INVESTIGATING NEURAL CORRELATES OF STIMULUS REPETITION USING FMRI

Hunar Ahmad Abdulrahman

Queens' College

University of Cambridge

MRC Cognition and Brain Sciences Unit

School of Clinical Medicine

This dissertation is submitted for the degree of Doctor of Philosophy

November 2017

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

The results in Chapter 2 were reported in the paper article by Abdulrahman and Henson, (2016). Also, the results in Chapters 4-6 were reported in the paper article by Alink, Abdulrahman and Henson (submitted).

I hereby state that this dissertation does not exceed the work limit set by the Degree Committee of the Faculties of Clinical Medicine and Veterinary Medicine.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor, Rik Henson. This entire thesis is the product of our weekly in-person meetings and discussions. His invaluable help and supervision covered everything a PhD student needs, from teaching neuroimaging essentials to writing Matlab code. I would also like to thank his lovely lab team, Rabble, for their encouragement and feedback. Thanks to my secondary supervisor, Niko Kriegeskorte, for his invaluable advice and feedback. Also, thanks to Arjen Alink and Dan Wakeman for sharing their fMRI data with me and the public.

This work was conducted at the MRC Cognition and Brain Sciences Unit, which reminds me of many other lovely people to whom I owe thanks, particularly the other postgraduates, the computer support team, the library staff and many more to name.

Finally, I would like to thank Cambridge Trust and Islamic Development Bank for jointly funding my PhD. I would also like to thank the Ministry of Health/Kurdistan Regional Government – Iraq for giving me the study leave for the duration of my PhD.

ABSTRACT

Examining the effect of repeating stimuli on brain activity is important for theories of perception, learning and memory. Functional magnetic resonance imaging (fMRI) is a non-invasive way to examine repetition-related effects in the human brain. However the Blood-Oxygenation Level-Dependent (BOLD) signal measured by fMRI is far removed from the electrical activity recorded from single cells in animal studies of repetition effects. Despite that, there have been many claims about the neural mechanisms associated with fMRI repetition effects. However, none of these claims has adequately considered the temporal and spatial resolution limitations of fMRI. In this thesis, I tackle these limitations by combining simulations and modelling in order to infer repetition-related changes at the neural level. I start by considering temporal limitations in terms of the various types of general linear model (GLM) that have used to deconvolve single-trial BOLD estimates. Through simulations, I demonstrate that different GLMs are best depending on the relative size of trial-variance versus scan-variance, and the coherence of those variabilities across voxels. To address the spatial limitations, I identify six univariate and multivariate properties of repetition effects measured by event-related fMRI in regions of interest (ROI), including how repetition affects the ability to classify two classes of stimuli. To link these properties to underlying neural mechanisms, I create twelve models, inspired by single-cell studies. Using a grid search across model parameters, I find that only one model ("local scaling") can account for all six fMRI properties simultaneously. I then validate this result on an independent dataset that involves a different stimulus set, protocol and ROI. Finally, I investigate classification of initial versus repeated presentations, regardless of the stimulus class. This work provides a better understanding of the neural correlates of stimulus repetition effects, as well as illustrating the importance of formal modelling.

CONTENTS

LIST OF FIGURES

Figure 2.3: Log of precision of Sample Correlation (PSC) for two randomly intermixed trial-types for LSA (A) and LSS (B). Log of ratio of PSC in panel (C) for LSA relative to LSS. See Figure 2.2 legend for more details.

Figure 2.4: Example of sequence of parameter estimates ($\hat{\beta}_j$) for 50 trials of one stimulus class with SOA of 2s (true population mean B=3) when trial-variability (SD=0.1) is less than scan-variability (SD=0.3; top row) or trial-variability (SD=0.3) is greater than scan-variability (SD=0.1; bottom row), from LSA (left panels, in blue) and LSS (right panels, in red). Individual trial responses β_j are shown in green (identical in left and right plots).

Figure 2.5: SVM classification performance for LSA (panels A + C + E) and LSS (panels B + D + F) for incoherent trial and scan variability (panels A + B), coherent trial-variability and incoherent scan variability (panels C + D) and incoherent trial variability and coherent scan-variability (panels E + F). Note colourbar is not log-transformed (raw accuracy, where 0.5 is chance and 1.0 is perfect classification). Note that coherent and incoherent cases are equivalent when trial-variability is zero (but LSA and LSS are not equivalent even when trial-variability is zero). See Figure 2.2 legend for more details.

