-model, like the original data from Tsunoda, can be overlapping. In the specific example shown, the overlap is between the whole object and its composing parts, but the same result is present for two different 'whole' objects that share a part.

2.5 How does plasticity in lateral interactions influence cortical dynamics?

A useful way to understand the influence of synaptic plasticity in the lateral interactions on cortical dynamics is to look at how different theories of cortical function based on either fixed or plastic connectivity are capable of explaining experimental data.

The first theory is a model of cortical map self-organisation proposed by Von der Malsburg in 1974 (Malsburg, 1973), which was reviewed in depth in Section 2.2. The main feature of the model is that each unit, arranged in a bidimensional sheet that approximates a patch of neocortex, is connected with its neighboring

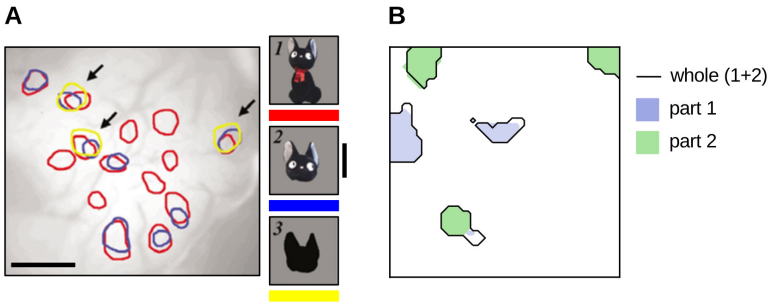


FIGURE 2.5: **Distributed representation of objects.** **A**. Optical imaging of the inferotemporal cortex (IT) of a macaque showing patches of neurons selective to parts of an object. The cortical representation of the whole object significantly overlaps with that of its parts. Adapted from Fig. 3b of (Tsunoda et al., 2001). **B**. Similar cortical organisation emerges in the L-model of (Stevens et al., 2013) in a simulation of 48 by 48 cortical units. The model units that are active by the object above a threshold of 0.3 are surrounded by a black line, and those which are active by presentation of either of its constituent parts are coloured blue or green. The threshold for visualization was chosen to reveal how the representations learnt by the model are distributed across the network. Note that a high threshold masks some units that participate in a given representation, whereas a low threshold exaggerates the contribution of units with poor stimulus tuning. Note that even though the L-model is here applied to high-level inferotemporal areas, it was originally designed as a model of early sensory cortices. The use of the L-model in this context is justified by similarities in the cortical circuits of the different areas, particularly in the presence of intra-area plastic lateral interactions and from experimental evidence that shows that higher cortical areas also feature a topographical map organization (e.g., Tanaka, (2003)). The figure is adapted from (Spigler and Wilson, 2017).

units by means of fixed lateral connections whose strength follows a "Mexican hat" profile, with short-range connections being excitatory and long-range ones being inhibitory. The model was successful in showing that such a fixed profile of lateral connectivity is sufficient to produce tuning preferences distributed in broad spatial patterns similar to those observed in cortical maps (reviewed in the next Section, 2.1).

Indiscriminate connectivity between neurons with different tuning is, however, in contrast with experimental evidence that suggests specific connectivity usually arranged in patches and linking neurons with similar tuning preferences (e.g., in the visual cortex (Gilbert and Wiesel, 1989; Weliky et al., 1995; Bosking et al., 1997); see Figure 2.6). The ring model of orientation tuning in the visual cortex (Ben-Yishai, Bar-Or, and Sompolinsky, 1995) addressed this problem by adopting a similar mexican hat profile of lateral connectivity, that was not however based on the bidimensional space of the cortex but rather in feature space, along the dimension of input line orientation. This different pattern of connectivity was found to be sufficient to produce contrast enhancement resulting, under appropriate parameters, in narrow tuning curves of the model units even in the case in which the afferent connectivity yielded low selectivity.

Intuitively, Von der Malsburg's model and the ring model can be integrated by allowing the lateral interactions to be modified by Hebbian plasticity, as this would allow learning connection strengths that depend on the similarity between the tuning preferences of pairs of model units. An example of this approach is the L-model (Stevens et al., 2013), that like Von der Malsburg's model simulates a patch of cortical surface as a bidimensional lattice of laterally-connected units. While plasticity in the lateral interactions can improve the model by letting it learn tuning-specific connectivity, thus removing the limitation due to fixed homogeneous connectivity in Von der Malsburg's model, it can also yield further advantages over the ring model where the connectivity is specific. In particular, it has been hypothesised that plastic lateral interactions

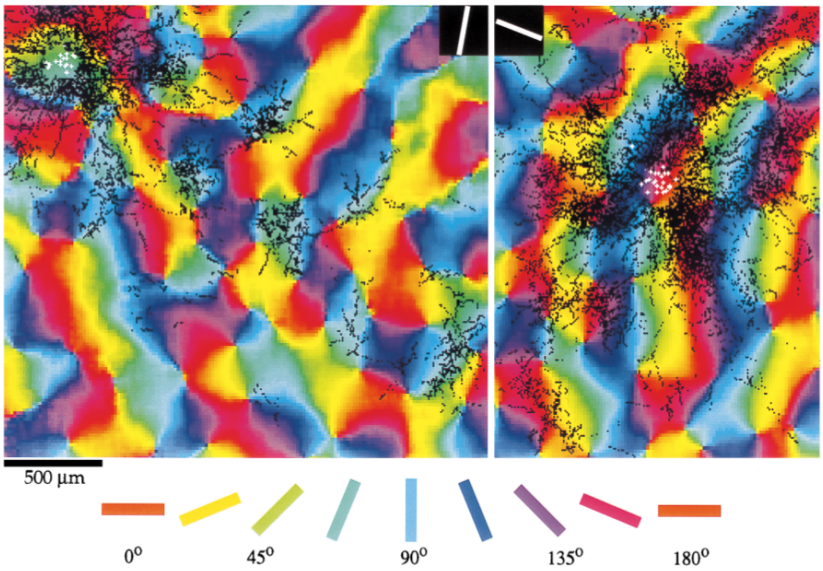


FIGURE 2.6: **Specific lateral connectivity in the primary visual cortex.** Distribution of synaptic boutons (labelled in black) for two cortical minicolumns (injection sites marked in white) in the primary visual cortex of a tree shrew. Near the target neurons, the connectivity is uniform and generic. At longer distances, however, the connectivity is found to be specific and restricted to afferent neurons with similar tuning preferences. The figure is from (Bosking et al., 1997).

can be advantageous to dynamically encode non-stationary changes in the distribution of the input statistics and use this information to decorrelate the tuning preferences of cortical units (Bernacchia and Wang, 2013; Bednar and Miikkulainen, 2000; King, Zylberberg, and DeWeese, 2013; Sippy and Yuste, 2013).

To understand how plasticity in lateral interactions affects cortical dynamics, the approach in this thesis is to examine short-term and medium-term cortical dynamics through the lens of the L-model, which has been calibrated through a number of studies to the development of cortical maps as emergent properties of cortical dynamics based on lateral interactions.

2.6 Plasticity in lateral interactions affects cortical representation

The effect of plasticity and adaptation in cortical lateral interactions on cortical representation has been indirectly observed in previous studies. In particular, it has been shown that plasticity and adaptation can explain the perceptual illusions of tilt aftereffect (Bednar and Miikkulainen, 2000) and the McCollough effect (Spigler, 2014).

The McCollough effect is a type of visual effect for which black and white gratings are made to perceptually appear colored via adaptation to an adapter grating of opposite color presented for a fixed time. The effect is contingent on the orientation of the adapter, with colorless test gratings at different orientations appearing more colored at orientations similar to the one of the adapter, and with the effect ultimately disappearing at orientations farther than 90° apart. The experimental setup thus introduces an artificial correlation between a specific orientation and color. In a previous study, I investigated how this correlation can produce plastic changes in the lateral interactions in the early primate visual areas using the L-model (Spigler, 2014), and I showed how the model could fit a wealth of psychophysical data. Specifically, the model explains the

McCollough effect by a Hebbian-like increase in the strength of the inhibition between neurons selective to the orientation and color of the adapters such that the neurons activated by a colorless test grating of similar orientation would inhibit the corresponding hue-sensitive neurons and thus produce a population-decoded perception of the opposite hue (under the assumption that a white grating would produce a baseline level of activity in neurons selective to all hues equally, such that a reduction in the activity of a specific subset of them would produce a pattern of unbalanced activity similar to that produced by a grating of opposite hue).

These examples provide a first intuition as per how plasticity in cortical lateral interactions may better fit available experimental data and explain the observed neural dynamics. These results are also particularly interesting as both the tilt aftereffect and the McCollough effect require relatively short adaptation times to work (on the order of minutes), but whose effect can persist for far longer times (up to days / months for the McCollough Effect (Jones and Holding, 1975)). Another interesting aspect of these studies is that they showed that a low-level mechanism of cortical dynamics such as plastic changes in lateral connectivity can affect the high-level perception of sensory stimuli, thus suggesting an initial bridge between models of cortical self-organisation and cognitive function. In particular, they relied on a population-based decoding of the distributed representations generated by the activity of the model units to predict the perceptual judgements made by the model. These studies also suggested that important information about the cortical representation of sensory stimuli may be encoded within the synapses of lateral connections.

However, while the previous studies did use the learnt synaptic connectivity to decode the activity of the network, in order to estimate perceptual judgements from the model, they did not explore the implications of these findings. Figure 2.7 shows the perceptual judgements decoded from the model in the two experiments, compared to similar judgements made by human participants.

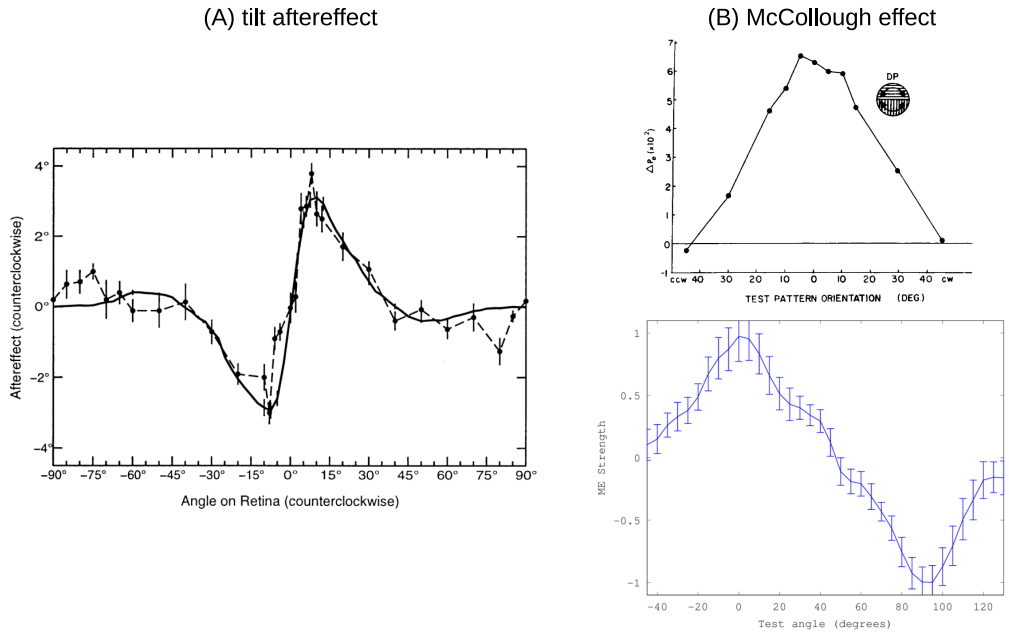


FIGURE 2.7: **Tilt Aftereffect and McCollough Effect.** **A**. Tilt aftereffect in human participants (dashed line) and in the L-model (solid line) (from (Bednar and Miikkulainen, 2000)). **B**. McCollough effect in human participants (top, only test angles in $[-45°, 45°]$ shown) and in the L-model (bottom, full range of test angles). The top panel is from (Ellis, 1977), the bottom from (Spigler, 2014).

In the next chapter we will investigate how the encoding of cortical representations in the synaptic weights of lateral interactions may lead to interference between distributed representations that share a large subset of neurons in the same cortical areas, and in particular how plastic and adaptative changes due to the presentation of one such stimulus may affect the perception of subsequent stimuli. Chapter 4 will further explore the high-level implications of this interference through a behavioural experiment.

Chapter 3

Inhibitory sharpening: simulations

The majority of this chapter has been adapted from (Spigler and Wilson, 2017).

3.1 Introduction

The more often we encounter an object, for example the more often we see a particular face or hear a particular voice, the more familiar it becomes. The first time we see a new face or hear a new voice, it evokes a *distributed* pattern of activity amongst neurons that otherwise participate in representing faces or voices with which we are already familiar. However, responses to familiar objects are usually more *localized*, to different degrees of sparsity and selectivity (Connor, 2005; Quiroga et al., 2008; Quiroga, Fried, and Koch, 2013). This chapter investigates how patterns of neural activity change as a novel object becomes familiar.

A distributed representation may be recovered from the responses in a population of neurons, as a linear combination of the features to which those neurons typically respond, weighted by their activities. For example, population activity in cat motor cortex has been found to represent the 3D location of the paw as a vector sum of the paw locations preferred by the individual active neurons (Ethier et al., 2006). A similar type of distributed representation has been found to encode complex shapes in primate visual area V4 (Pasupathy and

Connor, 2002), and to encode objects as a combination of simpler features and smaller parts in primate temporal cortex (Tsunoda et al., 2001). In general, areas representing complex objects like faces might use a population representation, constructed as the joint activity of neurons selective to similar objects or their parts (e.g., in representations of faces, neurons that are selective to specific eyes, mouths etc.).

If a novel object first evokes a distributed pattern of cortical activity amongst many neurons, then familiarization may correspond to a transition from an initial distributed representation to a more localist representation that involves the activity of a smaller subset of the original population.

This intuitive account of familiarization is indirectly supported by observations of *repetition suppression*, whereby repeated presentations of a stimulus reduce subsequent cortical responses to that stimulus (Kelly and Garavan, 2005). Repetition suppression has been demonstrated using fMRI, EEG, and single-neuron recordings, in humans and many other mammals (Li, Miller, and Desimone, 1993; Brown and Xiang, 1998; Henson and Rugg, 2003; Larsson and Smith, 2012; Henson, 2015), and it can be modulated by short-term neural habituation (Epstein, Parker, and Feiler, 2008), synchrony (Gotts, Chow, and Martin, 2012), expectation (Larsson and Smith, 2012), and attention and task-dependency (Henson et al., 2002; Henson, 2015). The opposite effect, *repetition enhancement*, can also be measured, especially at the level of single neurons, with suppression following shortly afterwards (Müller et al., 2012).

A plausible account of repetition suppression is offered by the *sharpening theory* (Desimone, 1996; Wiggs and Martin, 1998; Poldrack, 2000; Grill-Spector, Henson, and Martin, 2006), according to which a reduction in cortical activity reflects a narrowing of neuronal tuning curves and a silencing of the responses of the neurons least tuned to the stimulus. The assumptions of sharpening theory have been made explicit in a computational model (Norman and O'Reilly,

2003), in which synaptic weights in the *afferent* projections into a cortical network are modified by Hebbian learning, while neurons compete laterally to represent a given input pattern under a simple winner-take-all (k-WTA) operation, for which only the k most active units are allowed to remain active, while all the others are inactivated. The architecture of this model is consistent with a broad range of 'self-organising network' models that use similar local competitive learning mechanisms to explain the emergence of continuous topological map patterns resembling those measured in primary cortical areas (Malsburg, 1973; Dayan, 1993; Carreira-Perpinán and Goodhill, 2004; Wilson and Bednar, 2015).

Because strong lateral competition is a major component of Norman and O'Reilly's model, here we set to explore its role in a more biologically plausible manner, by using a model with explicit Hebbian-modifiable lateral interactions (Stevens et al., 2013) to investigate repetition suppression. The model accounts for the reduction in evoked cortical activity as a strengthening of lateral inhibitory interactions. Essentially, the more often a stimulus is presented the stronger the lateral inhibitory interactions between the responding neurons become, leading to an increase in the selectivity and a reduction in the spatial extent and magnitude of the response. The assumptions of this model are broadly consistent with those of sharpening theory, but the simulations presented herein suggest that plasticity in cortical afferents plays only a secondary role. Indeed, lateral plasticity alone is sufficient to account for repetition suppression. We see how this account can be falsified, by deriving a non-intuitive prediction from the model; repetition suppression for an 'adapter' object can be *disrupted* by intervening exposure to objects that produce activity that overlaps with that elicited by the adapter (i.e., by objects that have parts in common with the adapter).

A key prediction of the model is therefore that overlapping cortical representations interfere with one another. The prediction of interference offered by this account could be useful in interpreting data collected previously in a variety of

contexts, such as visual masking (Keysers and Perrett, 2002), and adaptive forgetting (Wimber et al., 2015), as well as in the context of interference between objects of different semantic categories (Cohen et al., 2014). Moreover, it is shown how this modelling prediction helps discriminate between theories of repetition suppression based on Hebbian plasticity and alternative theories, for example based on neural fatigue.

3.2 Methods

Repetition suppression has mostly been recorded in 'higher' cortical areas (Schacter and Buckner, 1998), which are characterized by large receptive fields and whose afferent input presumably represents stimuli with a degree of invariance to the lower level features represented in primary cortical areas. Our approach is to investigate repetition suppression in higher cortical areas by training the L-model (Stevens et al., 2013) with afferent input patterns that represent the minimal set of assumptions about the underlying network architecture that are required to reveal the effect. Therefore inputs to the L-model are derived from nine 'input units', with the activation of each corresponding to the presence of a particular stimulus feature such as a mouth or an eye. Each cortical unit has nine afferent weights A corresponding to the nine input units $x_i \in [0, 1]$.

In a period of pre-training, $10,000$ input patterns, each a vector with a randomly selected component set to 1 and the remaining eight set to values sampled uniform randomly in the range 0 to 0.3, were presented to the network to initialize the cortical sheet with a smooth map-like representation of the (nine-dimensional) input space. Note that pre-training was performed using individual 'parts' only, contrary to the main simulations presented here that used input stimuli composed by multiple parts. This was done to produce a parts-based population code similar to that suggested to be present in the monkey inferotemporal cortex (Tsunoda et al., 2001). Homeostatic plasticity was enabled during

this pre-training period to aid the development of continuous maps. However, as homeostatic plasticity is not a component of our account of repetition suppression it was then disabled for the simulations reported herein.

A sheet of 48 by 48 cortical units was simulated. Interaction strengths were set to $\alpha_A = 2.2$ (afferent), $\alpha_E = 1.2$ (lateral excitatory), and $\alpha_I = -2.3$ (lateral inhibitory) respectively, and the cortical dynamics were allowed to settle for $\tau = 16$ settling steps. Learning rates were $\epsilon_A = 0.1$, $\epsilon_E = 0$, and $\epsilon_I = 0.3$. All the parameters used are the same as (Stevens et al., 2013) with the exception of the interaction strengths, which were modified to increase the amount of effective inhibition. The simulations were found to be relatively robust to parameters changes. Note that maps generated by the model are indistinguishable regardless of whether or not plasticity is enabled in lateral excitatory connections (data not shown), hence plasticity in lateral excitatory connections was disabled ($\epsilon_E = 0$) to allow for a clear interpretation of the results in terms of plastic lateral inhibition (see (Miikkulainen et al., 2006; Stevens et al., 2013)). The full set of model parameters is reported in Appendix A. The model was implemented using the Topographica neural map simulator (Bednar, 2009).

3.3 Results

A set of simulations was run using the 'L-model' (Stevens et al., 2013) to investigate how repetition suppression might emerge from intracortical plasticity, according to which both afferent and lateral cortical connectivity is modifiable by Hebbian learning.

3.3.1 Plastic intracortical connectivity is sufficient to explain repetition suppression

The first simulation involved presenting the same 'adapter' pattern to the network for 100 simulated model iterations, while recording the sum of the activity

over all cortical units after the settling of the recurrent dynamics. The adapter was the pattern $\mathbf{x} = \{1, 1, 0, 0, 0, 0, 0, 0, 0\}$, which represents a simple 'object' as the configuration of two 'parts' (part x_1 and part x_2). The network clearly shows repetition suppression, i.e., a reduction in total cortical activity due to repeated presentation of the stimulus. Inspection of the pattern of activity generated by the network reveals why. Fig. 3.1 presents a comparison of the simulated cortical representation of the adapter stimulus before (blue) and after (red) repetition suppression, in which it is clear that the representation shrinks and 'sharpens' over time.

To investigate the relative contribution of afferent versus lateral plasticity to repetition suppression, the network was simulated in three cases. In the first case plasticity was enabled in both the afferent and inhibitory connections (afferent+inhibitory, i.e., the same procedure as in Fig. 3.1). In the second case plasticity was enabled only in the inhibitory connections, and the weights of afferent connections were kept fixed from time $t = 0$ (*inhibitory-only*). In the third case plasticity was enabled only in the afferent connections, and the weights of inhibitory connections were kept fixed from time $t = 0$ (*afferent-only*). As shown in Fig. 3.2, the decrease in the total activation of the model cortex depends heavily on the strengthening of the inhibitory interactions between units active in the same representation, and it occurs even when afferent plasticity is disabled. However, even though the case in which inhibitory plasticity is disabled does not cause a decrease in activity, it still produces *sharpening* in the representation of the adapter stimulus (Fig. 3.2B). Herein the model used will have both afferent and lateral plasticity enabled.