Figure 2.6: Log of ratio of LSA relative to LSS SVM classification performance (CP) for (A) incoherent trial variability and incoherent scan-variability, (B) coherent trial variability and incoherent scan variability and (C) incoherent trial-variability and coherent scan-variability. See Figure 2.2 legends for more detail. Note that only SOAs up to 10s are shown on the y-axes to clarify effects.

Figure 2.7: Illustration of coherent- trial and scan-variability across two voxels (SOA=2s and scan SD=0.2). Panels on top show parameters/estimates for 90 trials of each of two trial-types (trials 1-90 and 91-180 respectively) for each voxel (separate lines); bottom panels show difference between voxels for each trial (which determines CP). Left most Panels show true parameters (β_j), drawn from Gaussian with SD=0.3 and different means for each voxel. Middle and right Panels show corresponding parameter estimates ($\hat{\beta}_j$) from LSA and LSS models.

Figure 3.7: Compares the similarity of voxel pattern differences across the independent runs for each GLM (higher is better). Errors are CI 95%. Again note that error bars can be misleading for LSS and LSU as this is a paired t-test (see the text for the stats)

Figure 3.8: Compares the trial-wise classification accuracy among the GLMs using Linear SVM in FFA for A) F_init vs S_init and B) F_init vs F_imm. (Higher is better, error bars = 95% CI).

Figure 3.9:. Searchlight analyses of CP across subjects ($p<0.05$ FWE) for F_init vs S_init. Upper panels show one-sample t-tests vs chance classification; lower panels show paired t-test for LSA vs LSS.

Figure 3.10: Searchlight analyses of CP across subjects ($p<0.001$ uncorrected) for F_init vs F_imm. See Figure 3.9 legend for more details.

Figure 4.1: Shows the results of the six metrics in FFA. Bars reflect mean across participants for each condition (init = initial presentation; rep = repeated presentation), with error bars reflecting 95% confidence interval versus zero; diagonal line represents slope of linear contrast across conditions (red = positive; blue = negative) with dashed error margins reflecting 95% confidence interval of that slope (equivalent to pairwise difference when only two conditions) and p-value indicated above.

Figure 5.1: Example tuning curves along a stimulus dimension (ranging from 0 to X), both before (blue) and after (red) repetition of a single stimulus (with value X/4, shown by green line) according to the twelve different neural models of repetition suppression, created by crossing four mechanisms (rows) with three domains (columns). For illustrative purposes, only five neural populations are shown, equally-spaced along the stimulus dimension.

Figure 5.2: Distance functions, showing how amount of adaptation depends on distance between stimulus class (x-axis) and neural preference (here X/4). The *c* parameter is fixed to 0.5, while the *b* parameter is shown from 0.1 to ∞, though note that in our simulations, *b* only ranged from 0.1 to X/2.

Figure 5.3: Simulation results for each of the 12 models (columns) for each of the 6 data features (rows). Each coloured circle represents either no effect (flat; white), a decrease (blue), an increase (red), or some combination of these, when the model parameters *c*, *b*, and σ could take any value (within the grid search) for any data features (i.e parameters were not constrained to be same across data patterns; cf Figure 5.4).

Figure 5.4: The maximum number of data properties explained by each model when parameters are constrained to be equal across all data properties. Note that, for some models that can explain only 4 or 5 data properties with the same parameter values, there may be different subsets of the same number of data properties that can be explained (i.e., this figure only shows one such subset).

Figure 5.5: Predicted data features from averaging 18 simulation runs using the winning model (local scaling), with parameters: $a=0.7,\ b=0.3,\ \sigma=0.2$ (cf. data in Figure 4.1) to explain the behaviour of the average participant (fixed effect). Apart from the correlations, the Y-axes have arbitrary units. All trends were significant $p<0.001$; error bars are CI 95%.

Figure 6.1: Showing the experimental design, with the V1 mask from Alink et al (2013).

Figure 6.2: Results of the six data features in V1. Bars reflect mean across participants of each condition (init = initial presentation; rep = repeated presentation), with error bars reflecting 95% confidence interval versus zero; diagonal line represents slope of linear contrast across conditions (red = positive; blue = negative) with dashed error margins reflecting 95% confidence interval of that slope (equivalent to pairwise difference when only two conditions) and p-value indicated above.

Figure 6.3: Distance functions on circular stimulus dimension, showing how amount of adaptation depends on distance between stimulus class (x-axis) and neural preference (here X/4). The *c* parameter is fixed to 0.5, while the *b* parameter is

shown from 0.1 to ∞, though note that in our simulations, *b* only ranged from 0.1 to X/2.

Figure 6.4: Example tuning curves before (blue) and after (red) after adaptation to both orientations (X/4 and 3X/4), according to the twelve different neural models of adaptation. For illustrative purposes, we only show 4 neural populations equally-spaced along the stimulus dimension; in the simulations below, we sampled the population preferences randomly from a uniform distribution across 8 possible equally-spaced orientations (see Chapter 5 for details).

Figure 6.5: Simulation results for each of the 12 models (columns) for each of the 6 data features (rows). Each coloured circle represents either no effect (flat; white), a decrease (blue), an increase (red), or some combination of these, when the model parameters *c*, *b*, and σ could take any value (within the grid search) for any data features (i.e parameters were not constrained to be same across data patterns; cf below).

Figure 6.6: The maximum number of data properties explained by each model when parameters are constrained to be equal across all data properties. Note that, for some models that can explain only 4 or 5 data properties with the same parameter values, there may be different subsets of the same number of data properties that can be explained (i.e., this figure only shows one such subset).

Figure 6.7: Predicted data features from averaging 18 simulation runs using the winning model (local scaling), with parameters: *a=0.8, b=0.4,* σ*=0.4* (cf. data in Figure 6.2) to explain the behaviour of the average participant (fixed effect). Apart from the correlations, the Y-axes have arbitrary units. All trends were significant $p<0.001$; error bars are CI 95%.

Figure 6.8: Unpacks the AMS by ranking voxel bins by their absolute t-values for both the initial (init) and the repeated (rep) trial types and their difference (black bars). A & B shows the results in fMRI datasets while C & D shows the simulated results using the local scaling using the fitting parameters σ*=0.2 a=0.7 b=0.2* (Exp.1) and σ*=0.4, a=0.8, b=0.2* (Exp. 2).

Figure 6.9:. Example of two types of voxel possessing neural populations with narrow tuning widths ($\sigma=0.2$). Numbers at the bottom of each voxel represent the summed neuronal responses for S1 or S2 at the green line. Bars at the bottom ranked by selectivity and shows the summed response for S1 and S2 in both initial response (init) and the repeated response (rep). The black bars are the difference (init-rep) (see the text for details)

Figure 6.10:. Example of two types of voxel possessing neural populations that has the same distribution as Figure 6.9 but with broad tuning widths ($\sigma=0.4$). (see Figure 6.9 legends and the text for details)

Figure 7.1: Illustrates the effect of local scaling on two different voxels (voxel-1 and voxel-2) with different underling distributions of neural preferences. The large squares show the neural populations within each voxel in the initial phase (Panel-A) and after adaptation with a local scaling model (Panel-B). The small squares show the total voxel activities after summing the firing rate for all the neural populations at S1.

Figure 7.2: Illustrates the effect of local scaling model on Pattern Mean (PM) and Z-scored Pattern (ZP). Panel A) shows the effect of local scaling on two different voxels: voxel-1 (top row), and voxel-2 (bottom row). Each voxel has a different underlying distribution of neural preferences. (Note that the neurons in the right most panels have been adapted to both S1 and S2, though in this case the results were similar even if adaptation was released for S1). The small squares show the average voxel response coloured by their relative activity (hotter means more active). Panel B) plots the voxel responses for each stimulus before and after adaptation (left panel), then after averaging across the voxels to illustrate PM (middle panel) and after Z-scoring across the voxels to illustrate ZP (right panel). The solid lines are possible decision lines between the initial and the repeated presentations.

Figure 7.3: Compares the CP of ZP and PM in fMRI and simulation for both experiments. Parameter values used for simulations are the same as those reported previously in Figures 5.5 and 6.7.

Figure 7.4*: C*ompares the local scaling model prediction between the CP of PM and ZP for various levels of *a* and noise variance. The mean *a* was 0.7 for face paradigm and 0.8 for the grating paradigm. 18 subjects were simulated; differences are significant at $p<0.001$.

Figure 7.5: Shows the voxel pattern (x-axis is voxel number) averaged across the trials for initial (blue) and repeated (red) presentations in one example subject and for simulations of local scaling (using winning parameters values in Figure 6.7).
Figure 7.6 compares the CP of ZP and PM in fMRI and simulation for both experiments. Unlike Figure 7.3, Panels B & D now have coherent trial variability SD=0.05.

Supp. Figure 1: Efficiency comparison between LSS-N (here N = 2) and LSS-1 A) Ratio of PPM, B) Ratio of PSC LSS-2. See Figure 2.2 legend for more details.