3.3.2 Interfering representations disrupt repetition suppression

Several studies have investigated the effects of interrupting repetition suppression for an 'adapter' object by presenting a number of 'intervening' objects, and

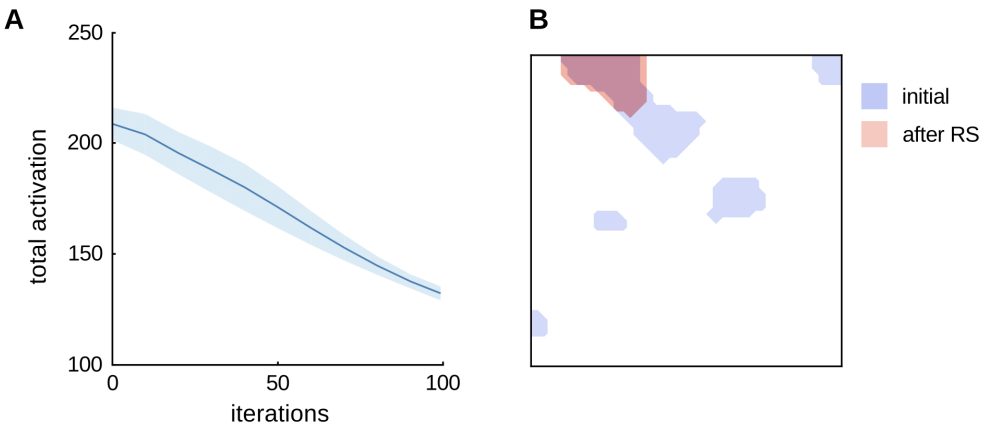


FIGURE 3.1: **Simulations showing repetition suppression.** **A**. The L model (Stevens et al., 2013) shows repetition suppression dynamics when a single input stimulus is presented to the network. The total activation is computed at each iteration as the sum of the activity of all units in the network. The plot is an average of 10 simulations ran with different random initial conditions, with the shaded area representing standard deviation. **B**. The cortical representation of the repeated stimulus is visualized by thresholding the activity of the network before (blue) and after (red) repetition suppression. Representations produced by the model are distributed across stable blobs of highly active units. After repetition suppression, the response is "sharpened" (Desimone, 1996), i.e., the sizes of blobs of super-threshold activity shrink.

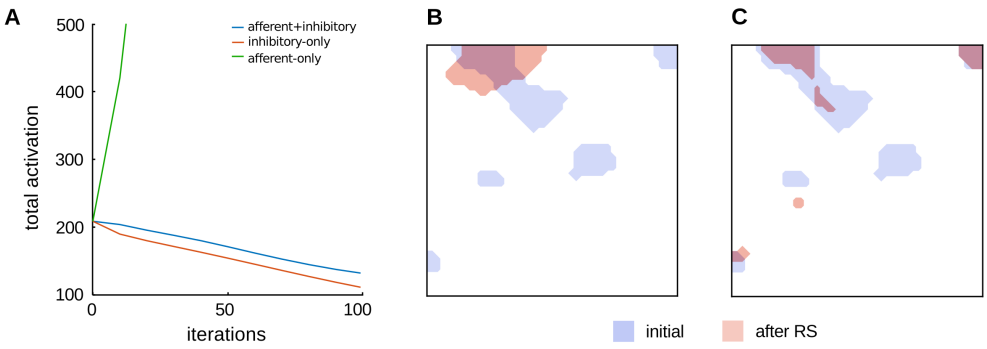


FIGURE 3.2: **Effect of afferent and inhibitory plasticity. A**. Comparison of the dynamics of repetition suppression in the L-model (Stevens et al., 2013) in three cases: plasticity enabled in both the afferent and inhibitory connections (afferent+inhibitory, i.e., the same as in Fig. 3.1); plasticity enabled only in the inhibitory connections (inhibitory-only); and plasticity enabled in the afferent connections, using fixed inhibitory connections (afferent-only). Plasticity in the inhibitory connections is revealed to be necessary to produce a decrease in the total activation of the network. We note that the case of afferent-only produces an increase in the total activation of the model, and that plasticity in the afferent connections alone is not guaranteed to decrease, as the activation depends on a balance between the magnitude of increase and decrease in activity of the individual units. We indeed observe both repetition suppression and enhancement of individual units, producing *sharpened* representations with repetition, as in previous work (Norman and O'Reilly, 2003). **B-C**. Comparison of the representation of the adapter stimulus in the afferent-only (**B**) and inhibitory-only (**C**) cases before and after repetition, which shows that regardless of the increase or decrease in activation, the representations do become *sharpened* and show a *decrease* in the number of active units.

then measuring the response to the original adapter presented again. It is interesting that there exists conflicting evidence on the effect of such designs. An early study of single neurons in primate inferotemporal cortex, for example, found that repetition suppression was unaffected by the presentation of more than 150 intervening stimuli between successive presentations of the adapter pattern (Li et al., (Li, Miller, and Desimone, 1993)). However, more recent fMRI studies with humans have reported significant differences between responses before and after interruption (Henson et al., (Henson et al., 2004; Henson, 2015)). The difference between these findings might be due to a variety of factors, from differences in the measured signals (single-neuron electrophysiology versus local field potential versus functional-MRI) to differences in protocol (stimulus type, duration, task, previous exposure to the adapters etc.) and species (human versus non-human primates).

Specifically, the hypothesis here is that intervening stimuli whose cortical representation overlaps significantly with that of the adapter (i.e., whose active neurons respond to both objects) may interfere with repetition suppression. Li et al., used stimuli less likely to produce overlapping cortical activations (line drawings of objects from various semantic categories), whereas the studies by Henson et al., used faces, that despite being unique and distinguishable from one another are processed in very localized regions of the neocortex (i.e., in the fusiform face area, FFA).

To explore these interactions further, the network was subjected to a three-phase design. In the first phase of the experiment, the adapter pattern was presented as input, thus producing repetition suppression dynamics as before. During the second phase, the network was shown a different, intervening stimulus. In the third phase, the original adapter was presented again. Each phase was run for 100 simulated steps. In what was called the 'non-overlap' condition, the intervening stimulus presented in phase 2 represented two new 'parts' that did not feature in the adapter stimulus (e.g., $x_4 = 1$ and $x_5 = 1$). In what is called

the 'overlap' condition, the intervening stimulus consisted of one part from the adapter stimulus and one new part (e.g., $x_1 = 1$ and $x_3 = 1$). The difference between these two conditions constitutes our hypothesis about the critical difference between the stimuli used by Li et al., (comparable with our 'non-overlap' condition) and Henson et al., (comparable with our 'overlap' condition); see Fig.3.3.

In the non-overlap condition, repetition suppression was not affected by presentation of the phase 2 stimulus (Fig.3.4A), and the cortical response to the adapter did not change between the final trial of phase 1 and the first trial of phase 3 (Fig.3.4B). The network response in the non-overlap condition is therefore consistent with the findings of Li et al. (Li, Miller, and Desimone, 1993). In the overlap condition, however, repetition suppression was strongly affected by the presentation of the intervening stimulus (Fig. 3.4C), which caused an increase in the response to the adapter at the beginning of phase 3, reflecting a re-organization of the representation of the adapter (Fig. 3.4D). A comparison presented in Fig. 3.5 reveals no significant difference in the total cortical response to the adapter before versus after phase 2 in the non-overlap condition (paired t-test, $t(18) = -2.0161$, $p > 0.05$), but a significant difference between the two responses in the overlap condition (paired t-test, $t(18) = -7.944$, $p < 0.0001$). Statistical tests were performed on data from 10 independent simulations, pre-trained and run with different random seeds and random initial conditions.

3.3.3 Role of plasticity in the afferent and inhibitory interactions

The effects of plasticity in the afferent and inhibitory interactions are further compared in the model by setting the learning rate of either one of the connection types to zero, and then replicating the three-phase protocol from the previous section. The comparison is reported in Fig. 3.6. We observe that the two models exhibit opposite dynamics after the intermediate phase, with a predicted

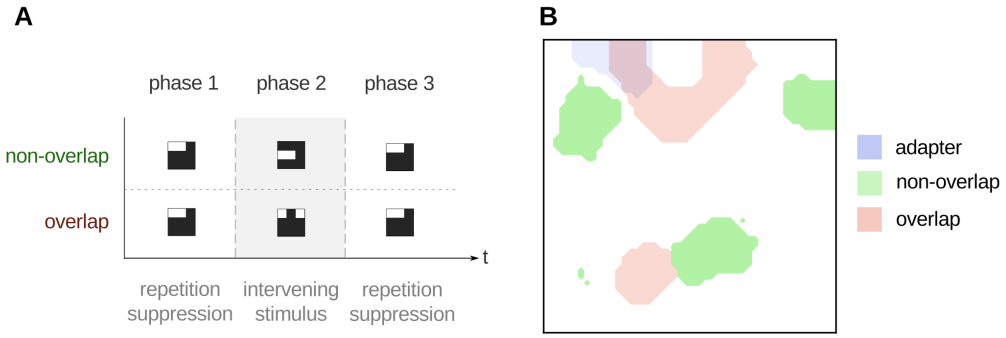


FIGURE 3.3: **Influence of intervening stimuli on the degree of cortical overlap. A**. The effect of intervening patterns in repetition suppression can be tested with a three phase protocol. First, an adapter object is presented to the network, in order to produce repetition suppression. In the second phase, the input is replaced with an intervening pattern (either overlap or non-overlap). Finally, the original adapter pattern is presented to the network again. Each phase consists of 100 iterations. The stimuli are nine-dimensional vectors (visualised here as 3 by 3 grids). **B**. Comparison of the cortical representations of the phase 1 and phase 2 stimuli, computed at the end of phase 1. At this point the network has learnt an explicit representation of the adapter (blue). However, no explicit representation of the overlap or non-overlap stimuli has emerged. The intervening stimuli (phase 2) use some ('overlap' object; red) or none ('non-overlap'; green) of the components of the adapter. The cortical representation to the non-overlap pattern has minimal to zero overlap with that of the adapter.

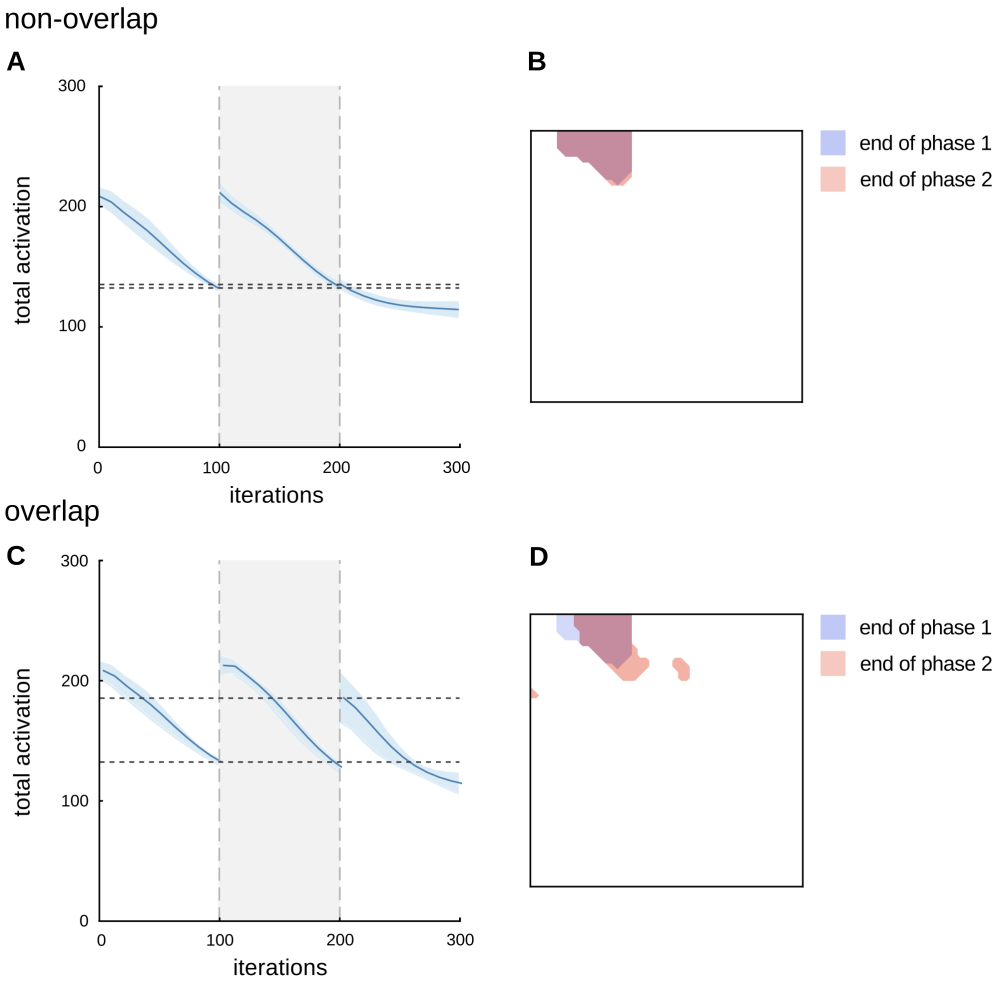


FIGURE 3.4: **Dynamics for the non-overlap versus overlap simulations.** **A**. The level of activity in the network is not affected by an intermediate phase in which a different (non-overlap) stimulus is presented. The plot shows the model activity averaged over 10 simulations pre-trained with different random initial conditions. The shaded area is the standard deviation. **B**. The cortical representation of the adapter does not change during the intermediate phase (iterations 100 to 200) when presented with the non-overlap stimulus. Indeed, there is no interaction between the representation of the adapter and intervening stimuli. **C**. The activity generated by the model is different when the overlap stimulus is used instead of the non-overlap stimulus. After the intermediate phase, the activity increases rather than remaining constant (as it does in panel A). **D**. The representation of the adapter stimulus changes during the intermediate phase when the overlap stimulus is used. Note that the total activation computed in panels A and C is the sum of the activity of all units (not the number of active units shown in the cortical representation; B,D).

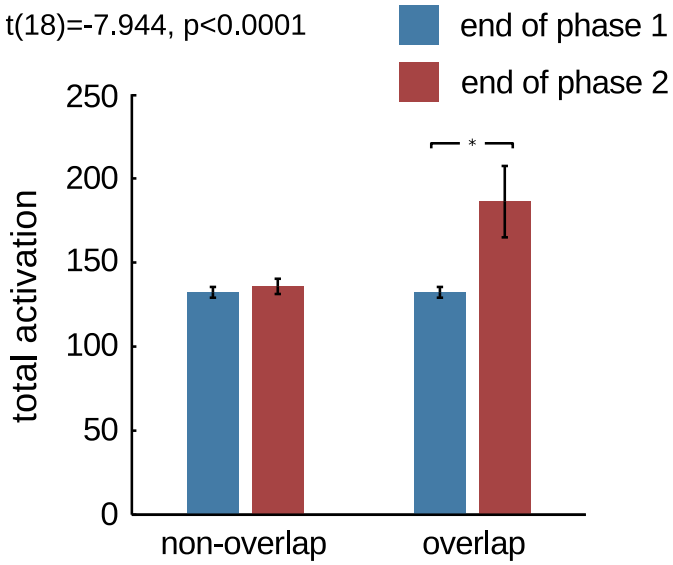


FIGURE 3.5: **Non-overlap versus overlap.** Comparison of the difference in activity produced by the adapter stimulus before and after the intermediate phase (iteration 100 versus 200), in the non-overlap and overlap simulations. The difference due to the overlap stimulus was statistically significant (paired t-test, $t(18) = -7.944$, $p < 0.0001$). Any difference due to the non-overlap stimulus was not significant (paired t-test, $t(18) = -2.0161$, $p > 0.05$). The tests were performed on data from 10 different simulations, using models pre-trained and ran with different random seeds.

increase in activity in the inhibitory-only case, similarly to the results from the combined model, and a further *decrease* in the afferent-only simulations. However, both models show a spreading of the representation of the adapter stimulus after the intermediate phase, as an un-sharpening of its representation. This is interesting as the afferent-only model shows a similar behavior in the first phase, producing sharpening (as seen from the changes in the model representation, see Fig. 4B in the main text), that is not accompanied by a decrease in the total activity. Indeed, while afferent plasticity continually improves the tuning of the model units to the adapter, thus increasing their activation, the lack of lateral inhibitory plasticity means that units that are active when the adapter is presented, but whose preferred features were not co-occurring before its presentation, have no means to inhibit each other and thus to compensate the increase in activity. This limitation, however, might not apply to other models based on fixed inhibition between the units, for example in Norman and O'Reilly (2003), where inhibition is modelled with a k-WTA operation uniform across all the units.

3.4 Discussion

The self-organising models of Stevens and colleagues (Stevens et al., 2013) was developed to model the emergence of topological maps in primate neocortex, and has been then used to model a variety of cortical dynamics. Notably, while the L-model was originally used to model long-term developmental processes, it was later successfully applied to dynamics with shorter timescales, for example the Tilt-Aftereffect (Bednar and Miikkulainen, 2000) and the McCollough Effect (Spigler, 2014). The distinguishing feature of the theory is that both afferent and lateral connectivity is updated using mechanisms of Hebbian plasticity. As a consequence, intra-cortical interactions strengthen between units that are co-active. In particular, repeated presentations of the same stimulus produce a

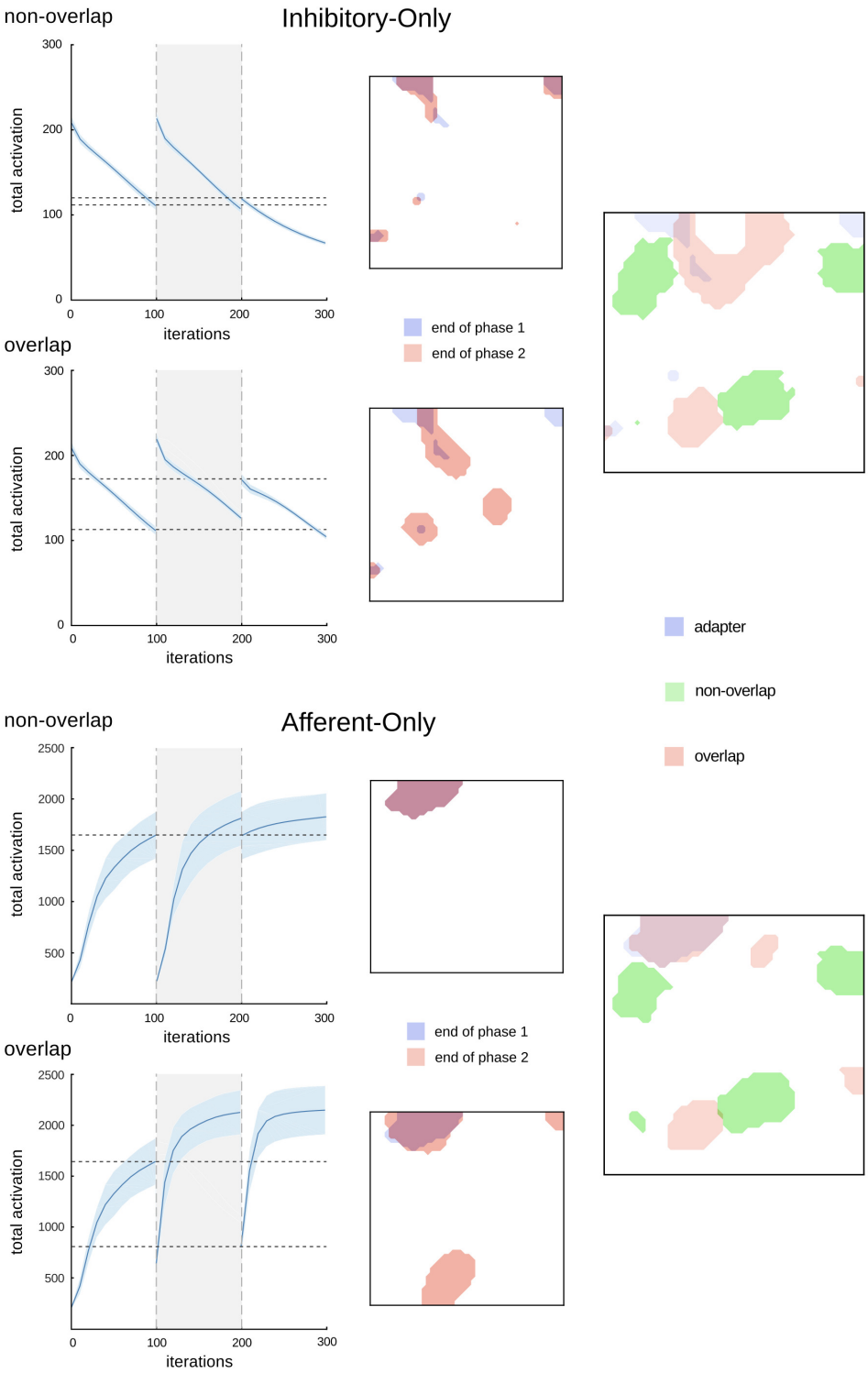


FIGURE 3.6: **Overlap vs Non-Overlap in Inhibitory-Only vs Afferent-Only.** Comparable with Fig. 3.4. The top panels present the Inhibitory-Only case (plasticity enabled only in the inhibitory lateral connections), while the bottom ones present the Afferent-Only one (plasticity enabled only in the afferent connections). The panels within each case are equivalent to those in Fig. 3.4. The two boxes at the right show a visualization of the activation patterns of the adapter, non-overlap and overlap stimuli at the end of the first phase, as in Figure 3.3 (B).

strengthening of the inhibitory interactions between the units that are recruited into its cortical representation (i.e., the units activated by the presentation of the stimulus), and thus a lowering of the overall level of activity. We suggest that such factors might underlie the phenomenon of repetition suppression.