Sup. Figure 2: Shows PPM for A) LSU, ratio of PPM for B) LSA relative to LSU, and C) LSS (LSS-2) relative to LSU. See Figure 2.2 legend for more details.

Sup. Figure 3: Simulations comparing the standard LSS beta estimates to that of LSA on a smoothed BOLD in a randomised design that has two trial types. The trials have been re-ordered in the picture where trials from 1 to 100 have a true beta magnitude of 3 while trials from 101-200 have a true beta magnitude of -3. Simulated scan variability =0.1 and trial variability =0.3

Sup. Figure 4: Showing A) adaptation effects were limited to the subruns and the BOLD activity at the start of each subrun were the same as the first trials in each subrun B) showing the effect of re-setting the adaptation factor between the independent runs in our simulation to match the empirical data results.

Sup. Figure 5: Local scaling prediction for all the 6 criteria after adding correlated neural activity to make BC positive. Around 10% of correlated neural activities (neurons with flat tuning curves) were added to the voxels for the face paradigm, and around 50% added for the grating paradigm. (init = initial, rep = repeated).

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AMA	Amplitude Modulation by Amplitude
AMS	Amplitude Modulation by Selectivity
BC	Between-class Correlation
BOLD	Blood Oxygen Level Dependent
CP	Classification Performance
EEG	Electro-encephalography
ES	Expectation Suppression
FFA	Fusiform Face Area
fMRI	Functional Magnetic Resonance
GLM	General Linear Model
HRF	Hemodynamic Response Function
LFP	Local Field Potential
LSA	Least Squares All
LSS	Least Squares Separate (Single)
LSU	Least Squares Unitary
MAM	Mean Amplitude Modulation

MEG	Magnetoencephalography
PM	Pattern Mean
MUA	Multi-Unit Activity
MVPA	Multi-Voxel Pattern Analyses
PE	Prediction Error
RE	Repetition Enhancement
ROI	Region of Interest
RS	Repetition Suppression
RSA	Representational Similarity Analyses
SD	Standard Deviation
SOA	Stimulus Onset Asynchrony
SVM	Support Vector Machine
TR	Time Resolution
V1	Primary Visual cortex
WC	Within-class Correlation
ZP	Z-scored Pattern

CHAPTER 1: INTRODUCTION

1.1 Repetition effects on the Brain activity

It is well known that repeating a stimulus has consequences for both behaviour and brain activity. Behaviourally, repetition typically leads to an increase in recognition accuracy and/or a decrease in reaction time on a given task; a phenomenon known as priming (e.g., Richardson-Klavehn & Bjork, 1988). In regard to its effect on brain activity, many studies have reported a reduction in the average neural activity after repetition; a phenomenon known as repetition suppression (RS) (for review see: Henson, 2003; Grill-Spector et al., 2006).

Repetition effects have also long been studied with single-cell recording, for example from ventral temporal lobe regions in nonhuman primates (Brown et al., 1987; Baylis and Rolls, 1987). These studies that have shown that RS can be stimulus-specific, persist for long times despite intervening stimuli, and plateau after multiple repetitions (e.g., Miller and Desimone 1993; Fahy et al., 1993; Lueschow et al., 1994; Desimone, 1996). Most neurons have a preferred stimulus, showing tuning curves that can be represented as Gaussian functions. Some studies have reported a uniform down-scaling of such tuning curves with repetition (Ringach & Bredfeldt, 2002; Swindale, 1998), whereas others like Kar and Krekelberg (2016) have reported both down-scaling and a sharpening of tuning curves. Yet other studies reported a shift in tuning curves either toward or away from the repeated stimulus (Dragoi et al., 2000; Dragoi et al., 2001; Bachatene et al., 2015; Jeyabalaratnam et al., 2013).

Figure 5.1 in Chapter 5 gives examples of each of these four types of repetition effects on neural tuning curves.

Human fMRI studies have also reported a reduction in the Blood Oxygenation Level Dependent (BOLD) signal following repetition (e.g., Buckner et al., 1998; George et al., 1999; Henson et al., 2000; James et al., 1999; Martin et al., 1995; Stern et al., 1996; Grill-Spector and Malach, 2001; Henson and Rugg, 2003). This RS normally occurs in brain regions responsive to the stimuli being repeated (relative to other stimulus types), and more recent studies have related the amount of RS to the selectivity of individual voxels to different stimulus classes (Chapter 4 provides a more detailed review). However, there is a large gap between neural tuning curves measured with single-cell recording and the voxels measured with fMRI, which is the topic of the formal models described in Chapter 5. There are also EEG (electroencephalography) and MEG (magnetoencephalography) studies of repetition effects, and occasional intracranial EEG recordings in humans (McDonald et al., 2010; Engel and McCarthy 2014; Henson, 2012; Vidal et al., 2014), but these are beyond the aims of the present thesis.

1.2 Theories of fMRI repetition suppression

The dominant theories of RS in fMRI conform to either "scaling" or "sharpening" theories (Grill-Spector et al., 2006; Blank & Davis 2016; Hatfield et al., 2016; Weiner et al., 2010; Kok et al., 2012; Spigler & Wilson 2017). These two are not necessarily exclusive and may co-occur in different brain regions, but are conceptually quite different mechanisms. I illustrate the basic concepts in Figure 1.1.

Consider an fMRI voxel containing 7x7 neuronal populations that each respond differentially to an external stimulus, S. For simplicity, I shall refer to these as "neurons" in this thesis, even though in reality they may be populations of many neurons with similar response profiles (e.g, cortical microcolumns). Upon repeated presentations (p1, p2, and p3) of S, some or all of the neurons respond less, so that the mean activity for the whole voxel decreases (i.e, RS). In the first scenario (Figure1.1-A), all neurons show reductions of activity with repetition, proportional to their initial response.

The second scenario (Figure1.1-B) proposes a selective suppression rather than a

global suppression. In this case, only the least active neurons are suppressed, leading to a sharpening of the representation of S. This idea of pruning non-selective neurons was hypothesised by Wiggs and Martin (1998). This sharpening will improve the signal-to-noise ratio for other brain regions that, for example, read-out the pattern of activity in the suppressed region, potentially enhancing perception of stimulus S.