A previous model has shown that sharpening can arise from plasticity in afferent projections alone, if strong competition between the model units is present (Norman and O'Reilly, 2003). However, repetition suppression was only measured in individual units, and the authors explain that the overall activation in the model is not guaranteed to decrease with repetition, as it depends on a balance between the magnitudes of suppression and enhancement of individual units. Further, this model approximated the net effects of recurrent inhibitory interactions in the neocortex using a simple winner-take-all operation, which may only account for few of the complex interactions that emerge from plasticity in real biological networks. Contrary to the the work by Norman and colleagues, in this study the recurrent cortical interactions that mediate local competition were explicitly simulated, and it was shown that plasticity in the lateral inhibition between cortical units is sufficient to account for repetition suppression, even without afferent plasticity (Fig.3.2). Our main simulations include both lateral and afferent plasticity, hence the present results do not challenge the idea that afferent plasticity contributes substantially to repetition suppression. Instead, we claim that repetition suppression reflects a combination of both afferent and lateral plasticity.

This account is broadly consistent with sharpening theory, according to which a reduction in the cortical response reflects a narrowing of tuning curves and therefore an increase in the selectivity of neuronal activity. The current model extends sharpening theory by emphasising also the role of intra-cortical plasticity. According to this extended 'inhibitory sharpening' model, tuning curves narrow due to the effects of both afferent and lateral plasticity. Essentially, the co-activation of units recruited in the representation of the adapter stimulus

causes a strengthening of mutual inhibition between them via Hebbian plasticity, and as this mutual inhibition builds over time the responses of individual units become more selective, the overall cortical response decreases, and the least selective neurons are silenced.

Alternatives to the sharpening theory are theories based on neural fatigue, according to which repetition suppression reflects a depletion in the resources required by neurons in order to spike (Li, Miller, and Desimone, 1993; Grill-Spector, Henson, and Martin, 2006). Neural fatigue theories seem to be supported by single-unit studies showing that the greatest reduction in cortical activity is attributable to the neurons that respond most strongly to the first presentation of an adapter stimulus. However, the inhibitory sharpening account provides an alternative explanation. By Hebbian association, the units that happen to be most active upon first presentation of the adapter stimulus subsequently develop the strongest mutual inhibition.

A further note is that the dependency of the inhibitory sharpening theory on plastic lateral connectivity makes its dynamics consistent with the predictive coding framework, which also offers an alternative interpretation of repetition suppression compared with theories based on neural fatigue and sharpening (Rao and Ballard, 1999; Huang and Rao, 2011; Friston, 2005). According to predictive coding, each cortical area predicts the incoming sensory signal, and makes the unpredicted portion of the signal (the prediction error) available to subsequent processing areas. Repeated presentation of a stimulus leads to synaptic changes that improve the ability to predict future stimuli, reducing the prediction error and thus reducing levels of cortical activation (Grotheer and Kovács, 2016; Auksztulewicz and Friston, 2016).

Interestingly, when a neural mass model of cortical dynamics was inverted to fit empirical data, the assumption of an intrinsic (intra-area) and extrinsic (inter-area) cortical connectivity which reduced exponentially with stimulus presentations could explain most of the suppression (though an additional phasic term

helped increase the fit) (Garrido et al., 2009). The authors interpreted the changes in intrinsic and extrinsic cortical circuitry in terms of the perceptual and plastic components of the computations required for predictive coding. Specifically, they reported a consistent decrease in coupling in the intrinsic connectivity following the first stimulus presentation, which is broadly consistent with the Hebbian buildup of recurrent lateral inhibition predicted by the present model. An extension of the present model to include extrinsic interactions between cortical areas, guided by the predictive coding framework, may allow for a mechanistic account of the contribution of plastic recurrent cortical interactions to hierarchical cortical computation. Moreover, as the cortical interactions simulated in the present model are known to subserve topological map formation, this approach could provide a theoretical bridge between predictive coding (acting on psychophysical timescales) and map development (acting on developmental timescales).

Other avenues for future research include establishing the relationship between inhibitory sharpening and other known accounts of repetition suppression framed in terms of increased speed of processing (James and Gauthier, 2006) and enhanced neural synchronization (Gotts, Chow, and Martin, 2012).

An interesting of our model, and hence of the extended 'inhibitory sharpening' theory that it represents, is demonstrated in Figs. 3.4 and 3.5. Experimental confirmation of the prediction that repetition suppression may be modulated and disrupted by stimuli with a cortical representation that overlaps that of the adapter (e.g., comprising a subset of the features of the adapter stimulus), would constitute preliminary evidence in support of the inhibitory sharpening theory. In contrast, in the same protocol neural fatigue would likely predict a further decrease in cortical activity, as the shared units would undergo further repetition suppression independently in each of the three phases. Sharpening would also predict a further suppression of the activity due to the overlap stimulus, but the decrease could be minimal or absent depending on the narrowing of the tuning

of the neurons selective to the adapter stimulus after its first repetition. Still, even a positive verification of the proposed predictions would require further supplementary investigation to determine whether the effect was due to lateral inhibitory plasticity, as in the inhibitory sharpening model, or whether other dynamics could explain the same data. For example, predictive coding might exhibit more complex dynamics, such that the presentation of the overlap stimulus could introduce a new statistical co-occurrence between the parts/features shared by the adapter and the overlap stimuli, and the novel parts of the overlap stimulus. Such co-occurrence would not be observed on the second repetition of the adapter, in the third phase of the protocol, which would result in an increased error due to the un-predicted mismatch and thus an increase in activity similar to that from inhibitory sharpening. This is however not too surprising, as plastic recurrent lateral connections can learn the statistical co-occurrence of features (Bednar, 2012; Bednar and Miikkulainen, 2000; Miikkulainen et al., 2006). Different types of dynamics may still however exhibit similar dynamics even without requiring lateral interactions, for example if higher cortical areas were to feed back novelty detection signals.

To understand why the L-model predicts an increase in activation after the intermediate phase, it is useful to look at the changes in the representations of the stimuli before and after each of the three phases (Fig. 3.7). During the first phase of repetition suppression, the co-activation of units recruited in the representation of the adapter stimulus led to a strengthening of their mutual inhibition by Hebbian plasticity, and thus to a suppression of responses in a subset of units (Fig. 3.1B). However, presentation of a second stimulus sharing features of the adapter increased the inhibition between units selective only to the second stimulus and units responding to both, and further led to some of the units responsive to both to drop out of the representation of the adapter stimulus and into the representation of the second stimulus. Thus, some of the units in the representation of the adapter were suppressed, while others had

the distribution of their inhibitory inputs shifted towards units selective for the second stimulus (see Figures 3.8 and 3.9), due in part to divisive normalization of the synaptic weights (Eq. 2). When the adapter pattern was presented again in a third phase, the total inhibition received by the units suppressed during the first phase was reduced. This is because either the inhibitory weights between units representing the adapter had decreased or the inhibiting units were no longer active. The variety of inhibitory interactions is illustrated in Fig. 3.7, which shows the change in the influence of one pre-synaptic unit over four post-synaptic units. Another example is shown in Fig. 3.8.

The model can account for why some studies have found that repetition suppression is affected by intervening patterns whereas others have not, in terms of differences in the choice of the stimuli. In particular, Li and colleagues (Li, Miller, and Desimone, 1993) used stimuli that were sufficiently different from one another, which could therefore have produced cortical responses with little overlap and hence little interference. Henson et al. (Henson et al., 2004; Henson, 2015), on the other hand, used pictures of faces (famous versus unfamiliar), that despite their individual differences could have elicited overlapping representations whose effect would have been further amplified by the large number of intervening stimuli (around 100). In support of the model's account of these effects, is it interesting to note that Sawamura and colleagues (Sawamura, Orban, and Vogels, 2006) found that the firing rates of neurons in monkey infero-temporal cortex depend on whether a preceding stimulus was the same (causing repetition suppression), different but capable of making the same neuron fire (causing a response similar to the prediction in our overlap condition), or different and not capable of making the neuron fire (causing a response similar to the prediction in our non-overlap condition).

A limitation of the current modelling framework is that due to the discrete timescales of the settling steps of the recurrent dynamics, and of the onset of new iterations, neurophysiological timescales in the model are difficult to reconcile

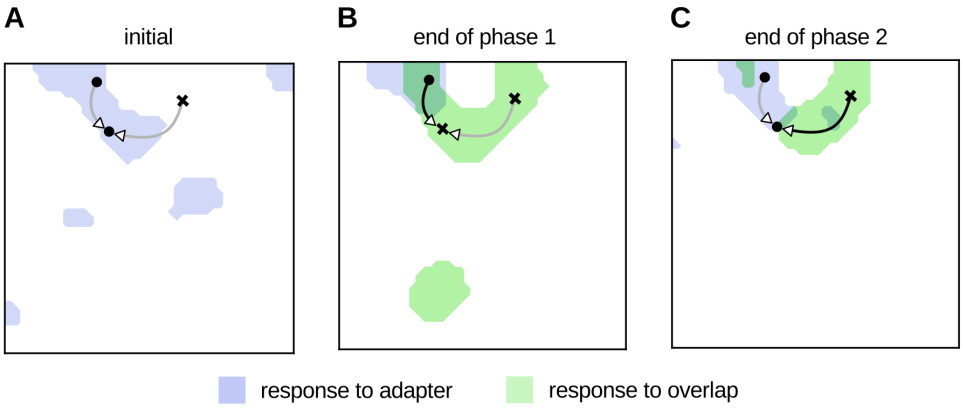


FIGURE 3.7: **Changes in lateral connectivity underlying repetition suppression.** This figure exemplifies the changes in the effective contribution of four different inhibitory connections from the same pre-synaptic neuron during each experimental phase. The width and color of the lines indicates the strength of inhibition received by the post-synaptic units; the product of the pre-synaptic activity and the weight of the inhibition. Each panel shows the cortical representation of the adapter and overlap stimuli, on the first iteration of each phase. **A**; iteration 0, **B**; iteration 100, and **C**; iteration 200. During the first phase the strength of the inhibition between the co-active units (i.e., those belonging to the same representation) increases, leading to a subset of them being suppressed. In the intermediate phase, however, the inhibition between the shared units (overlap/adapt) and the units selective only to the overlap pattern increased, and further led to some of the shared units to be removed from the representation of the adapter in favor of the overlap stimulus. Thus, some of the units in the representation of the adapter became suppressed, while others shifted the distribution of their inhibitory inputs to units in the representation of the overlap stimulus. When the adapter was presented again, the total amount of inhibition that the units that had been suppressed during the first phase received was reduced, as either the inhibitory weights between units in that representation had decreased or the units that inhibited them were not active anymore. Another example is shown in Fig. 3.8.

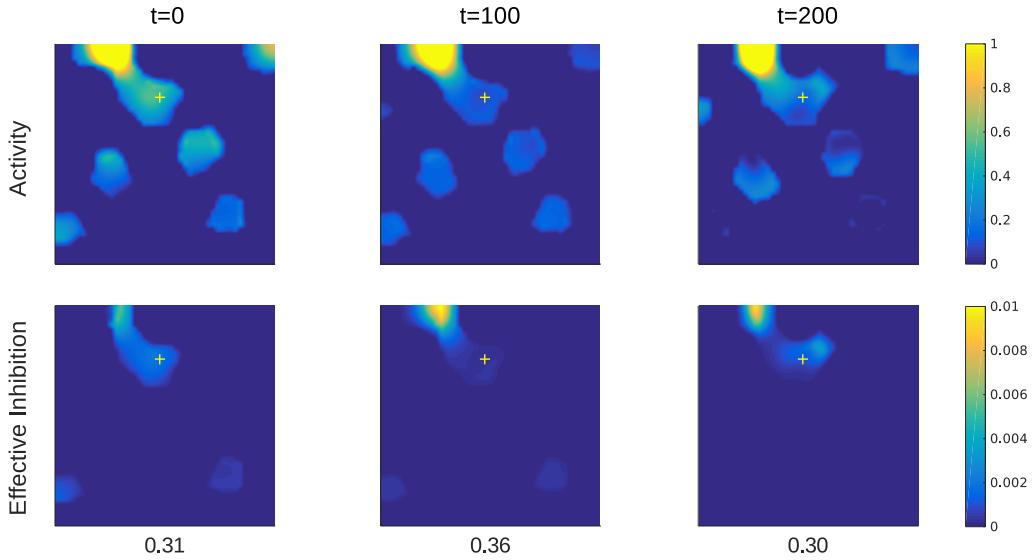


FIGURE 3.8: **Activation and effective inhibition in the three phases of the protocol** A snapshot of the activation produced by the combined model (afferent+inhibitory) at the beginning of the three phases in the overlap condition (top row), together with the effective inhibition received by the unit marked with a cross (bottom row). Effective inhibition is computed as the product of the strength of the inhibitory connection between the unit marked with a cross and each other unit, and the activation of the second unit. Below each plot is the value of the total effective inhibition received by the marked unit (i.e., 0.31 at the beginning of the simulations, 0.36 after the first phase of repetition suppression, and 0.30 after the intermediate/overlap phase). As discussed in the main text, the L-model produces an increase in inhibition between the units active in the representation of the adapter (i.e., here from 0.31 to 0.36). The presentation of the overlap stimulus, however, shifts the inhibitory weights of the units active in both the representation of the adapter and overlap stimuli to the units active in the representation of the overlap stimulus only, thus reducing the amount of inhibition that the adapter-selective units receive when the adapter is presented again (phase 3) and leading to an increase in their activity.

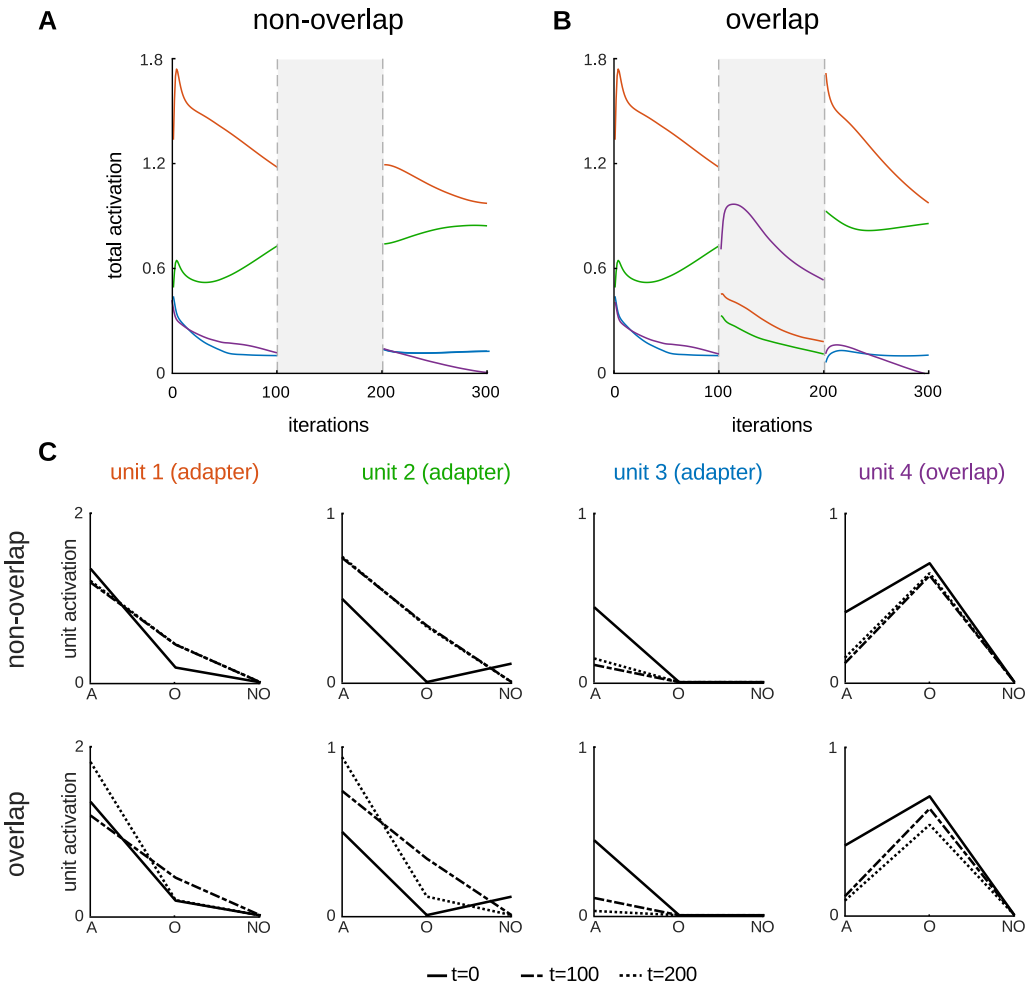


FIGURE 3.9: **Timecourse and tuning curves of sample units. A-B**. Activation of three units in the model in the two conditions (**A** *non-overlap*, and **B** *overlap*) in the combined model (afferent+inhibitory). The red and blue units show repetition suppression, with the blue unit reaching a minimum value in the first phase of the protocol. The green unit exhibits repetition enhancement dynamics. The red and green units show selectivity for the overlap stimulus, and they are both suppressed during the intermediate phase. The purple unit is more selective to the intervening/overlap stimulus. **C**. Tuning curves of the three units with a fourth that is selective to the overlap stimulus, in both the non-overlap and overlap conditions, at the beginning of the three phases ($t = 0$,$t = 100$,$t = 200$). The tuning curves were computed by presenting the three stimuli **A**-adapt, **O**-overlap and **NO**-non-overlap and allowing the dynamics of the network network to settle. The units 1, 3 and 4 exhibit repetition suppression, while unit 2 shows enhancement. It is interesting to observe that, in contrast to the sharpening theory, units 1 and 2 actually broaden their tuning during the first phase. This is due to a dis-inhibition of the units shared in the representations of the stimuli. Indeed, the presentation of the adapter leads to a strengthening of the inhibitory interactions from the units that are active in its representation, which in turn produces a decrease in the strength of those originating from units active only in the representation of the overlap stimulus, by means of weight normalization. The effect of dis-inhibition is further supported by the decrease in activity with repetition by units that responded strongly to the adapter. It is interesting to note that this broadening of the tuning curves has been found experimentally (e.g., macaque area MT [Kar and Krekelberg, (2016)]) and has been investigated in computer models of the primary visual cortex based on recurrent inhibition [Teich and Qian, 2003]. In any case, we observe sharpening in the intermediate phase for units 1, 2 and 4, with a stronger decrease in response for the less preferred stimuli.

precisely with psychological timescales. A single presentation of a stimulus on a psychologically relevant timescale corresponds to multiple simulated 'iterations' of the model. To reconcile stimulus presentations and model iterations approximately, the model was run for an extended period of 1000 iterations. The longer-term dynamics conformed to an exponential decay fit to the total activation ($y(t) = ae^{-bt} + c$), and thus match the empirically observed dynamics of repetition suppression (Li, Miller, and Desimone, 1993; Sayres and Grill-Spector, 2006). Although there is significant variability between studies of repetition suppression regarding the number of repetitions after which activation plateaus, which may depend on differences in protocol, species and recording techniques, the various estimates in the literature agree broadly that a plateau is reached within $5 - 10$ repetitions. Thus, drawing a parallel to the fitted exponential curve in Fig. 3.10, which reaches a plateau within the 1000 iterations displayed, it may be possible to consider the 100 iterations used throughout this manuscript as roughly corresponding to a single repetition of the adapter stimulus.

The mechanistic account of repetition suppression (and enhancement) offered by the inhibitory sharpening theory may be challenged further by investigating the effects of interference in perceptual discrimination tasks, using the degree of similarity and overlap in cortical representations to quantify the 'confusion' between similar stimuli. For example, multi-voxel fMRI analysis such as representational similarity analysis (Kriegeskorte, 2009; Kriegeskorte and Kievit, 2013), could be used to measure the similarity and overlap between representations, and multi-voxel pattern analysis (Norman et al., 2006) could be used to correlate the performance of a classifier built on the representations of the stimuli (measured by fMRI imaging) with the behavioural discrimination accuracy.

Experimental confirmation of the predictions of the model would provide

evidence that repetition suppression reflects a transition in the cortical representation of stimuli from a distributed to a localist encoding.