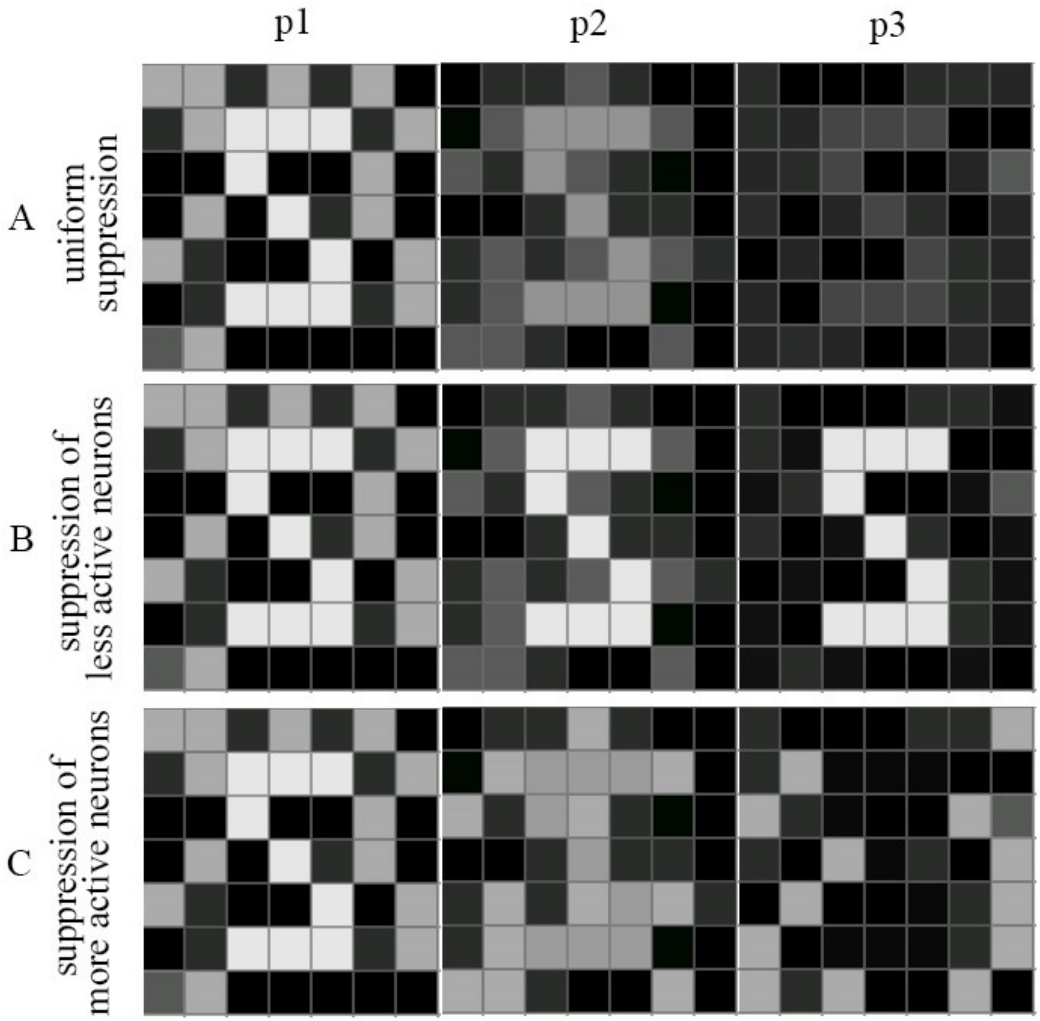


Figure 1.1: Illustrates how suppression of mean activity within a brain region of 7x7 hypothetical neurons can be achieved via global suppression of all the units or selective suppression of some neurons (in reality, each neuron represents a large number of neurons with similar response profiles). Lighter colours indicate higher activity. p number = presentation number.

The third scenario is opposite to sharpening, i.e. the most active neurons are the ones that are suppressed most (Figure1.1-C). This suppression of past stimuli could be beneficial for focusing the brain on novel stimuli (i.e, perceptual adaptation). This is

the pattern expected by predictive coding theories of perception, where the activity of neurons that respond to predicted stimuli are suppressed, and repetition improves predictions (Henson, 2003; Ewbank and Henson, 2012).

These are not necessarily the only mechanisms that could produce a decrease in the BOLD response recorded by fMRI. Indeed, because fMRI integrates over several seconds of neural activity, RS could also be caused by a reduced duration of activity, without any change in the pattern of activity over neurons (Henson, 2003). However, since the focus of this thesis is fMRI, I do not consider such temporal models any further.

Dissociating among the various theories of RS is important for understanding basic brain mechanisms. For instance, in neural network models, competitive learning uses the concept of winner-takes-all (WTA), where the maximally responsive neurons suppress weaker ones until only the strongest remains (with the remaining neurons potentially reflecting the best representation of a stimulus). If prior presentation of a stimulus strengthens the feedforward weights to the winning neurons, and/or strengthens the inhibitory connections between the winning neurons and other neurons, then this competition will be resolved more quickly and efficiently when that stimulus is presented again (Spigler & Wilson, 2017). This corresponds to the above sharpening account (Figure 1.1-B). In fact, when neurons are topologically organised, such WTA mechanisms have been used for unsupervised clustering (Bacciu & Starita, 2008).

As alluded to above, down-scaling of the most responsive neurons is consistent with another basic mechanism, that of "predictive coding". Predictive coding normally assumes a generative hierarchical network, in which higher layers constantly predict activity in lower layers (Rao & Ballard, 1999; Rao, 1999; Friston & Kiebel, 2009; Whittington & Bogacz, 2017). In return, the activity propagated from the lower layer to the higher layer is proportional to the difference between the predicted and actual activity, i.e., prediction error (PE). If repetition improves the top-down predictions (by changing synapses between layers), then the neurons signalling PE will be suppressed, leading to RS. Thus, unlike the above WTA mechanism, the predictive coding account claims that it is the neurons coding the predicted stimuli that are most suppressed (Figure 1.1-C). A further difference is that in WTA models, the relevant

synaptic changes occur in feedforward or within-layer (lateral) connections, whereas in predictive coding models, the synaptic changes occur primarily in backward connections (though in more sophisticated hierarchical models of predictive coding, forward, lateral and backward connections are all altered by experience, Friston, 2003).

There have been studies that attempt to manipulate "predictions" in terms of the expected probability of repetition (e.g, Summerfield et al., 2008), although it is unclear whether this type of contextual expectation is the same as the automatic predictions assumed between layers of hierarchical predictive coding networks (see Henson, 2016). Nonetheless, experiments that compare expected with unexpected stimuli do find BOLD response reductions (so-called "expectation suppression", ES). While some have argued for dissociable mechanisms for RS and ES (e.g., Grotheer & Kovacs, 2016), it is worth noting that authors like Kok et al., (2012) interpreted their ES in terms of voxel-sharpening (for reasons explained in Chapter 4).

Thus there is still much debate about the neural mechanisms underlying fMRI RS. Later in this thesis, I will simulate simple models, based on the single-cell data reviewed above, in their ability to explain the effects of repetition on multiple aspects of the fMRI response, including both the mean response and patterns over multiple voxels. First however, it is important to consider the temporal and spatial limitations of fMRI, and how the estimation of BOLD responses to individual trials can be optimised.

1.3 Pros and Cons of fMRI

Although single-cell studies can give precise measures about the change in the neural firing rate associated with repetition, most such studies are performed on animals that have been trained extensively to attend to stimuli. Moreover, single-cell recording is normally restricted to a few neurons in one (or at most a few) regions of interest (ROIs), and those neurons are normally excitatory cells with the largest action potentials, which may not be representative of other cells (e.g, in different layers of cortex). fMRI has the advantage of recording many brain regions simultaneously and in humans where paradigms can be more easily adjusted by changing task instructions. However, at the same time, fMRI essentially measures gross metabolic

demand within several millimetres of cortex, which can include contributions of many cell types, including inhibitory cells. The latter means that it is possible in principle to observe an increase in BOLD response in conjunction with a decrease in firing rates of the type of excitatory cells recording in single-cell studies, owing to the greater overall metabolic demand of inhibitory interneurons (or conversely, BOLD RS could be associated with increased firing rate of the subset of cells that are excitatory; for review see: Logothetis et al., 2008). It is also important to keep in mind that fMRI signals are dominated by changes in local field potentials (Goense & Logothetis, 2008), rather than the action potentials in large pyramidal cells that are normally measured in single-cell studies. Having said this, in practice, studies have reported a positive relationship between firing rate and BOLD signal (Heeger et al., 2000; Rees et al., 2000).

1.3.1 Addressing spatial limitations of fMRI

Despite the fact that the BOLD signal in a single fMRI voxel represents the average metabolic demand of millions of neurons (and that some of that signal may even come from upstream in the vasculature of the brain, several millimetres away, depending on the fMRI acquisition used), recent developments in the analysis of patterns across voxels offers the possibility of gaining higher spatial resolution. For example, it is believed to be possible to measure columnar-scale neuronal population codes (which exist a sub-voxel scales, e.g. in visual cortex) by virtue of unbalanced sampling of such columns across voxels (Kriegeskorte et al, 2010). This random variation allows statistical classifiers to decode the visual properties of a stimulus from the pattern of fMRI response across voxels – so-called "multivariate pattern analysis" (MVPA) – even if the mean response of the brain region as a whole does not differ according to those properties (Haxby et al., 2001; Norman et al., 2006). This has been demonstrated for edge orientation preferences in visual cortex for example (Kamitani and Tong 2005). It is worth noting that there is controversy around the ability of MVPA to decode fine-scaled information at the voxel level, given that spatial smoothing of fMRI data does not necessarily impair classification (de Beeck, 2010). Kamitani and Swahata (2010) proved that, although smoothing smears the fine-scaled features, the multivariate information is still retained, and the non-smoothed data can be recovered completely if the smoothing is invertable (and

the space is infinite). So smoothing, unlike subsampling, does not generally reduce the total multivariate information content of the data.