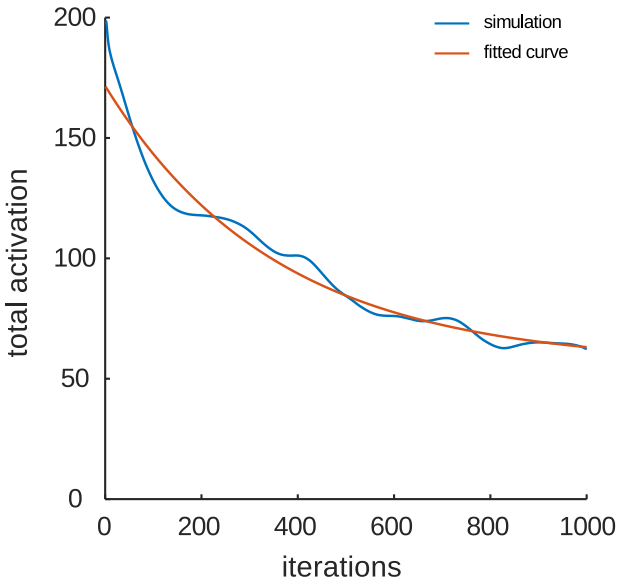


FIGURE 3.10: **Longer term dynamics of repetition suppression.** Repetition suppression in the model was simulated for an extended period of 1000 iterations. The longer-term dynamics conform to an exponential decay fit to the total activation ($y(t) = ae^{-bt}+c$), approximating the general form of the dynamics of repetition suppression measured by e.g., (Li, Miller, and Desimone, 1993; Sayres and Grill-Spector, 2006).

Chapter 4

Perceptually similar stimuli disrupt recognition performance

4.1 Introduction

Catastrophic interference has been observed in humans during sequential associative learning, for which the process of learning new associations can disrupt previously acquired knowledge (Barnes and Underwood, 1959; Postman and Underwood, 1973). It has also been observed and studied in the context of the plasticity/stability dilemma for which artificial and biological neural networks have to balance acquisition of knowledge and forgetting of previous information. In artificial neural networks, catastrophic forgetting due to sequential learning has been suggested to depend on the high degree of overlap between the distributed intermediate representations produced (French, 1992).

The aim of this chapter is to explore the predictions of the 'inhibitory sharpening' theory (Spigler and Wilson, 2017) and the L-model of cortical self-organization (Stevens et al., 2013), upon which it is based, in the cognitive domain. Specifically, we test the key prediction that stimuli that are hypothesized to produce overlapping patterns of cortical activity can interfere with one another during the process of familiarization. The prediction fits in the context of catastrophic interference and may provide insight into its underlying dynamics if found to be correct.

The study presented in this Chapter also serves as a basis for the work of Chapter 5, to test whether the set of stimuli used does produce measurable cognitive effects, that could then be looked for in an fMRI experiment.

A reduction in cortical activity caused by the continued exposure to an adapter stimulus (repetition suppression) has been linked to perceptual learning and priming, and thus with a decrease in reaction time and an increase in the recognition accuracy of the adapter stimulus (Henson, 2015; Henson and Rugg, 2003; Henson, Shallice, and Dolan, 2000), although the strength of this relationship is debated (Sayres and Grill-Spector, 2006). A variety of changes in the distribution of cortical activity have been observed during task learning (Kelly and Garavan, 2005). Here, we investigate whether the increase in cortical activity due to interference between perceptually similar stimuli predicted by the inhibitory sharpening theory can be observed as a decrease in recognition accuracy.

In this chapter, we show that the recognition accuracy can be modulated by using intermediate stimuli designed to disrupt learnt cortical representations, in line with the predictions from the sharpening theory introduced in Chapter 3.

4.2 Methods

4.2.1 Participants

Thirty-four healthy volunteers (15 female, aged 18-68, mean age 27.8) participated in the present study. The study was approved by the University of Sheffield, Department of Psychology Ethics Sub-Committee, and carried out in accordance with the University ethics guidelines. All volunteers provided informed consent to take part in the study. Data from seven participants were not included in the analysis due to poor performance in the first testing phase. Even though the threshold for acceptance was set to 80% average accuracy in the first part of the experiment, the discarded participants had scores lower than 40% and sometimes 20%, close to chance level for the specific task. The performance of the

remaining participants was instead never lower than $90\% - 95\%$ except for a single participant who scored 80% in a single condition. Twenty-one volunteers participated in the experimental condition while six participated in a control condition.

4.2.2 Materials

The procedure for presenting experimental stimuli was programmed in Python, using the PsychoPy library (Peirce, 2007), and stimuli were presented to the participants on a computer screen. The stimuli were faces generated by combining two parts, a top and a bottom half of different female faces from the Chicago Face Database (Ma, Correll, and Wittenbrink, 2015) and separated by a gap of 2 pixels, for a total size of 732 pixels in width and approximately 1000 pixels in height, to produce novel faces exploiting by exploiting the Face Composite Effect (Young, Hellawell, and Hay, 1987). No stimulus in the experiment was composed with parts taken from the same face. The stimuli were pre-processed using a combination of custom scripts to segment and cut them in such a way that random composition of the different top and bottom parts always resulted in a good alignment. Similar composite faces have been used in previous studies to investigate the difference in parts-based versus holistic representation of faces in the human visual cortex (Schiltz and Rossion, 2006; Schiltz et al., 2010) and are known to affect the perception of the individual face parts (Young, Hellawell, and Hay, 1987).

Each participant was presented with a set of five face stimuli sampled from seven different pre-generated sets for the group in the experimental condition, and five pre-generated sets of face stimuli for the group in the control condition. The five stimuli used in the experiments were split into two groups, the first comprising stimuli referred to as $F1$, $F2$, and C, and the second comprising stimuli referred to as $H1$ and $H2$. The first three stimuli ($F1$, $F2$ and C) were

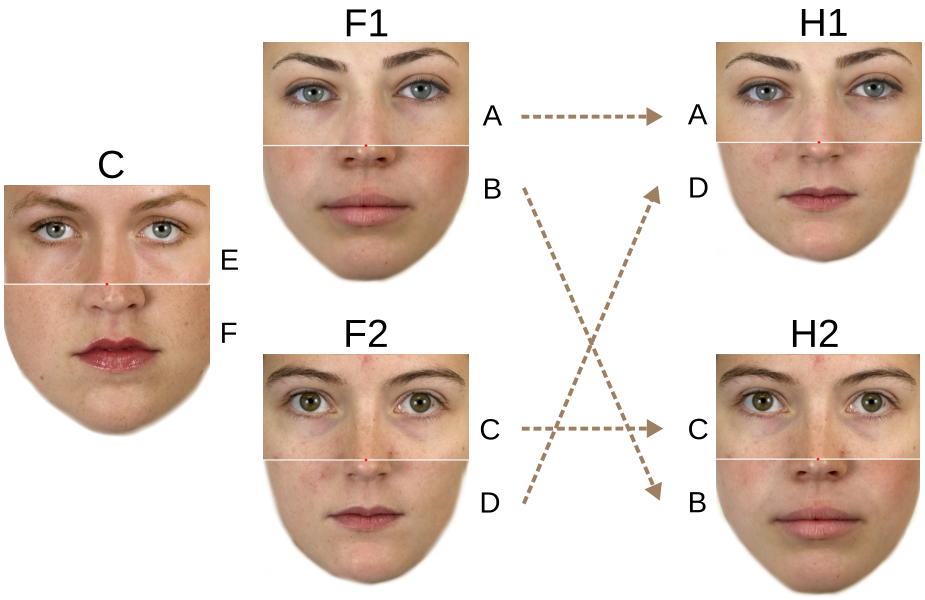
constructed in a similar way, by matching top and bottom parts belonging to different faces and without any condition specific differences. The last two stimuli however were designed differently for each condition. In the control condition both $H1$ and $H2$ were composed in the same way as the first three faces using parts from novel faces, distinct from the others. In the experimental condition however the $H1$ and $H2$ faces were composed by swapping the top and bottom parts of the $F1$ and $F2$ stimuli used in the same run, so that the top part of $H1$ was set as the top half of $F1$ and the bottom part of $H1$ was set as the bottom half of $F2$.

An example of the stimuli used is shown in Fig. 4.1. The participants watched the stimuli from a normal viewing distance (approximately 50-100cm from the computer screen), from which the stimuli subtended a visual angle of around $11^\circ - 16^\circ$ in width and $13 - 19^\circ$ in height.

4.2.3 Experimental Design

Each participant took part in four experimental phases, two 'training phases' in which a sequence of composite face stimuli was presented in conjunction with an integer label assigned to each face, alternated with two 'testing phases', one following each of the training phases, in which participants viewed the faces and were asked to recall the corresponding labels by pressing a keyboard button.

The experiment was based on a 2x2 mixed factorial design. A within-groups factor with two levels was defined by the comparison of recall accuracy between the first and second testing phases. A between-groups factor with two levels was defined by the composition of the face stimuli that were used in the second training phase. For participants assigned to the experimental condition, two face stimuli introduced in the second training phase shared a top or bottom half with two of the face stimuli presented in the first training phase. For participants assigned to the control condition, two face stimuli introduced in the second training phase were entirely novel.



- Phase 1

  Train on {F1, F2, C} x 8
  e.g., F1 C F1 F2 F2 C F2 C C F1 F2 C ....

  Test on {F1, F2, C} x 15
  e.g., C F2 F1 F1 F2 C C F1 C C F1 C F2 ....

- Phase 2

  Train on {H1, H2} x 3
  e.g., H1 H1 H2 H1 H2 H2

  Test on {F1, F2, C, H1, H2} x 15
  e.g., H1 C F1 F1 F1 H2 C C F2 C H2 ....

FIGURE 4.1: **Stimuli and protocol of the behavioural experiment.** Sets of 5 composite faces, composed of separate top and bottom parts, were used in the experiment. The experimental procedure involved training the participants to associate a number (1 to 5) to each face, first by training on the first three distinct faces (F1, F2 and C), that did not share any top or bottom part, and subsequently by presenting the final two stimuli, H1 and H2. The second set of faces presented was composed by either using novel halves, distinct from the first stimuli, in the control condition, or by swapping the top and bottom parts of the F1 and F2 stimuli in the experimental condition. During the first training phase, the first 3 faces were shown for 8 repetitions each in a randomized sequence. The second training phase presented the last 2 faces for only 3 repetitions each. Testing was performed after each of the two training phases, and involved presenting all the 5 faces (second test) or just the first 3 (first test) for 15 repetitions each, in a randomized sequence.

Dependent variables were recall accuracy and reaction times for each composite face stimulus in the two testing phases. The experiment was designed to test the hypothesis that learning to recall two composite face stimuli in the second training phase (H1 and H2), which share an upper or lower half with previously learnt stimuli (F1 and F2), will reduce recall accuracy for the two previously learnt stimuli (F1 and F2), compared to recall for a stimulus (C) with no shared features as a within-participants control, and compared to a between-participants control in which two entirely novel faces substituted H1 and H2.

4.2.4 Experimental Procedure

In the first phase, referred to as 'training phase 1', participants were shown a sequence of 24 stimuli. Each stimulus was a composite face image from a selection of three composite faces ($F1$, $F2$ and C) followed by a corresponding integer label. Stimuli were presented in a pseudo-random order such that each face was presented eight times in total.

In the second phase, referred to as 'testing phase 1', participants were shown faces from the same set of three faces, and they were requested to press a keyboard button corresponding to the integer label that was associated with each face (no longer presented on screen). Forty five faces were presented in total, with each face appearing fifteen times in pseudo-random order.

In the third phase, referred to as 'training phase 2', participants were shown two new face stimuli ($H1$ and $H2$), followed by the corresponding integer label. Six stimuli were presented in total, with each stimulus presented three times.

In the fourth and final phase, referred to as 'testing phase 2', participants were shown all five face stimuli, without the integer labels, and were required to respond by pressing the keyboard buttons 1, 2, 3, 4 or 5. Seventy-five stimuli were presented in a pseudo-random order such that each stimulus was presented fifteen times.

The association between each of the five composite faces and its integer label was randomized for each participant, as was the composition of each composite face.

Figure 4.1 shows a summary of the protocol together with an example set of faces used.

4.3 Results

4.3.1 Experiment

The recognition accuracy of human participants was recorded for each face stimulus before and after training to recognize the disruptive stimuli (H1 and H2). The results are shown in the bar plot in Figure 4.2A, averaged across the participants in each condition. While the control C was unaffected by the disruptive training, the stimuli F1 and F2 were partially forgotten due to the interference of the disruptive training, and suffered a significant drop in recognition accuracy, with a mean decrease of 26.35% for $F1$ and 23.49% for $F2$. All the stimuli were learnt to above chance level (33% in the first testing phase, 20% in the second) as confirmed by a Wilcoxon Signed Rank test ($p < 0.001$ in each condition).

A Kruskal-Wallis test found significant differences between the recognition accuracy of the three faces F1, F2 and C after disruption, $\chi^2(2) = 26.77$, $p < 0.0001$. A post-hoc Tukey-Kramer analysis showed no difference between F1 and F2 ($p = 0.97$) but a significant difference between F1, F2 and C ($p < 0.0001$). A similar test on the accuracies before disruption reported no difference between the conditions ($\chi^2(2) = 3.86$, $p = 0.14$), as was expected due to the lack of differences between the faces F1, F2 and C prior to disruption. Further, a Wilcoxon Ranked Sum test confirmed the significant decrease in performance in recognizing the faces F1 and F2 ($Z = 4.94$, $p < 0.0001$ and $Z = 4.43$, $p < 0.0001$) but no change in the control ($Z = -1.4$, $p = 0.16$). A similar difference was found in the reaction times after training to recognize the last two stimuli ($\chi^2(2) = 21.32$,

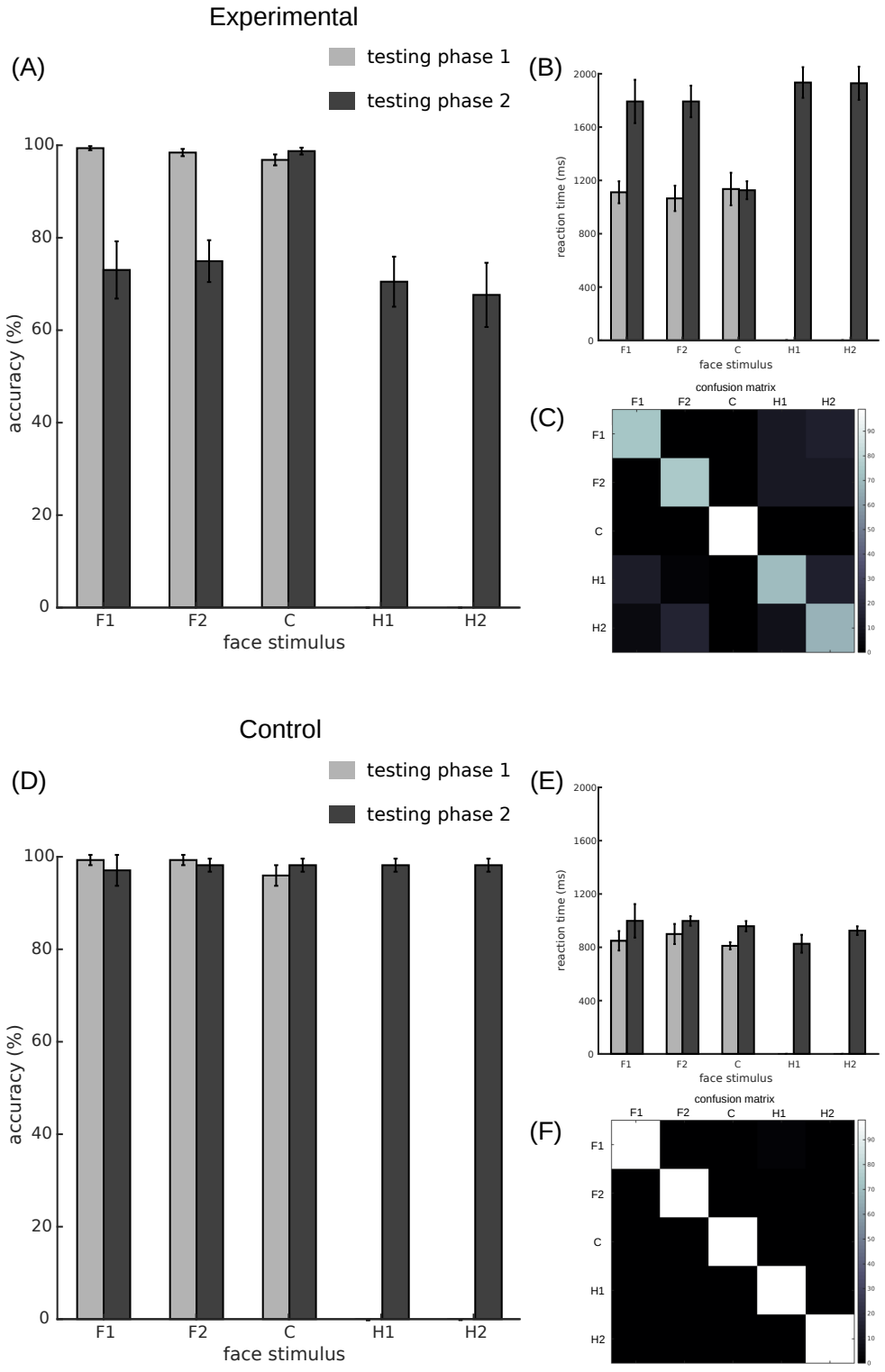


FIGURE 4.2: **Results of the experiment. A,D** Recognition accuracy for each face stimulus *before* and *after* the second block of training with the disruptive stimuli (A: experimental condition, D: control condition). The data shown is averaged across participants and bars denote standard error. **B,E** Reaction time for each face stimulus. **C,F** Confusion matrix relative to the last phase of testing over all the stimuli. Faces F1 and F2 are often confused with the distruptors H1 and H2 in the experimental condition, where the stimuli H1 and H2 are composed by swapping the top and bottom halves of the F1 and F2 faces, while no confusion between stimuli is observed in the control condition.

$p < 0.0001$), with a significant increase for the F1 and F2 faces but no increase for the face C, while no difference was present between the reaction times before the second training phase ($\chi^2(2) = 0.59$, $p = 0.74$).

Inspection of the confusion matrix for the final testing phase (Figure 4.2C) shows that the decrease in recognition performance to the F1 and F2 stimuli is due to mis-classifications between these stimuli and the half stimuli H1 and H2.

Further, the decrease in recognition performance cannot be attributed to the increase in the difficulty of the task due to learning to recognize five faces instead of three. The control condition, indeed, showed that participants can achieve high recognition accuracy in the same protocol if the last two stimuli H1 and H2 are composed with novel parts, contrary to the experimental condition in which they were made by swapping the top and bottom halves of the F1 and F2 stimuli. A Kruskal-Wallis test found no significant differences between the recognition accuracy of the three faces F1, F2 and C both before ($\chi^2(2) = 2.41$, $p = 0.3$) and after training on the last two faces ($\chi^2(2) = 0.23$, $p = 0.89$). The same test also found no differences between the reaction times (before, $\chi^2(2) = 0.61$, $p = 0.74$, and after $\chi^2(2) = 1.09$, $p = 0.58$). It is interesting to observe that even if not specific for particular cases, an increase in reaction times is observed even in the control condition. The increase, likely due to the increased difficulty of the task (recognizing five faces instead of three), is however more limited than in the experimental condition (17.6%, 10.9% and 18.2% for the control condition, compared to 61.3% and 68.3% in the experimental condition). The reaction time to the C stimulus in the experimental condition did not change significantly (-0.8% mean change), which may be due to the fact that, due to it being the only stimulus that did not share any parts with the other faces, it was the easiest to identify in the second testing phase.

4.4 Discussion

We investigated how the perception of visual objects can be affected by interference due to stimuli hypothesized to generate overlapping cortical responses under the assumption that the similarity between the patterns used, especially in sharing significant parts and features, may correlate with similarity and overlap in evoked cortical activations, at least in higher cortical visual areas (e.g., as partially observed in (Tsunoda et al., 2001)). In particular, previously learned objects were found to be partially forgotten in the process of learning to recognize different objects that share a subset of their component parts and features. The process of forgetting manifests as a decrease in recognition accuracy and an increase in the reaction time for the faces that were targeted for disruption. That is, the faces F1 and F2 in the experimental condition that were disrupted by H1 and H2. The use of a control face C that was not affected by the predicted disruption showed that the effect was selective and was not simply due to the increased difficulty of the task due to the larger number of faces to be recognized in the second testing phase (five faces versus the original three). A control condition was then explored by replacing the H1 and H2 stimuli with novel faces that did not share significant features with the first three. In this case, the performance on the task was found not to be affected by the second training phase, further showing that participants are capable of learning to recognize all five faces without interference. Together, the evidence presented suggest that the effect of decrease in recognition performance observed in the experimental condition is to be attributed to the disrupting stimuli.

The potential link between repetition suppression and performance in perceptual learning tasks makes the results of this experiment compatible with the prediction from the inhibitory sharpening theory on the effect of overlapping cortical representations. Of particular interest is that the L-model could explain similar effects of confusion between similar objects by relying on the capability of downstream neurons to decode cortical representations based on linear

decoding of the patterns of activation (for example, Pitkow et al., 2015).

It is interesting to note that similar results were found in previous work, that however used a different protocol (Barnes and Underwood, 1959; Postman and Underwood, 1973). The previous experiment involved training a group of volunteers to recognize word pair associations from a list A-B (pairing words in A with the corresponding one in B). Catastrophic interference was then observed by training them on a different set of associations A-C that linked the same words in the first list with a novel association, similarly to the interference observed in our experiment.

Chapter 5

Neuroimaging investigation of the inhibitory sharpening theory

5.1 Introduction

A useful way to study cortical dynamics in a large area of the neocortex without directly recording from the individual neurons involved is to exploit the phenomenon of *repetition suppression*, where the repeated presentation of a stimulus leads to an overall decrease in the level of stimulus evoked cortical activation (Li, Miller, and Desimone, 1993; Kelly and Garavan, 2005; Grill-Spector, Henson, and Martin, 2006). The effect has been measured using a variety of recording techniques (fMRI, EEG, single-neuron) and in a variety of mammalian species (Li, Miller, and Desimone, 1993; Brown and Xiang, 1998; Henson and Rugg, 2003; Larsson and Smith, 2012; Henson, 2015) and has been shown to be affected by attention and task-dependency (Henson et al., 2002; Henson, 2015), short-term neural adaptation (Epstein, Parker, and Feiler, 2008), expectation (Larsson and Smith, 2012), and synchrony (Gotts, Chow, and Martin, 2012). Repetition suppression measured with neuroimaging can then be used to infer the underlying neural mechanisms by means of computational models that predict fMRI responses from patterns of cortical activity (Alink, Abdulrahman, and Henson, 2017; Spigler and Wilson, 2017), as for example was done in a simplified way in Chapter 3 by simply averaging the activation of all the model units.

Different theories have been proposed to explain the neural dynamics that produce repetition suppression (Grill-Spector, Henson, and Martin, 2006). For example, the sharpening theory explains repetition suppression in terms of a narrowing of the tuning curves of the neurons involved, which leads to an increase in the selectivity of neural activity (Desimone, 1996; Wiggs and Martin, 1998). The fatigue theory instead suggests that repetition suppression reflects a depletion in the resources required by neurons in order to spike, leading to a decrease in the overall activation (Ringo, 1996). Predictive coding has also been proposed as an underlying mechanism, with the prediction error due to a novel sensory stimulus decreasing during repetition and familiarization (Grotheer and Kovács, 2016; Auksztulewicz and Friston, 2016).

Chapter 3 presented the inhibitory sharpening theory as an extension to sharpening, focusing on the contribution of plasticity in lateral cortical interactions on the effect (Spigler and Wilson, 2017). The theory has been made explicit in a computational model, leading to the central prediction that the magnitude of repetition suppression should be affected by the intermediate presentation of specially designed stimuli in between repetitions of the adapter stimulus. In particular, intermediate stimuli designed to produce a pattern of cortical activity that overlaps with that produced by the adapter should produce a disruption of the effect of repetition suppression, resulting in a smaller magnitude of suppression compared to a control case in which the intermediate stimulus produces little overlap.

This prediction can be considered in a more general context. Most areas in the mammalian neocortex exhibit distributed patterns of activation in response to sensory stimuli (Tsunoda et al., 2001; Pasupathy and Connor, 2002; Pouget, Dayan, and Zemel, 2000; Connor, 2005; Georgopoulos, Schwartz, and Kettner, 1986). A direct consequence of this is that different stimuli can trigger activation in a common subset of neurons. The single presentation of a stimulus can be sufficient to produce lasting plastic changes in the synapses of the neurons

involved, as seen for example in at least some cases of repetition suppression. There should therefore be an interaction between the responses to stimuli that activate a large sub-set of shared neurons, that is, whose cortical representations overlap. This is in contrast with the current assumption implicit in typical experiments that the effect of repetition suppression for each stimulus is independent from the others. As well as discriminating between theories of repetition suppression, it is thus also important to characterise whether a sequence of stimuli can produce interference due to the activation of a shared sub-set of neurons.

The primary aim of this study is to test the prediction of the inhibitory sharpening theory in a neuroimaging experiment, by replicating the essential components of the experimental design that was used to derive it in a simulation. In any case, even a positive verification of the predictions would require further supplementary investigation to determine whether the effect was due to lateral inhibitory plasticity, as hypothesized in the inhibitory sharpening theory, or whether other mechanisms could explain the same results.

A second more general aim is to investigate how stimuli capable of producing overlap in their cortical patterns of activity may affect the dynamics of repetition suppression. One of three potential outcomes is expected from the experiment. We should see either i) a lack of changes in the magnitude of repetition suppression attributable to cortical overlap, ii) an *increase* in the amount of repetition suppression as predicted by, e.g, the fatigue theory, due to the continued adaptation of the shared neurons, or iii) a *reduction* in the amount of suppression, as predicted by the inhibitory sharpening theory.

As in Chapter 4, we will use face stimuli because face processing in the human and monkey brain has been studied extensively (Kanwisher, McDermott, and Chun, 1997; Haxby, Hoffman, and Gobbini, 2000; Tsao et al., 2006; Tsao and Livingstone, 2008) and the relevant cortical areas involved are known to exhibit repetition suppression (Henson, Shallice, and Dolan, 2000; Henson et al., 2004; Gilaie-Dotan and Malach, 2007; Goffaux et al., 2013; Henson, 2015). In particular,

two cortical areas have been identified to respond preferentially to face stimuli, the "Occipital Face Area" (OFA, in the Inferior Occipital Cortex, IOG) and the "Fusiform Face Area" (FFA, in the lateral Middle Fusiform Gyrus, MFG), with right hemispheric dominance. The neural encoding of faces in those areas is still debated, though evidence was found for a sparse population code (Young and Yamane, 1992). Further, face processing in humans is also thought to have separate parts-based and holistic processing components, with a tendency for the first to dominate in the OFA and the second in the FFA (Schiltz and Rossion, 2006; Schiltz et al., 2010). Moreover, it is possible to produce novel faces by combining a "top" and a "bottom" part of different faces, taking advantage of the composite face effect (CFE) (Young, Hellawell, and Hay, 1987), for which the same "top" part of two faces looks slightly different if the "bottom" parts are different, as long as the two halves are well aligned. The composite face effect has been used by others to differentiate between parts-based and holistic processing of faces in the context of repetition suppression (Schiltz and Rossion, 2006; Schiltz et al., 2010), and thus appears to be a good candidate to produce the type of overlap of cortical activations that we are interested in. Indeed, cortical areas that represent the stimuli by parts would be expected to produce similar patterns of activation when the same part (top or bottom) is re-used in a different face, while holistic processing areas might lead to similar activations depending on the overall degree of similarity of stimuli sharing the same part.

5.2 Materials and Methods

5.2.1 Participants

Thirteen healthy volunteers (4 female, aged 18-28 years, mean age $= 22.8$) with normal or corrected to normal vision participated in the study. The study was approved by the University of Sheffield, Department of Psychology Ethics Sub-Committee, and carried out in accordance with the University ethics guidelines.

All volunteers provided informed consent to take part in the study.

5.2.2 Stimuli

The stimuli used in the study are faces generated by combining two parts, the top and bottom halves of different faces from the Chicago Face Database (Ma, Correll, and Wittenbrink, 2015), matched by gender, and separated by a gap of 2 pixels. No stimulus in the experiment was composed of parts taken from the same face, and each part was only used in a single stimulus. Similar composite faces have been used in previous studies to investigate the difference in parts-based versus holistic representations of faces in the human visual cortex (Schiltz and Rossion, 2006; Schiltz et al., 2010) and are known to affect the perception of the individual face parts (Young, Hellawell, and Hay, 1987). Examples of the stimuli are shown in Fig. 5.2. The images used to generate the composite faces are 84 white female and 75 white male faces, while 78 latina female faces were used to compose the targets. The targets were shown upside-down and were 85% the size of the other images. The faces were processed using a combination of custom scripts to segment and cut them in such a way that random selection of the top and bottom parts always resulted in a good alignment. Throughout the experiment, the stimuli were presented on a monitor at a distance of approximately $220 - 240cm$ that the volunteers could see through a single-mirror setup from inside the scanner. The stimuli were approximately $4.75° - 5.15°$ in width and $5.9° - 6.6°$ in height.

5.2.3 Experimental Design

Each participant took part in three scanning sessions, with a brief setup interval in between. An event-related design was used, in which a sequence of composite face stimuli was presented grouped in consecutive triplets.

The experiment used a within-subjects design with three levels. The independent variable was the configuration of the second stimulus in a sequence

of three face stimuli. The dependent variable was the difference in response to identical stimuli presented as the first and third faces in the sequence. Three conditions were defined by the choice of the second face stimulus in the sequence. In the 'Same' condition, the second face was identical to the first and third. In the 'Different' condition, the second face was different from the first and third. In the 'Half' condition, the second face was identical to the first and third in either its upper or lower half only. Sequences of face triplets from each condition were pseudo-randomly interspersed, and haemodynamic responses were recorded using functional magnetic resonance.

The experiment was designed to test the hypothesis that the magnitude of repetition suppression (response to the first minus the response to the third in each triplet) would be strongest for the Same condition, weaker for the Different condition, and weakest for the Half condition. Specifically, the hypothesis reflects the prediction of the inhibitory sharpening theory in this protocol that the use of the specially designed intermediate stimuli in the Half condition should result in a lower magnitude of repetition suppression compared to the Different condition.

Indeed, different theories of cortical function predict contrasting differences in the magnitude of repetition suppression in the three conditions, especially in the Half versus Different condition, or even a lack of differences in case cortical overlap was predicted not to affect repetition suppression. A comparison of the predictions made by three different theories, *fatigue*, *inhibitory sharpening*, and *lack of any overlap-dependent effects* is shown in Fig. 5.1. The fatigue theory based on neural habituation predicts a stronger amount of suppression for those neurons that are shared when presenting a stimulus that produces overlap in cortical activity in between successive repetitions of the adapter (i.e., Half condition), compared to the case in which overlap is absent (Different condition). On the opposite end, the inhibitory sharpening theory has produced specific

predictions in this context by means of computer simulations (Spigler and Wilson, 2017), predicting an opposite effect of *decrease* in the amount of suppression compared to a control in which the intervening stimulus does not produce significant amounts of overlap. A similar dynamics could also be derived in the context of predictive coding, as we briefly discussed in previous work (Spigler and Wilson, 2017), although specific simulations should be developed to make more accurate predictions in the context of the protocol used here.

5.2.4 Experimental Procedure

A graphical summary of the experimental procedure is shown in Fig. 5.2. Each of the three experimental sessions was split into 5 blocks in which a sequence of stimuli was shown on a monitor, alternated by a fixation pause lasting $20s$. Stimuli were either targets (a single inverted face) or were grouped into consecutive sequences of three faces, together corresponding to an experimental trial. Each stimulus was displayed for a fixed time of $900ms$ following a small cross at the center of the screen, leading the stimulus by a uniform random time between $400ms$ and $600ms$. The Inter-Stimulus Interval (ISI, i.e., the time between the disappearance of one stimulus and the onset of the next) was sampled uniformly at random between $2500ms$ and $2700ms$. This introduces a small jitter in the presentation of the stimuli and improves fitting of the haemodynamic response function. The participants were instructed to look at the fixation dot at the center of the screen, that was present in between presentations of the stimuli before changing to the cross that informed the participants of the imminent presentation of the next stimulus. The pause blocks were signalled by replacing the dot with a small empty circle. During each session, the participants were instructed to press a button whenever they saw a target, in order to provide an attentionally demanding task that was orthogonal to the experiment (as done in, e.g., (Henson, Shallice, and Dolan, 2000)).

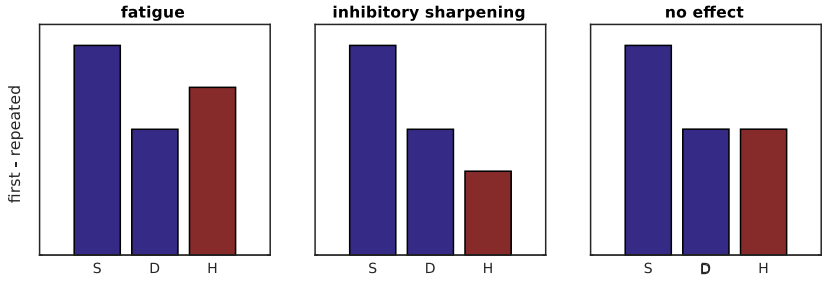


FIGURE 5.1: **Predicted magnitude of RS in the different conditions.** Different theories of cortical function predict contrasting differences in the magnitude of repetition suppression in the three conditions used in the protocol. For example, the fatigue theory predicts that the cortical units shared between the stimuli in the *Half* condition would undergo further adaptation during the presentation of the middle stimulus. This results in an increased magnitude of suppression in the Half condition compared to the Different condition. The inhibitory sharpening theory (Spigler and Wilson, 2017), on the contrary, predicts an opposite effect for which the shared cortical units become dis-inhibited and increase in activation, resulting in a weaker magnitude of suppression when the adapter is repeated. A third hypothesis is finally reported for which no effect due to cortical overlap is present. Note that the difference between the *Same* and *Different* conditions is well established in the literature, as two repetitions of an adapter are known to produce stronger repetition suppression than a single one (for example (Li, Miller, and Desimone, 1993; Sayres and Grill-Spector, 2006)).

The main part of the experiment consisted of three different trial types. Each trial was composed by a sequence of three faces, the last one being always a repetition of the first one (henceforth called the *adapter* stimulus). The trials differed in the composition of the middle face, which was either the same as the other two (*Same* trials), different (*Different* trials) or was generated by keeping either the top or bottom half of the adapter and replacing the complementary part with a unique one (*Half* trials). The three sessions combined resulted in 34 Same trials, 34 Different trials, and 38 Half trials, half of which were generated by keeping the top part fixed and half by keeping the bottom part fixed. 46 targets were inserted at random positions in the gaps between trials. The sequence of stimuli and the choice of the individual parts composing the stimuli were randomized between different sessions and subjects. Presentation of visual stimuli was controlled using Python and PsychoPy (Peirce, 2007).

5.2.5 Scanning Parameters

Participants were scanned in a $3T$ MRI scanner (Achieva 3T, Philips Healthcare, Best, NL) with a 32-channels head coil at the University of Sheffield. A high-resolution T_1-weighted 'structural' MRI was acquired for each participant at the beginning of the experiment using an MPRAGE (Magnetization Prepared Rapid Acquisition Gradient Echo) sequence (repetition time TR=3000ms, echo time TE=4.4ms, flip angle=8°, 256x256 matrix, field of view 240x240x170mm). $T_2^\star$-weighted Echo-Planar Imaging (EPI) was then performed using the BOLD contrast. During each scan, twenty-five axial slices were acquired in ascending order (TR=2000ms, TE=35ms, flip angle=90°, matrix size=128x128, field of view 230x230x93.75mm, in-plane voxel size 3x3mm with 3.75mm slice thickness and no gap). The $T_2^\star$ scans did not cover the whole brain, and excluded the inferior cerebellum and superior frontal/parietal cortices. The field of view was positioned on a midline T1 sagittal plane, centered on, and angled to, the anterior and posterior genu of the Corpus Callosum. The Shim box was positioned at

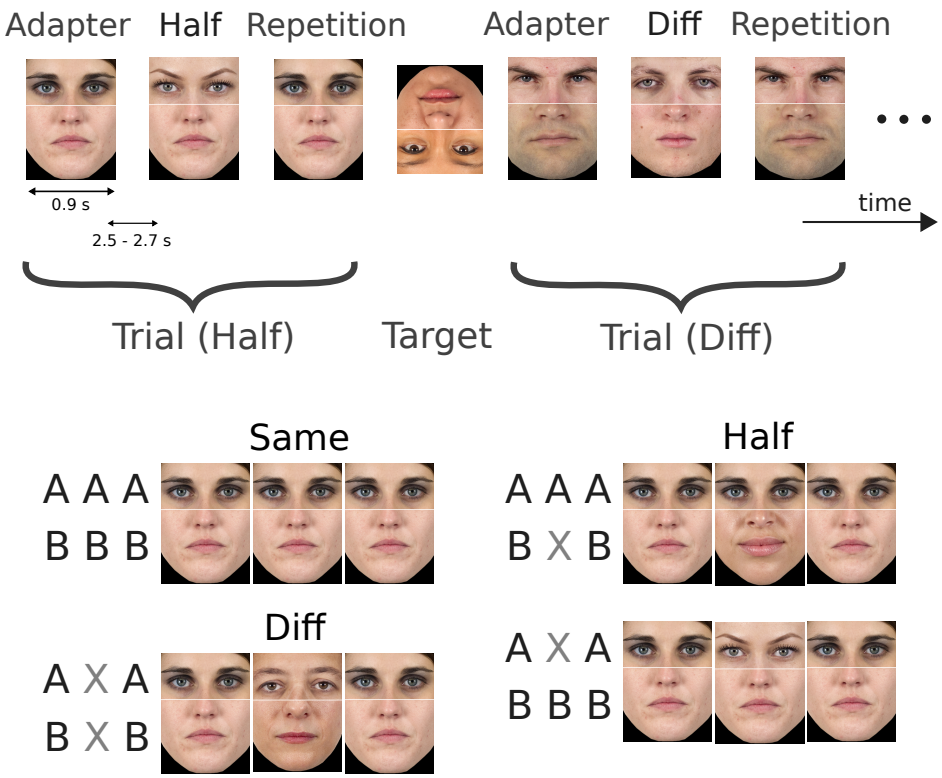


FIGURE 5.2: **fMRI experimental protocol.** Each scanning session consisted of 5 blocks consisting in a sequence of face stimuli, interleaved with a 20 seconds fixation baseline. The stimuli presented were either grouped in trials of three stimuli or were inverted faces presented in between trials as targets, that the participants were instructed to recognize by pressing a button. The three faces in each trial shared a similar structure, with the third stimulus being a repetition of the first one, the adapter. The trials were divided into three conditions, differing in the type of stimulus used for the second face in each triplet. In the *Same* trials, all the three faces were identical repetitions of the adapter, while in the *Different* trials, the middle stimulus was different from the other two. The middle stimulus in the *Half* trials was constructed by using either the same top or bottom halves of the adapter and matching it with a novel part.

a slightly increased angle to the slices, to avoid the frontal sinuses. 268, 256 and 251 volumes were acquired respectively in 3 runs, excluding three initial volumes that were discarded to allow saturation of T1 effects. The number of volumes acquired in the three sessions was 268, 256 and 251.

5.2.6 Region of Interest Analysis

Region of Interest (ROI) analysis was performed using the MarsBaR toolbox (Brett et al., 2002). Unbiased masks for left and right "Occipital Face Area" (OFA) and "Fusiform Face Area" (FFA) were computed using a separate, previously published open access dataset (Wakeman and Henson, 2015) with a group-level *Face* > *Scrambled Face* contrast (Family-Wise-Error-corrected, threshold at $p = 10^{-5}$, minimum 20 voxels per cluster). The peak coordinates and size of the resulting clusters are reported in Table 5.1.

Region	Size (# voxels)	Peak (mm)
Right OFA	113	39 −82 −11
Right FFA	133	39 −46 −17
Left OFA	45	−36 −85 −14
Left FFA	63	−39 −49 −20

TABLE 5.1: **Masks used for the region of interest (ROI) analysis.**

5.2.7 Data Analysis

Analysis of the data was performed using the SPM12 software (SPM12, Wellcome Trust Centre for Neuroimaging, `http://www.fil.ion.ucl.ac.uk/spm/software/spm12`) and custom MATLAB scripts. The data was pre-processed to correct for head movement and slice timing (synchronized to the middle slice, slice 13). For each subject, the T_1-weighted structural image was segmented, co-registered to the mean BOLD image, and normalized to the MNI template. All the resulting BOLD scans were re-sampled to $3mm$ isotropic voxels and smoothed using a 3D Gaussian kernel with FWHM=$8mm$.

The time series of the BOLD scans were high-pass filtered at a cutoff frequency of $\frac{1}{128}s^{-1}$ and an AR(1) auto-regression model was used to account for temporal correlations in the data. A General Linear Model (GLM) was fitted for each subject using the canonical haemodynamic response function with 10 regressors: first presentation of faces, middle stimuli and repetition of the adapters, separately for each of the three conditions, and targets. Noise covariates were estimated using the GLMDenoise toolkit (Kay et al., 2013) and added to the GLM, which was then solved using Ordinary Least Squares. Group-level analysis was conducted using a random-effects model with the summary statistics computed from each subject.

5.3 Results

A second-level random effects analysis showed statistically significant suppression of the BOLD signal with repetition in the bilateral occipital visual areas, including the Occipital Face Area (OFA), and in the Fusiform gyri, including the Fusiform Face Area (FFA). Figure 5.3 shows the thresholded statistical parametric map (SPM) for the difference between the first presentations of the adapter stimuli and their repetition across all conditions. For this analysis the threshold was set at $p = 0.01$ without correction for multiple comparisons and a minimum cluster size was set at 20 voxels. The statistical parametric map was restricted to the voxels in which the positive effect of the first presentation of the adapters across conditions was higher than a threshold of $p = 10^{-5}$. The thresholded SPM of the positive effect of the first presentation of faces is reported in Appendix B as Fig. B.1. The estimated haemodynamic response function was next computed as a peri-stimulus time histogram (PSTH) of the BOLD response to first and second faces in each trial (immediate repetition) in the *Same* condition, averaged across voxels in the right OFA and FFA areas respectively. The estimated haemodynamic response functions match the previous literature and

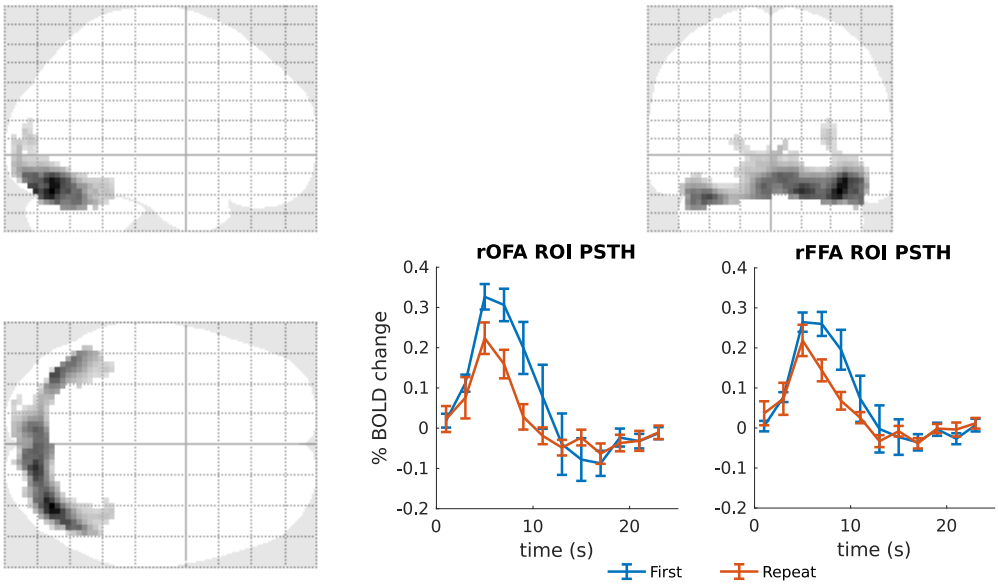


FIGURE 5.3: **Repetition suppression: single-voxel analysis and ROI PSTHs.** Second-level thresholded statistical parametric map showing repetition suppression across all conditions (difference between the first face and its repetition, no family-wise error correction, threshold at $p = 0.01$ and minimum of 20 voxels, masked by the positive effect of the first presentation of faces across conditions with a threshold of $p = 0.00001$, as shown in Appendix B as Fig. B.1). **Lower-right**: peri-stimulus time histogram (PSTH) of the BOLD response in the right OFA and FFA areas to the first presentation and consecutive repetition of the same stimulus, averaged across all the participants. Bars show the standard error.

show repetition-dependent suppression.

As the first presentation of the adapters in the different conditions is always a novel face carrying no condition-specific information, no significant differences should be expected between their regressors in each condition, although noise may be present due to different attentional effects and other perturbations in each trial. This was confirmed by a one-way ANOVA in each area (lFFA $F(2, 36) = 0.04$, $p = 0.96$; lOFA $F(2, 36) = 0.16$, $p = 0.85$; rFFA $F(2, 36) = 0.11$, $p = 0.90$; rOFA $F(2, 36) = 0.23$, $p = 0.80$). A bar plot of the regressors compared in the ANOVA is included in Appendix B as Fig. B.2.

To perform the analysis of the condition-specific differences in activation, we then computed the magnitude of suppression for each stimulus in each condition (Same, Different and Half) by subtracting the beta value of each regressor (second and third stimuli in each trial) from its corresponding first presentation (e.g., $RS_{repeated,same} = \beta_{first,same} - \beta_{repeated,same}$). Figure 5.4 shows the computed values, averaged across all the participants. The original beta values for each regressor are included in Appendix B as Fig. B.3.

The main analysis was conducted on the magnitude of suppression of the adapter on its repetition (third stimulus in each trial) across the different conditions. This corresponds to the contrasts in Figure 5.4 (left). A one-way ANOVA between the computed magnitudes of repetition suppression found no statistically significant differences between the conditions in areas lFFA ($F(2, 38) = 0.22$, $p = 0.8$), lOFA ($F(2, 38) = 0.79$, $p = 0.46$) and rFFA ($F(2, 38) = 0.28$, $p = 0.76$). Differences in the magnitude of repetition suppression between conditions in the right OFA were found to be near borderline significance with $F(2, 38) = 3.21$, $p = 0.052$. A Fisher's Least Significant Differences post hoc test further revealed a significant difference between the *Same* and *Half* condition ($p = 0.02$) in the right OFA. While the difference between the *Different* and *Half* conditions was not found to be significant ($p = 0.36$), a trend in the non-significant differences seems consistent with the inhibitory sharpening theory,

for which repetition suppression in the *Half* condition should be weaker than in the *Same* and *Different* conditions. We observe that a similar trend is also consistently present in all the other areas, albeit with a high variance. It is thus possible that using a larger number of participants may reduce the noise and reveal the statistical difference corresponding to this trend. The post hoc test further revealed significant effects of repetition for each condition in each area (lFFA: *Same* $p < 0.01$, *Different* $p < 0.05$, *Half* $p < 0.05$; rFFA: *Same* $p < 0.01$, *Different* $p < 0.05$, *Half* $p < 0.05$; rOFA: *Same* $p < 0.001$, *Different* $p < 0.001$, *Half* $p < 0.05$), except for the left OFA (*Same* $p = 0.08$, *Different* $p = 0.5$, *Half* $p = 0.95$).

Further analysis focused on the potential differences in the trials of the *Half* condition depending on whether the part that was replaced was the top or the bottom part. Applying a paired t-test to the magnitude of repetition suppression did not reveal any significant differences between the two subsets of trials in the condition in any area, regardless of position within the trial triplet (middle stimulus in lFFA $t(24) = -0.83, p = 0.42$, rFFA $t(24) = -0.29, p = 0.78$, lOFA $t(24) = 0.72, p = 0.48$, rOFA $t(24) = 0.15, p = 0.88$ and repeated stimulus in lFFA $t(24) = -0.46, p = 0.65$, rFFA $t(24) = 0.58, p = 0.56$, lOFA $t(24) = -0.69, p = 0.50$, rOFA $t(24) = -0.50, p = 0.62$).

It is also of interest to look at the differences between magnitudes of suppression for the stimuli in the middle position of the triplets. The value of the regressors, subtracted from the corresponding first presentation of the adapter stimuli, is shown in Fig. 5.4 (right). A one-way ANOVA between the computed magnitudes of suppression found no statistically significant differences between the conditions in any area (lFFA $F(2, 38) = 1.21$, $p = 0.31$, lOFA $F(2, 38) = 0.6$, $p = 0.56$, rFFA ($F(2, 38) = 1.49$, $p = 0.24$)), except for the right OFA that was near the borderline of significance ($F(2, 38) = 2.7, p = 0.08$). A Fisher's Least Significant Differences post hoc test further revealed a significant difference between the *Same* and *Different* conditions in the right OFA area ($p < 0.05$). The

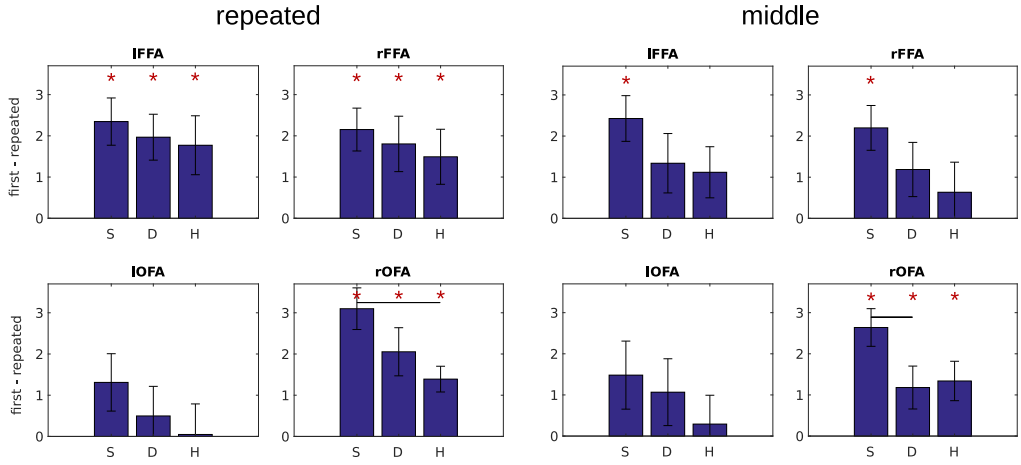


FIGURE 5.4: **Repetition suppression in the different conditions.** Magnitude of repetition suppression for each condition, measured as the difference between the beta values of first presentation and repetition (left) and the middle stimulus (right) separately for each condition (*Same*, *Different* and *Half*, averaged across the mean values for each participant. Vertical lines show the standard error, stars denote statistically significant repetition suppression, and horizontal lines denote significant differences in the magnitude of suppression. Despite the lack of statistically significant differences, interesting trends seem to be present in a way compatible with the inhibitory sharpening theory (Spigler and Wilson, 2017).

difference between the *Same* and *Half* conditions was found to be near the borderline of significance ($p = 0.07$), and it is thus possible that the lack significance can be attributed to the small number of participants compared to the amount of noise in the data. A similar difference between the *Same* and *Half* conditions was also found to be near significance in the right FFA area ($p = 0.086$). Finally, a one-sample t-test for each condition determined a significant suppression of activity in the *Same* condition in all areas (lFFA $t(12) = 4.37$, $p < 0.001$; rFFA $t(12) = 4.04$, $p < 0.01$; rOFA $t(12) = 5.77$, $p < 0.001$) except for the left OFA ($t(12) = 1.79$, $p = 0.1$). Interestingly, the right OFA also showed significant suppression in both the *Different* and *Half* conditions (Different, $t(12) = 2.26$, $p < 0.05$; Half, $t(12) = 2.79$, $p < 0.05$).

5.4 Discussion

Sensory stimuli have been shown to produce broad patterns of distributed activity in the corresponding areas of mammalian neocortex (Tsunoda et al., 2001; Pasupathy and Connor, 2002; Pouget, Dayan, and Zemel, 2000; Connor, 2005; Georgopoulos, Schwartz, and Kettner, 1986). The single presentation of a stimulus has been found to be sufficient to produce short-term and even long-term changes in the responses of the selective neurons active during its presentation. Here was investigated the effects of short-term to medium-term synaptic changes due to a sequence of stimuli designed to produce overlap in their cortical response patterns, that is, to stimuli thought to activate a shared subset of neurons. In this study, repetition suppression to novel face stimuli was used to measure the effect, as it is known to produce short-term plasticity or adaptation effects (Henson, 2015) in face-selective cortical areas of monkeys and humans. The face stimuli were designed using the Composite Face Effect (Young, Hellawell, and Hay, 1987), giving us freedom to compose new faces that produce a holistic perception but allow for the composition of different parts or

features.

To test the effect of cortical overlap on the strength of repetition suppression we adopted a protocol in which each trial consists of a sequence of three faces, an adapter repeated in the first and last position, and a middle face whose composition was varied between conditions. The middle face could be identical to the adapter (*Same* condition), different and thus not generally sharing parts or features with the adapter (*Different* condition), or composed by keeping either the same top or bottom half of the adapter and matching it with a complementary part taken from a separate set of new faces (*Half* condition), so that the evoked pattern of activity might show a higher degree of overlap with that of the adapter than in the Different condition.

Repetition suppression to novel faces was significant in the bilateral OFA and FFA areas, as found in previous studies (Henson, 2015), except for the left OFA in which it was only present as a trend, possibly due to its small effect size. Condition specific differences observed in the magnitude of repetition suppression (as approximated by the difference in the value of the regressors of the first minus the third stimulus in each triplet) only reached statistical significance in the right OFA, where the *Half* condition produced significantly weaker suppression than the *Same* condition (see Figure 5.4). It is interesting to note that all the cortical areas, and in particular the right OFA, showed a clear trend in agreement with the predictions of the inhibitory sharpening theory (compare Figure 5.1 with Figure 5.4 (left)), which predicts the strongest suppression in the *Same* trials, intermediate suppression in the *Different* trials and the weakest suppression in the *Half* trials. This trend was not statistically significant, so strong conclusions cannot be drawn. However, the amount of variance was found to be high compared to the differences between the three conditions, so the lack of statistical significance may be due to a small sample size. Future experiments with more participants are required to draw stronger conclusions. In any case, the significant difference in the *Same* and *Half* conditions could decrease the

likelihood that the interaction due to cortical overlap was caused by neural fatigue, as in that case the magnitude of suppression in the Half condition would be predicted to be larger than in the Different condition. We also note that previous studies (for example, Li, Miller, and Desimone; Sayres and Grill-Spector, 1993; 2006) showed that the difference between the *Same* and *Different* conditions should be significant, as the *Same* condition features two repetitions of the same stimulus versus the single repetition in the Different condition. The lack of statistical significance for this difference in any area may thus further support our conclusion that more data is required to reveal the true statistical differences between the different conditions. A separate test found no difference within the *Half* condition, between trials that retained the same top or the same bottom parts.

Further inspection of the difference in the BOLD signal between the first and middle faces in each trial found a significant difference between the *Same* and *Different* conditions in the right OFA. This was expected, as the *Same* condition corresponds to an immediate repetition of the adapter in this case, while the *Different* condition would not be expected to produce any repetition effects. It is interesting to note however that the middle face in both the *Different* and *Half* conditions did produce statistically significant suppression of activity in the right OFA and a strong but non-significant trend in all the others, which may be due to some degree of adaptation of similar facial features even when different faces are used. This was observed in previous studies that exploited repetition suppression to faces constructed with the Composite Face Effect to investigate holistic versus parts-based encoding of faces in the neocortex (Schiltz and Rossion, 2006; Schiltz et al., 2010). This result has implications for the design of standard repetition suppression protocols, as mild suppression may be present due to the sequential presentation of faces, even if they are not repeated.

The choice of stimuli to use in the experiment in order to produce overlap in their cortical activations in a controlled way is, however, not trivial. The choice

of the Composite Face Effect is motivated by the idea of using different objects that share common features or sub-parts, exploiting a high-level population coding of objects as a combination of parts and features as found in primate area V4 (Pasupathy and Connor, 2002), IT (Tsunoda et al., 2001) and OFA (Schiltz and Rossion, 2006). Other ways to achieve overlap in the cortical representation of sensory stimuli could be to use small changes in properties of the stimuli (e.g., direction of motion of moving dots), or more generally via image morphing algorithms. This might be especially useful in early sensory cortices where population codes based on a weighted average of the tuning of the activated neurons is usually found for some encoded variables (e.g., direction of motion or orientation of lines) (Georgopoulos, Schwartz, and Kettner, 1986; Pouget, Dayan, and Zemel, 2000; Purushothaman and Bradley, 2005; Bednar and Miikkulainen, 2000; Spigler, 2014).

In conclusion, this study has demonstrated how the effects of interference in neural dynamics, such as repetition suppression, and higher level cognitive processes, such as perception and memory, can be investigated using stimuli that are thought to produce overlap in their patterns of cortical activity. While the results of the functional neuroimaging experiment presented here were not found to provide conclusive evidence for the inhibitory sharpening theory (Spigler and Wilson, 2017), possibly due to small effect sizes and a need for a larger amount of data, interesting trends were observed. These new data, in the context of the various neuroimaging studies surveyed in this discussion, suggest that further investigation may allow the different predictions of neural fatigue, sharpening and inhibitory sharpening theories to be tested.

Chapter 6

Discussion and conclusion

6.1 General discussion

The principal aim of this thesis was to work towards a complete theory of how the human brain builds and maintains representations of the world. Specifically, the focus of this work was on the role of lateral cortical interactions in shaping cortical responses to perceptual stimuli on a timescale between the short-term dynamics associated with perception and the longer-term dynamics associated with the longer-term development of cortical maps.

Lateral connectivity between cortical minicolumns is a main feature in the mammalian neocortex. Lateral interactions mediated by this dynamic pattern of connectivity subserve important functions in cortical processing, as we have reviewed in Chapters 1 and 2. Further, previous results (Bednar and Miikkulainen, 2000; Spigler, 2014) suggest that important aspects of the cortical representation of sensory stimuli in the cortex are encoded in the synaptic strengths of such lateral interactions. In this thesis, we explored whether a neurobiologically plausible computational model of cortical self-organisation could be used to investigate how synaptic plasticity and adaptation in lateral cortical interactions modifies the structure of pre-existing cortical representations and how it affects their decoding.

To this end, we started by using the L-model (Stevens et al., 2013), building upon an established program of existing computational neuroscience research. We first designed and run a protocol for computer simulations using the L-model to investigate the role of plastic changes in lateral interactions in the cortical representation of stimuli, and we decided to focus on the medium-term dynamics of the phenomenon of repetition suppression as a measure of their effect. This investigation resulted in the 'inhibitory sharpening' theory that is capable of explaining dynamics of repetition suppression by an increase in the strength of the inhibitory interactions between cortical units co-active during the presentation of the same stimulus due to Hebbian learning. The theory was then used to produce the novel prediction that repetition suppression for an adapter object can be disrupted (i.e., the magnitude of suppression can be reduced) by intervening exposure to objects that produce activity that overlaps with that elicited by the adapter, its cortical representation.

While the L-model was originally developed to model experimental data from early sensory areas, it is here applied to higher cortical areas to model complex cortical representations of sensory stimuli. The choice is justified by similarities in their neural circuits (e.g., the presence of intra-area lateral inhibition) and representation of features (e.g., cortical maps, population coding, etc...). However, while the main assumptions of the L-model may hold for higher sensory cortical areas, further experimental evidence will be required to properly verify the validity of the hypothesis.

It was next decided to test the predictions of the inhibitory sharpening theory by taking two approaches, a neuroimaging experiment to measure the actual magnitude of repetition suppression in a protocol compatible with that used in the simulations, and a behavioural experiment to achieve both a separate test of the same predictions and to explore whether the predicted interference between stimuli that are hypothesized to produce overlapping patterns of cortical activity can affect perception and behavior.

Specifically, the behavioural experiment was designed to investigate how the perception of objects can be affected by the interference due to stimuli whose cortical representation is suspected to be overlapping with theirs, and at the same time to further test the predictions of the inhibitory sharpening theory in the cognitive domain. The results showed a significant effect of partial forgetting of previously learnt stimuli in such a protocol, measured as a significant decrease in their recognition accuracy and an increase in the reaction time to recognize them. The results from the control group further showed that if the intervening stimuli used are not designed to produce disruption, human participants are capable of learning to recognize the full set of stimuli used in the experiments, showing that the decrease in accuracy that was found was not due to an increase in the complexity of the task.

Finally, the neuroimaging experiment was designed and run to test the predictions of the inhibitory sharpening theory directly by measuring the magnitude of repetition suppression in a protocol inspired to the one used in the computer simulations, with the objective of measuring the possible predicted *decrease* in the magnitude of suppression due to disruptive stimuli. The experiment used a sequence of faces grouped in triplets where the same adapter stimulus was repeated in the first and third position and the composition of the middle stimulus determined the three trial conditions (Same, Different and Half). The magnitude of repetition suppression to the adapter stimulus would then be different in the three conditions as predicted by different theories. Figure 5.1 was used to highlight three possible outcomes. According to the *fatigue* theory of repetition suppression, the magnitude of suppression in the Half condition is predicted to be higher than in the Different condition, as the neurons shared in the representation of the adapter and the disruptive stimuli undergo continued adaptation throughout the trial. On the contrary, the *inhibitory sharpening* theory predicts a decrease in suppression as per the dynamics discussed in Chapter 3. Finally, it could have been possible that no effect of interaction existed, and that

thus the magnitude of suppression would be identical in the Different and Half conditions. It is still to be noted, however, that each of the three outcomes is compatible with a number of theories, and the results of the experiment would thus only provide a coarse discrimination between different sets of competing theories. All the cortical areas that were analysed (bilateral FFA and OFA areas) showed a clear trend in agreement with the predictions of the inhibitory sharpening theory, although the trend did not reach statistical significance in any area. However, it may be possible that the effect size is small compared to the variance we found in the data, so the lack of statistical significance may be due to a too limited amount of data. To further support this conclusion, we note that previous studies (Li, Miller, and Desimone, 1993; Sayres and Grill-Spector, 2006) showed that the difference between the Same and Different conditions should be significant, although small in magnitude, as the Same condition corresponds to two repetitions of the adapter versus a single repetition in the Different condition. This difference was however not found to be statistically significant in the experiment presented here, suggesting that significant effects may still be hidden by the large variance in the data due to a too limited number of participants in the study. Future experiments with a larger number of participants will be required to draw stronger conclusions. In any case, the trend that was found in the data seems to disagree with the fatigue theory, that predicts opposite dynamics.

In the next sections of this chapter we will discuss some general themes of interest related to the present work.

6.2 Parts-based population coding

As we reviewed in Chapter 2, population coding can be used to represent sensory stimuli or perceptual variables (e.g., orientation of a line or colour of a patch) as a distributed pattern of activity across multiple neurons. The population codes that we explored in this thesis can be decoded as a linear combination

of the preferred features of the active neurons, weighted by their activations. For example, population activity in the cat motor cortex has been found to represent the 3D location of the paw as a vector sum of the paw locations preferred by the individual active neurons (Ethier et al., 2006). A similar type of distributed representation has been found to encode complex shapes in primate visual area V4 (Pasupathy and Connor, 2002), and to encode objects as a combination of simpler features and smaller parts in primate temporal cortex (Wachsmuth, Oram, and Perrett, 1994; Tsunoda et al., 2001).

In Chapters 2 and 3 we discussed how the L-model can support both distributed and localist representations, to different degrees depending on the strength of the inhibitory connections between the model units. Indeed, different balances between afferent and excitatory input on one hand, and lateral inhibition on the other, change the equilibrium to which the network settles after exposure to a new stimulus. In this thesis we have shown how the L-model in particular can produce an approximate parts-based code of stimuli as in primate area V4 (Pasupathy and Connor, 2002) and inferotemporal cortex (Wachsmuth, Oram, and Perrett, 1994; Tsunoda et al., 2001), while previous work has shown how its intrinsic population code could be decoded to read out perceptual information such as the perceived orientation of gratings (Bednar and Miikkulainen, 2000) and perceived colour (Spigler, 2014).

Another interesting way to produce parts-based representations that was explored early in the stages of this work is to compute a Non-Negative Least Squares (NNLS) fit of the input stimulus as a linear combination of the afferent weights of the cortical units in the network weighted by non-negative mixing coefficients. The weights of the networks could then be updated by Hebbian Learning. It is interesting that this approach still relies on implicit competition between the model units, as the optimal combination of units is computed out of the many possible ones. A more complete description of the preliminary experiments using NNLS are reported in Appendix C. Future work should investigate

the relation between the two approaches, and how the L-model or other models of cortical dynamics may approximate aspects of the NNLS optimization problem.

6.3 Bridging mechanistic models of cortical dynamics with high-level cognition

An interesting result of the work in this thesis is the connection between low-level modeling of cortical dynamics and behaviour that depends on high-level cognition such as perception and memory. The link is made possible by the use of patterns of cortical activity as an intermediate stage connecting the two domains. In particular, different types of cortical models may be used to predict the organization of distributed representations of stimuli based on a population code hypothesized to be used in the specific cortical area of interest. This was explored in Chapters 2 and 3, setting the bases by reviewing and modifying the L-model of cortical self-organization which is capable of producing a mixture of distributed and localist cortical representations similar to those observed in the mammalian sensory cortices. Chapter 4 then linked an hypothesized similar parts-based population code to perception and behaviour.

As was done in Chapter 4, the approach explored here can be used in complex protocols by further predicting changes in the cortical representation of stimuli due to neural plasticity and adaptation. This was explored in this thesis to explain perceptual changes due to interference from stimuli that produced overlapping patterns of activation. Previous work, however, had already taken advantage of similar ideas to explain the changes in perception observed in the tilt aftereffect (Bednar and Miikkulainen, 2000) and the McCollough effect (Spigler, 2014). In both cases of this thesis and previous work, the cortical representations produced by the model are used as an intermediate bridge between cortical dynamics and cognition.

Finally, a similar approach was recently used to explain the dynamics underlying visual abnormalities in schizophrenia (Silverstein, Demmin, and Bednar, 2017), thus bridging the low-level dynamics produced by a mechanistic model of cortical function with cognitive phenomena. However, contrary to the the work presented in this thesis, the recent work explored the level of efferent activity in parts of the models and a change in the model parameters to explain the differences between healthy and clinical data, rather than investigating the changes in the generated representation patterns.

6.4 Effect of cortical overlap in cognition

In this thesis we have investigated the effect of cortical overlap between different stimuli both directly with a neuroimaging experiment (Chapter 5) and indirectly in a set of behavioural experiments (Chapter 4). Here we argue that this idea can be further explored in relation to previous studies in the literature that have produced results that are compatible with the account of cortical overlap given in this thesis, both in neuroimaging experiments and in cognitive studies of perception and memory.

A previous study by Webster and colleagues (Webster et al., 2004) showed that watching a sequence of faces can strongly affect subsequent perceptual judgements on new faces, dynamically changing the boundary between categories such as perceived gender, ethnicity and facial expressions. For example, watching a sequence of male faces was found to shift the perceptual boundary on gender such that a subsequent test neutral face was more likely to be perceived as being female. This effect may be explained in the context of the present study as adaptation of neurons selective to features and parts of faces that thus modify the population decoded perceptual judgement on subsequent faces that rely on such shared features and thus shared neurons in the same cortical representation.

The potential interference between overlapping cortical patterns may also have been observed in working memory. Cohen and colleagues (Cohen et al., 2014) presented a theory based on experimental data for which the processing capacity of multiple visual objects decreases with the amount of overlap between their neural response patterns, although differently from here the degree of overlap was measured directly from the fMRI signal and considered across multiple visual cortical areas rather than within a single area. This result may be further linked to a set of studies that showed that visual masking is more effective when the category of the mask matches the one of the masked stimulus (e.g., masking faces with faces rather than random noise) (Aguado, Serrano Pedraza, and García Gutiérrez, 2014), which is more likely to activate shared sub-sets of neurons, thus producing a higher degree of interference. The effect was then correlated to direct measurements of overlap in the BOLD fMRI signal using a larger number of visual categories (Cohen et al., 2015). A similar type of interference in visual masking was also indirectly explored in the "neural competition" theory by Keysers (Keysers and Perrett, 2002), which is based on a competition between co-existing neural representations within the same cortical areas. Competition between patterns of neural activity corresponding to different stimuli may be also implicated in adaptive forgetting due to competition between memory traces (Wimber et al., 2015).

The effect of interference due to overlapping distributed representations was also proposed to address the plasticity/stability dilemma in artificial and biological neural networks, and to explain the problem of catastrophic forgetting in connectionist models (French, 1992). It is interesting that the results from Chapter 4 match previous data in this context very well.

Finally, the results presented in Chapter 4 can be used to potentially link a mechanistic description of neural dynamics with confusion between perceptually similar stimuli, by assuming a linear decoding of the patterns of cortical activity by downstream neurons (for example, (Pitkow et al., 2015)). In this regard,

it is interesting that the results from Chapter 3 predicted a tendency for the cortical representation of stimuli to change from an initial broadly distributed representation to a more localist one during the process of familiarization, which has the effect of reducing the overlap between pairs of stimuli and thus makes them more easy to discriminate with linear template matching. This phenomenon could thus prove particularly useful for cortical computation.

6.5 How can plasticity of lateral interactions affect cortical representation?

In conclusion, this thesis investigated the role of plasticity of lateral interactions in shaping cortical responses to perceptual stimuli on a timescale between the short-term dynamics associated with perception and the longer-term dynamics associated with cortical self-organisation. In particular, we have shown how the plasticity-dependent changes in the cortical representation of sensory stimuli is compatible with the observed phenomenon of repetition suppression and to behaviour.

The theory developed in this thesis along with the results from a set of experiments, both behavioural and by means of neuroimaging, was finally used to perform a preliminary investigation of the effect of interference due to stimuli that are designed to produce overlapping cortical representations, that is to activate a large number of shared neurons.

6.6 Future Work

As was discussed in Chapter 5, it is possible that the lack of statistical significance found in the neuroimaging experiment was due to a small effect size that was could not be captured with the number of participants that was recruited

(13 volunteers). Future work should try to extend the experiment or use similar designs in order to further test the inhibitory sharpening theory and better isolate the specific contribution of plasticity in lateral cortical interactions.

Also, future work is required to isolate the specific contribution of lateral inhibitory plasticity on repetition suppression and on the cortical representation of stimuli. Single-neuron studies may be necessary. Optical imaging experiments may prove useul as well; for example, a setup similar to that used by Tsunoda et al., (2001) may be used to test new predictions from the L-model on the familiarization-induced dynamic changes to the activity blobs associated to the presentation of sensory stimuli.

We expect that studying the effect of overlapping cortical representations, for example using Multi-Voxel Pattern Analysis (MVPA) and machine learning, will prove fruitful to probe cortical dynamics and to bridge them with high-level cognition and behaviour.

At the same time, it will be useful to further develop the models of cortical development presented in this thesis in order to generate and test novel predictions about perception. For example, it would be interesting to integrate the L-model with to support some degrees of supervised learning, which would allow a direct modelling of the data from Chapter 4.

Indeed, we suggest that this approach will help bridge the gap in our understanding of how we represent the world, between psychological descriptions of perception and neurobiological mechanisms of cortical dynamics on developmental and perceptual timescales.

Appendix A

Parameters of the simulations

The table below shows the parameters used for the simulations in Chapter 3. The learning rates represent the total amount of change across all the model units. The learning rate of each unit is computed as $\epsilon = \frac{\epsilon_p}{N_{units}}$, where N_{units} is the number of units in the network ($N_{units} = 48 \cdot 48$).

Parameter	Value
# Units	$48 \cdot 48$
Afferent Strength (α_A)	2.2
Excitatory Strength (α_E)	1.2
Inhibitory Strength (α_I)	2.3
Afferent Learning Rate (ϵ_A)	0.1
Excitatory Learning Rate (ϵ_E)	0.0
Inhibitory Learning Rate (ϵ_I)	0.3

Parameters used to pre-train the models.

Parameter	Value
Afferent Strength (α_A)	1.5
Excitatory Strength (α_E)	1.2
Inhibitory Strength (α_I)	2.3
Homeostatic learning rate (λ)	0.01
Homeostatic smoothing (β)	0.991
Homeostatic initial average ($\overline{\eta_j}(0)$)	0.15
Homeostatic target activity (μ)	0.024

Appendix B

Supplementary Results for the neuroimaging investigation

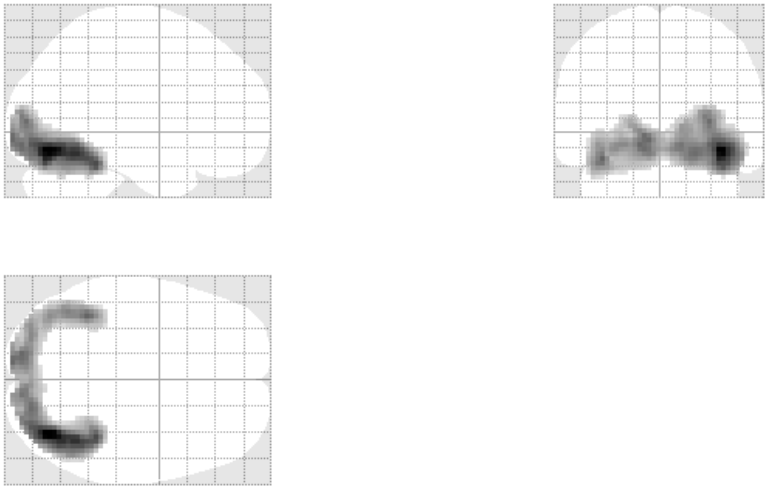


FIGURE B.1: **Single voxel analysis: face localizer.** The figure shows single-voxel results of the main effect of first presentations of faces across all the conditions at a threshold of $p < 0.001$ with Family-Wise Error (FWE) correction and clusters larger than 20 voxels.

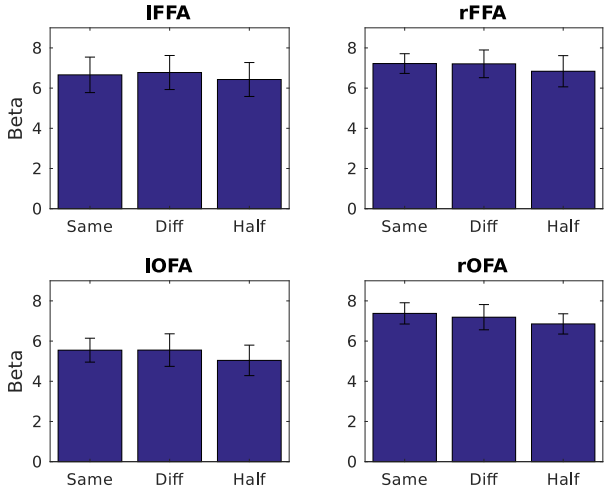


FIGURE B.2: **Beta values for each first face**. Beta values for each first presentation of faces in each condition and ROI, averaged across all the participants. Bars show the standard error.

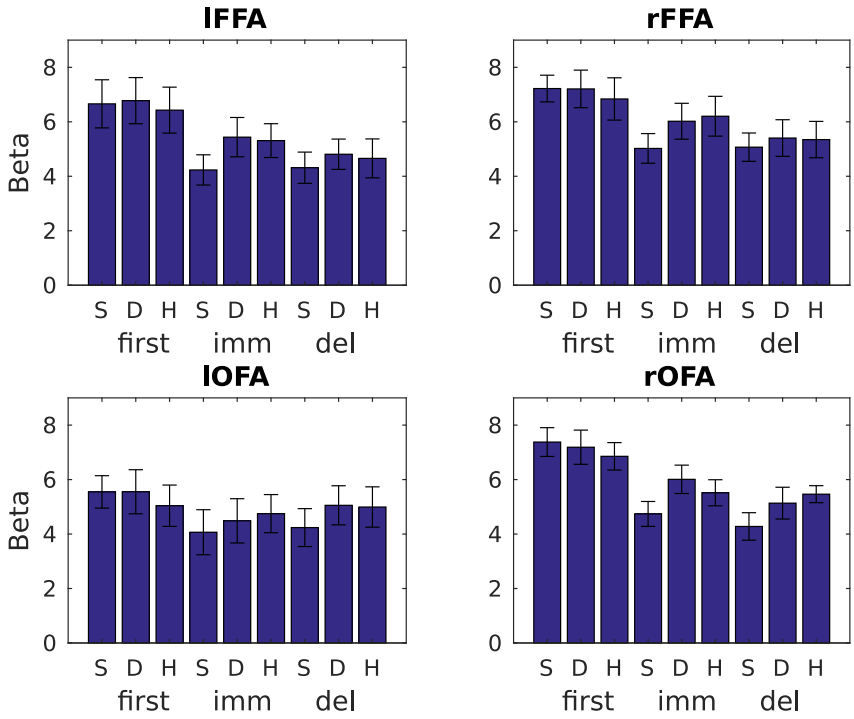


FIGURE B.3: **Beta values for each regressor and ROI**. Beta values for each regressor (Condition x Repetition; conditions being Same, Different and Half, and repetitions being First, Immediate and Delayed) in each ROI, averaged across all the participants. Bars show the standard error.

Appendix C

Preliminary experiments with Non-Negative Least Squares (NNLS)

The simulations presented in this thesis involved the high-level encoding of sensory stimuli as a parts-based population code compatible with that suggest to be used in primate area V4 (Pasupathy and Connor, 2002) and inferotemporal cortex (Wachsmuth, Oram, and Perrett, 1994; Tsunoda et al., 2001). In particular, the simulations presented in Chapter 4 used a fixed high-level encoding of stimuli as a collection of parts, specifically as binary vectors $\mathbf{x} \in \mathbb{R}^D$, representing the presence or absence of D possible object part that could be observed. For example, this encoding could represent an object A with the first 3 parts present out of 5 possible ones as

$$\mathbf{x_A} = (1, 1, 1, 0, 0)^T$$

Part of the work of this thesis showed that repetition suppression dynamics can reflect a re-organisation of cortical activation from a initially distributed representations to more localist ones, during the process of familiarization (see Chapter 3). Preliminary work, however, used a mathematically more accurate method for estimating the activation of the model units, at the expense of a

lower biological plausibility and higher computational cost. In this regard, the L-model may be seen as approximating a similar distributed population code in a biologically realistic way.

The method described here relies on the Non-Negative Least Squares optimization (NNLS) (Chen and Plemmons, 2010) to compute the target representation of input stimuli as a vector of non-negative mixing coefficients such that the input stimulus can be reconstructed as a linear combination of the template weights of the cortical units in the model, weighted by their corresponding activation.

We next present a simple model of a high-level sensory cortex based on these ideas, and we show how it can produce dynamics of repetition suppression dependent only on the specific type of neural coding used, together with simple Hebbian learning. The model is based on a Kohonen Self-Organising Map with N units. The weights of the map's units, that is their projective fields or templates, can be represented as D-dimensional columns $\mathbf{w_i}$ of a matrix $C \in \mathbb{R}^{D \times N}$.

However, contrary to regular Kohonen SOMs, the model activation $\boldsymbol{\alpha}(t) \in \mathbb{R}^N$ in response to each new stimulus $\mathbf{x(t)}$ is produced by computing a population-coded representation of the stimulus as a non-negative linear combination of the templates of the units in the model, solving the Non-Negative Least Squares optimization

$$\tilde{\boldsymbol{\alpha}}(t) = \underset{\boldsymbol{\alpha} \geq 0}{\operatorname{argmin}} \|C\boldsymbol{\alpha} - \mathbf{x}(t)\|_2$$

where $\boldsymbol{\alpha} \geq 0$ constrains each component of $\boldsymbol{\alpha}$ to be non-negative, and $\|\cdot\|_2$ denotes the Euclidean norm.

After the response to a stimulus has been computed, the weights are updated according to the traditional Kohonen SOM learning rule, by picking the most active unit $a = \operatorname{argmax} \tilde{\boldsymbol{\alpha}}$ and updating all the model units in its neighborhood as

$$\mathbf{w_i}(t+1) = \mathbf{w_i}(t) + h(a,i,t) \cdot \eta(t) \cdot (\mathbf{x}(t) - \mathbf{w_i}(t))$$

where $\eta(t)$ is the learning rate and $h(a,i,t)$ is a Gaussian neighborhood function that depends on the distance between units a and i on the 2D model map.

We finally measure the model dynamics, and in particular the transition from population-coded distributed representation to localist ones tracking stimulus familiarization (as in Chapter 3), by computing the sum of the activity of all the model units

$$\text{TotalActivation} = \sum_i \alpha_i$$

Figure C.1 shows the amount of total activation in the model following the repeated presentation of a single adapter stimulus, in a protocol compatible to that used in Chapter 3. The results reported were computed using a SOM map with N = $20 \times 20 = 400$ units and input objects composed of $D = 5$ possible parts, and an adapter stimulus $\mathbf{x}_{\text{adapter}} = (1,0,1,0,1)$. The weights of the model units were initialized to random values (uniform in $[0,1]$).

The results show a reduction in the total amount of activation in the model due to the strongest units becoming strongly tuned to the adapter stimulus via Hebbian learning and competitive dynamics implicit in the activation computed by the NNLS optimization. Specifically, single units highly tuned to the adapter become capable of reconstructing the stimulus without requiring the co-activation of other complementary units, in a 'grand-mother neuron'-like fashion.

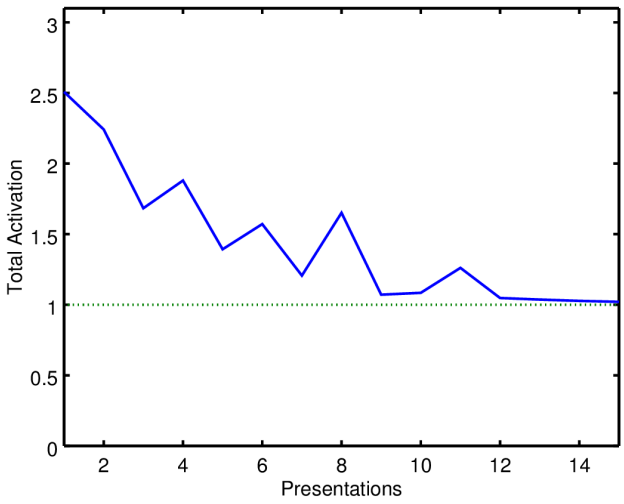


FIGURE C.1: **Results of the NNLS-based Kohonen SOM simulation**. Results of the NNLS-based Kohonen SOM simulation, showing the model's total activation during repeated exposure to a single adapter stimulus $\mathbf{a}_{\text{adapter}} = (1, 0, 1, 0, 1)$.

Bibliography

Aguado, Luis, Ignacio Serrano Pedraza, and Ana García Gutiérrez (2014). "A comparison of backward masking of faces in expression and gender identification". In: *Psicológica* 35.2.

Alink, Arjen, Hunar Abdulrahman, and Richard N. Henson (2017). "From neurons to voxels - repetition suppression is best modelled by local neural scaling". In: *bioRxiv*. DOI: 10.1101/170498. eprint: http://www.biorxiv.org/content/early/2017/07/31/170498.full.pdf. URL: http://www.biorxiv.org/content/early/2017/07/31/170498.

Auksztulewicz, Ryszard and Karl Friston (2016). "Repetition suppression and its contextual determinants in predictive coding". In: *cortex* 80, pp. 125–140.

Barlow, HB and P Foldiak (1989). "The computing neuron". In: *Adaptation and decorrelation in the cortex*, pp. 54–72.

Barnes, Jean M and Benton J Underwood (1959). ""' Fate" of first-list associations in transfer theory". In: *Journal of experimental psychology* 58.2, p. 97.

Bednar, James A. (2009). "Topographica: Building and Analyzing Map-Level Simulations from Python, C/C++, MATLAB, NEST, or NEURON Components". In: *Frontiers in Neuroinformatics* 3, p. 8. URL: http://dx.doi.org/10.3389/neuro.11.008.2009.

Bednar, James A (2012). "Building a mechanistic model of the development and function of the primary visual cortex". In: *Journal of Physiology-Paris* 106.5, pp. 194–211.

Bednar, James A and Risto Miikkulainen (2000). "Tilt aftereffects in a self-organizing model of the primary visual cortex". In: *Neural Computation* 12.7, pp. 1721–1740.

Bednar, James A and Stuart P Wilson (2015). "Cortical Maps". In: *The Neuroscientist*, p. 1073858415597645.

Ben-Yishai, R, R Lev Bar-Or, and H Sompolinsky (1995). "Theory of orientation tuning in visual cortex". In: *Proceedings of the National Academy of Sciences* 92.9, pp. 3844–3848.

Bernacchia, Alberto and Xiao-Jing Wang (2013). "Decorrelation by recurrent inhibition in heterogeneous neural circuits". In: *Neural computation* 25.7, pp. 1732–1767.

Blasdel, Gary G (1992). "Orientation selectivity, preference, and continuity in monkey striate cortex". In: *Journal of Neuroscience* 12.8, pp. 3139–3161.

Blasdel, Gary G, Guy Salama, et al. (1986). "Voltage-sensitive dyes reveal a modular organization in monkey striate cortex". In: *Nature* 321.6070, pp. 579–585.

Bosking, William H. et al. (1997). "Orientation Selectivity and the Arrangement of Horizontal Connections in Tree Shrew Striate Cortex". In: *The Journal of Neuroscience* 17.6, pp. 2112–2127.

Brett, Matthew et al. (2002). "Region of interest analysis using the MarsBar toolbox for SPM 99". In: *Neuroimage* 16.2, S497.

Brown, MW and J-Z Xiang (1998). "Recognition memory: neuronal substrates of the judgement of prior occurrence". In: *Progress in Neurobiology* 55.2, pp. 149–189.

Carreira-Perpinán, Miguel A and Geoffrey J Goodhill (2004). "Influence of lateral connections on the structure of cortical maps". In: *Journal of neurophysiology* 92.5, pp. 2947–2959.

Chen, Donghui and Robert J Plemmons (2010). "Nonnegativity constraints in numerical analysis". In: *The birth of numerical analysis*. World Scientific, pp. 109–139.

Choe, Yoonsuck (2001). *Perceptual grouping in a self-organizing map of spiking neurons*. University of Texas.

Cohen, Michael A et al. (2014). "Processing multiple visual objects is limited by overlap in neural channels". In: *Proceedings of the National Academy of Sciences* 111.24, p. 8955.

Cohen, Michael A et al. (2015). "Visual Awareness Is Limited by the Representational Architecture of the Visual System". In: *Journal of Cognitive Neuroscience* 27.11, pp. 2240–2252.

Connolly, Michael and David Van Essen (1984). "The representation of the visual field in parvicellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey". In: *Journal of Comparative Neurology* 226.4, pp. 544–564.

Connor, Charles E (2005). "Neuroscience: Friends and grandmothers". In: *Nature* 435.7045, pp. 1036–1037.

Dayan, Peter (1993). "Arbitrary elastic topologies and ocular dominance". In: *Neural Computation* 5.3, pp. 392–401.

Desimone, Robert (1996). "Neural mechanisms for visual memory and their role in attention". In: *Proceedings of the National Academy of Sciences* 93.24, pp. 13494–13499.

Dong, Dawei W (1996). "Associative decorrelation dynamics in visual cortex". In: *Lateral interactions in the cortex: structure and function*.

Dow, Bruce M. (2002). "Orientation and Color Columns in Monkey Visual Cortex". In: *Cerebral Cortex* 12.10, pp. 1005–1015.

Edelman, S (1996). "Why have lateral connections in the visual cortex?" In: *Lateral Interactions in the Cortex: Structure and Function*.

Ellis, Stephen R (1977). "Orientation selectivity of the McCollough Effect: Analysis by equivalent contrast transformation". In: *Perception & Psychophysics* 22.6, pp. 539–544.

Epstein, Russell A, Whitney E Parker, and Alana M Feiler (2008). "Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms". In: *Journal of Neurophysiology* 99.6, pp. 2877–2886.

Ermentrout, Bard, Daniel J Simons, and Peter W Land (2009). "Subbarrel patterns in somatosensory cortical barrels can emerge from local dynamic instabilities". In: *PLoS computational biology* 5.10, e1000537.

Ethier, Christian et al. (2006). "Linear summation of cat motor cortex outputs". In: *The Journal of Neuroscience* 26.20, pp. 5574–5581.

Evans, Benjamin D and Simon M Stringer (2015). "STDP in lateral connections creates category-based perceptual cycles for invariance learning with multiple stimuli". In: *Biological cybernetics* 109.2, pp. 215–239.

Fischer, Tobias (2014). *Model of All Known Spatial Maps in Primary Visual Cortex*. Master's thesis, The University of Edinburgh, UK.

Fisken, Roger A, LJ Garey, and TPS Powell (1975). "The intrinsic, association and commissural connections of area 17 of the visual cortex". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, pp. 487–536.

French, Robert M (1992). "Semi-distributed representations and catastrophic forgetting in connectionist networks". In: *Connection Science* 4.3-4, pp. 365–377.

Friston, Karl (2005). "A theory of cortical responses". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 360.1456, pp. 815–836.

Garrido, Marta I et al. (2009). "Repetition suppression and plasticity in the human brain". In: *Neuroimage* 48.1, pp. 269–279.

Georgopoulos, Apostolos P, Andrew B Schwartz, and Ronald E Kettner (1986). "Neuronal population coding of movement direction". In: *Science* 233.4771, pp. 1416–1419.

Gilaie-Dotan, Sharon and Rafael Malach (2007). "Sub-exemplar shape tuning in human face-related areas". In: *Cerebral Cortex* 17.2, pp. 325–338.

Gilbert, Charles D and Torsten N Wiesel (1989). "Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex". In: *Journal of Neuroscience* 9.7, pp. 2432–2442.

Goffaux, Valerie et al. (2013). "Local discriminability determines the strength of holistic processing for faces in the fusiform face area". In: *Frontiers in psychology* 3, p. 604.

Gotts, Stephen J, Carson C Chow, and Alex Martin (2012). "Repetition priming and repetition suppression: A case for enhanced efficiency through neural synchronization". In: *Cognitive Neuroscience* 3.3-4, pp. 227–237.

Graziano, Michael (2008). *The intelligent movement machine: an ethological perspective on the primate motor system*. Oxford University Press.

Graziano, Michael SA and Tyson N Aflalo (2007). "Mapping behavioral repertoire onto the cortex". In: *Neuron* 56.2, pp. 239–251.

Grill-Spector, Kalanit, Richard Henson, and Alex Martin (2006). "Repetition and the brain: neural models of stimulus-specific effects". In: *Trends in Cognitive Sciences* 10.1, pp. 14–23.

Grotheer, Mareike and Gyula Kovács (2016). "Can predictive coding explain repetition suppression?" In: *Cortex* 80, pp. 113–124.

Haxby, James V, Elizabeth A Hoffman, and M Ida Gobbini (2000). "The distributed human neural system for face perception". In: *Trends in cognitive sciences* 4.6, pp. 223–233.

Helmstaedter, Moritz, Bert Sakmann, and Dirk Feldmeyer (2009). "L2/3 interneuron groups defined by multiparameter analysis of axonal projection, dendritic geometry, and electrical excitability". In: *Cerebral Cortex* 19.4, pp. 951–962.

Henson, R, T Shallice, and R Dolan (2000). "Neuroimaging evidence for dissociable forms of repetition priming". In: *Science* 287.5456, pp. 1269–1272.

Henson, Richard N (2015). "Repetition suppression to faces in the fusiform face area: A personal and dynamic journey". In: *Cortex*.

Henson, RN et al. (2004). "The effect of repetition lag on electrophysiological and haemodynamic correlates of visual object priming". In: *Neuroimage* 21.4, pp. 1674–1689.

Henson, RNA and MD Rugg (2003). "Neural response suppression, haemodynamic repetition effects, and behavioural priming". In: *Neuropsychologia* 41.3, pp. 263–270.

Henson, RNA et al. (2002). "Face repetition effects in implicit and explicit memory tests as measured by fMRI". In: *Cerebral Cortex* 12.2, pp. 178–186.

Hermann, Ludimar (1870). "Eine erscheinung simultanen contrastes". In: *Pflügers Archiv European Journal of Physiology* 3.1, pp. 13–15.

Hirsch, Judith A and Charles D Gilbert (1991). "Synaptic physiology of horizontal connections in the cat's visual cortex". In: *Journal of Neuroscience* 11.6, pp. 1800–1809.

Huang, Yanping and Rajesh PN Rao (2011). "Predictive coding". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 2.5, pp. 580–593.

Hubel, David H and Torsten N Wiesel (1974). "Sequence regularity and geometry of orientation columns in the monkey striate cortex". In: *Journal of Comparative Neurology* 158.3, pp. 267–293.

Imaizumi, Kazuo et al. (2004). "Modular functional organization of cat anterior auditory field". In: *Journal of neurophysiology* 92.1, pp. 444–457.

James, Thomas W and Isabel Gauthier (2006). "Repetition-induced changes in BOLD response reflect accumulation of neural activity". In: *Human brain mapping* 27.1, pp. 37–46.

Jones, Paul D and Dennis H Holding (1975). "Extremely long-term persistence of the McCollough Effect." In: *Journal of Experimental Psychology: Human Perception and Performance* 1.4, p. 323.

Kanwisher, Nancy, Josh McDermott, and Marvin M Chun (1997). "The fusiform face area: a module in human extrastriate cortex specialized for face perception". In: *Journal of neuroscience* 17.11, pp. 4302–4311.

Kay, Kendrick N et al. (2013). "GLMdenoise: a fast, automated technique for denoising task-based fMRI data". In: *Frontiers in neuroscience* 7.

Kelly, AM Clare and Hugh Garavan (2005). "Human functional neuroimaging of brain changes associated with practice". In: *Cerebral Cortex* 15.8, pp. 1089–1102.

Keysers, Christian and David I Perrett (2002). "Visual masking and RSVP reveal neural competition". In: *Trends in Cognitive Sciences* 6.3, pp. 120–125.

King, Paul D, Joel Zylberberg, and Michael R DeWeese (2013). "Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1". In: *Journal of Neuroscience* 33.13, pp. 5475–5485.

Kohonen, Teuvo (1982). "Self-organized formation of topologically correct feature maps". In: *Biological Cybernetics* 43.1, pp. 59–69.

Koulakov, Alexei A and Dmitri B Chklovskii (2001). "Orientation preference patterns in mammalian visual cortex: a wire length minimization approach". In: *Neuron* 29.2, pp. 519–527.

Kriegeskorte, Nikolaus (2009). "Relating population-code representations between man, monkey, and computational models". In: *Frontiers in Neuroscience* 3, p. 35.

Kriegeskorte, Nikolaus and Rogier A Kievit (2013). "Representational geometry: integrating cognition, computation, and the brain". In: *Trends in Cognitive Sciences* 17.8, pp. 401–412.

Larsson, Jonas and Andrew T Smith (2012). "fMRI repetition suppression: neuronal adaptation or stimulus expectation?" In: *Cerebral Cortex* 22.3, pp. 567–576.

Law, Judith S (2009). *Modeling the development of organization for orientation preference in primary visual cortex*. The University of Edinburgh.

Li, Lin, Earl K Miller, and Robert Desimone (1993). "The representation of stimulus familiarity in anterior inferior temporal cortex". In: *Journal of Neurophysiology* 69.6, pp. 1918–1929.

Ma, Debbie S, Joshua Correll, and Bernd Wittenbrink (2015). "The Chicago face database: A free stimulus set of faces and norming data". In: *Behavior Research Methods*, pp. 1–14.

Malsburg, Chr Von der (1973). "Self-organization of orientation sensitive cells in the striate cortex". In: *Kybernetik* 14.2, pp. 85–100.

Martin, Kevan AC (2002). "Microcircuits in visual cortex". In: *Current opinion in neurobiology* 12.4, pp. 418–425.

Miikkulainen, Risto et al. (2006). *Computational maps in the visual cortex*. Springer Science & Business Media.

Moore, Christopher I, Sacha B Nelson, and Mriganka Sur (1999). "Dynamics of neuronal processing in rat somatosensory cortex". In: *Trends in neurosciences* 22.11, pp. 513–520.

Müller, Notger G et al. (2012). "Repetition Suppression versus Enhancement - It's Quantity That Matters". In: *Cerebral cortex*, bhs009.

Nauhaus, Ian et al. (2012). "Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex". In: *Nature neuroscience* 15.12, pp. 1683–1690.

Norman, Kenneth A and Randall C O'Reilly (2003). "Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach." In: *Psychological Review* 110.4, p. 611.

Norman, Kenneth A et al. (2006). "Beyond mind-reading: multi-voxel pattern analysis of fMRI data". In: *Trends in Cognitive Sciences* 10.9, pp. 424–430.

Obermayer, Klaus, Gary G Blasdel, and Klaus Schulten (1992). "Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps". In: *Physical Review A* 45.10, p. 7568.

Obermayer, Klaus, Helge Ritter, and Klaus Schulten (1990). "A principle for the formation of the spatial structure of cortical feature maps". In: *Proceedings of the National Academy Of Sciences* 87.21, pp. 8345–8349.

Ohki, Kenichi et al. (2005). "Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex". In: *Nature* 433.7026, p. 597.

Page, Mike (2000). "Connectionist modelling in psychology: A localist manifesto". In: *Behavioral and Brain Sciences* 23.04, pp. 443–467.

Pasupathy, Anitha and Charles E Connor (2002). "Population coding of shape in area V4". In: *Nature Neuroscience* 5.12, pp. 1332–1338.

Peirce, Jonathan W (2007). "PsychoPy - psychophysics software in Python". In: *Journal of neuroscience methods* 162.1, pp. 8–13.

Penfield, Wilder and Edwin Boldrey (1937). "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation." In: *Brain: A journal of neurology*.

Pennartz, Cyriel MA (2015). *The Brain's Representational Power: On Consciousness and the Integration of Modalities*. MIT Press.

Pitkow, Xaq et al. (2015). "How can single sensory neurons predict behavior?" In: *Neuron* 87.2, pp. 411–423.

Poldrack, Russell A (2000). "Imaging brain plasticity: conceptual and methodological issues - a theoretical review". In: *Neuroimage* 12.1, pp. 1–13.

Postman, Leo and Benton J Underwood (1973). "Critical issues in interference theory". In: *Memory & Cognition* 1.1, pp. 19–40.

Pouget, Alexandre, Peter Dayan, and Richard Zemel (2000). "Information processing with population codes". In: *Nature Reviews Neuroscience* 1.2, pp. 125–132.

Purushothaman, Gopathy and David C Bradley (2005). "Neural population code for fine perceptual decisions in area MT". In: *Nature Neuroscience* 8.1, pp. 99–106.

Quiroga, R Quian et al. (2008). "Sparse but not 'grandmother-cell' coding in the medial temporal lobe". In: *Trends in cognitive sciences* 12.3, pp. 87–91.

Quiroga, Rodrigo Quian, Itzhak Fried, and Christof Koch (2013). "Brain cells for grandmother". In: *Scientific American* 308.2, pp. 30–35.

Rao, Rajesh PN and Dana H Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature neuroscience* 2.1, pp. 79–87.

Ren, Ming et al. (2007). "Specialized inhibitory synaptic actions between nearby neocortical pyramidal neurons". In: *Science* 316.5825, pp. 758–761.

Ringo, James L (1996). "Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey". In: *Behavioural brain research* 76.1, pp. 191–197.

Sabatini, Silvio P (1996). "Recurrent inhibition and clustered connectivity as a basis for Gabor-like receptive fields in the visual cortex". In: *Biological cybernetics* 74.3, pp. 189–202.

Sawamura, Hiromasa, Guy A Orban, and Rufin Vogels (2006). "Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the FMRI adaptation paradigm". In: *Neuron* 49.2, pp. 307–318.

Sayres, Rory and Kalanit Grill-Spector (2006). "Object-selective cortex exhibits performance-independent repetition suppression". In: *Journal of neurophysiology* 95.2, pp. 995–1007.

Schacter, Daniel L and Randy L Buckner (1998). "Priming and the brain". In: *Neuron* 20.2, pp. 185–195.

Schiltz, Christine and Bruno Rossion (2006). "Faces are represented holistically in the human occipito-temporal cortex". In: *Neuroimage* 32.3, pp. 1385–1394.

Schiltz, Christine et al. (2010). "Holistic perception of individual faces in the right middle fusiform gyrus as evidenced by the composite face illusion". In: *Journal of Vision* 10.2, pp. 25–25.

Schwark, HD and EG Jones (1989). "The distribution of intrinsic cortical axons in area 3b of cat primary somatosensory cortex". In: *Experimental Brain Research* 78.3, pp. 501–513.

Schwartz, Eric L and Alan S Rojer (1994). "Cortical hypercolumns and the topology of random orientation maps". In: *Pattern Recognition, 1994. Vol. 2-Conference*

B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on. Vol. 2. IEEE, pp. 150–155.

Silberberg, Gilad and Henry Markram (2007). "Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells". In: *Neuron* 53.5, pp. 735–746.

Silverstein, Steven M, Docia L Demmin, and James A Bednar (2017). "Computational modeling of contrast sensitivity and orientation tuning in first-episode and chronic schizophrenia". In: *Computational Psychiatry*.

Sippy, Tanya and Rafael Yuste (2013). "Decorrelating action of inhibition in neocortical networks". In: *Journal of Neuroscience* 33.23, pp. 9813–9830.

Sirosh, Joseph (1996). "A self-organizing neural network model of the primary visual cortex". In:

Somers, David C et al. (1998). "A local circuit approach to understanding integration of long-range inputs in primary visual cortex." In: *Cerebral Cortex* 8.3, pp. 204–217.

Somers, DC et al. (1996). "Variable gain control in local cortical circuitry supports context-dependent modulation by long-range connections". In: *Lateral Interactions in the Cortex. University of Texas, Austin*.

Spigler, Giacomo (2014). *Neural modelling of the McCollough Effect in color vision*. Master's thesis, The University of Edinburgh, UK.

Spigler, Giacomo and Stuart P. Wilson (2017). "Familiarization: A theory of repetition suppression predicts interference between overlapping cortical representations". In: *PLOS ONE* 12.6, pp. 1–17. DOI: 10.1371/journal.pone.0179306. URL: https://doi.org/10.1371/journal.pone.0179306.

Stemmler, Martin, Marius Usher, and Ernst Niebur (1995). "Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics". In: *Science* 269.5232, p. 1877.

Stevens, Jean-Luc R et al. (2013). "Mechanisms for stable, robust, and adaptive development of orientation maps in the primary visual cortex". In: *The Journal of Neuroscience* 33.40, pp. 15747–15766.

Tanaka, Keiji (2003). "Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities". In: *Cerebral cortex* 13.1, pp. 90–99.

Tootell, Roger B et al. (1982). "Deoxyglucose analysis of retinotopic organization in primate striate cortex". In: *Science* 218.4575, pp. 902–904.

Tsao, Doris Y and Margaret S Livingstone (2008). "Mechanisms of face perception". In: *Annu. Rev. Neurosci.* 31, pp. 411–437.

Tsao, Doris Y et al. (2006). "A cortical region consisting entirely of face-selective cells". In: *Science* 311.5761, pp. 670–674.

Tsunoda, Kazushige et al. (2001). "Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns". In: *Nature Neuroscience* 4.8, pp. 832–838.

Turing, Alan Mathison (1952). "The chemical basis of morphogenesis". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 237.641, pp. 37–72.

Wachsmuth, E, MW Oram, and DI Perrett (1994). "Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque". In: *Cerebral Cortex* 4.5, pp. 509–522.

Wakeman, Daniel G and Richard N Henson (2015). "A multi-subject, multi-modal human neuroimaging dataset". In: *Scientific data* 2.

Webster, Michael A et al. (2004). "Adaptation to natural facial categories". In: *Nature* 428.6982, pp. 557–561.

Weliky, Michael et al. (1995). "Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns". In: *Neuron* 15.3, pp. 541–552.

Wiggs, Cheri L and Alex Martin (1998). "Properties and mechanisms of perceptual priming". In: *Current Opinion in Neurobiology* 8.2, pp. 227–233.

Wilson, Mary Ann et al. (2000). "Neonatal lead exposure impairs development of rodent barrel field cortex". In: *Proceedings of the National Academy of Sciences* 97.10, pp. 5540–5545.

Wilson, Stuart P and James A Bednar (2015). "What, if anything, are topological maps for?" In: *Developmental Neurobiology* 75.6, pp. 667–681.

Wilson, Stuart P et al. (2010). "Modeling the emergence of whisker direction maps in rat barrel cortex". In: *PloS One* 5.1, e8778.

Wimber, Maria et al. (2015). "Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression". In: *Nature Neuroscience* 18.4, pp. 582–589.

Wolf, Fred (2005). "Symmetry, multistability, and long-range interactions in brain development". In: *Physical Review Letters* 95.20, p. 208701.

Woolsey, Clinton N et al. (1951). "Patterns of localization in precentral and supplementary motor areas and their relation to the concept of a premotor area." In: *Research publications-Association for Research in Nervous and Mental Disease* 30, pp. 238–264.

Young, Andrew W, Deborah Hellawell, and Dennis C Hay (1987). "Configurational information in face perception". In: *Perception* 16.6, pp. 747–759.

Young, Malcolm P and Shigeru Yamane (1992). "Sparse population coding of faces in the inferotemporal cortex". In: *Science* 256.5061, pp. 1327–1331.