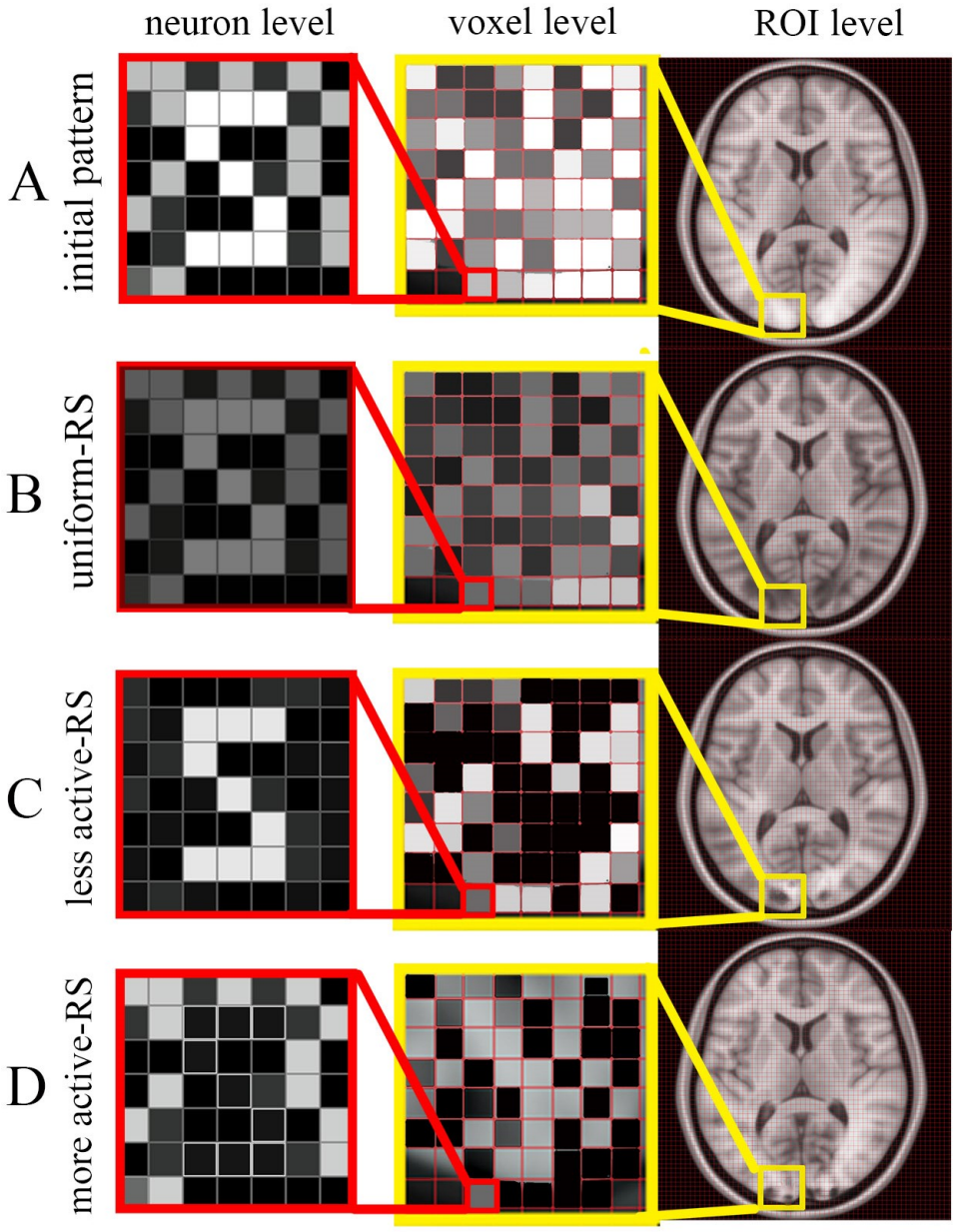


Figure 1.2: Shows the correspondence between neural pattern and the voxel pattern and compares the effects of uniform RS (Panel B) versus different types of non-uniform RS (Panels C & D) on the initial voxel pattern (see the text for more details).

Figure 1.2 illustrates the effect of RS on some of the neural patterns presented in Figure 1.1, but now situated within the context of multiple voxels within an ROI.

Across each scale there is loss of pattern information. At the highest level, averaging signal across all voxels within an ROI (shown in yellow) loses any pattern information, so cannot distinguish many plausible neural mechanisms of RS. At the level of a single voxel (shown in red), Figure 1.2 already made the point that its response cannot always distinguish uniform (global) (Figure1.2-B) from non-uniform (selective) (Figure 1.2-C & D) neural mechanisms of RS. However, if different voxels contain neurons with different distributions of selectivities, the pattern of responses across voxels can differ for different RS mechanisms (as elaborated in Chapter 5). Thus MVPA at the level of voxels can be used to constrain mechanisms at the level of neurons.

This is further illustrated in the top row of Figure 1.3. Suppose the ROI contains just two voxels that show a different response to the initial presentations of two stimuli, S1 and S2. A uniform down-scaling across the voxels (Figure 1.3-A, left panel) will always reduce the distance between S1 and S2 within the two-dimensional voxel space, hence MPVA classification of those two stimuli (indicated by the coloured decision boundary) will become more difficult after repetition (given additive noise in the data). On the other hand, down-scaling that is non-uniform across the voxels can potentially increase the separation between S1 and S2, such that classification of the two stimuli can improve after repetition (Figure 1.3-A, middle panel). This is sometimes called a "sharpening of voxel representations", though it is important to note that this does not imply that the neural tuning curves within each voxel are sharpened (i.e, does not necessarily imply that the least selective neurons are being suppressed within each voxel): as Chapter 5 will demonstrate, sharpening at the voxel level can arise from by down-scaling the most selective neurons within each voxel. This illustrates the potential confusion that can arise when applying concepts like "sharpening" to different levels of neuroscientific measurement.

Although MVPA may potentially distinguish uniform from non-uniform effects of repetition, non-uniform effects on neurons do not necessarily improve MVPA classification: depending on properties of the neural tuning curves and effects of neuronal adaptation (described in Chapter 5), repetition can also impair classification performance (Figure 1.3-A, right panel). This reinforces the potential for misinterpretation when linking fMRI findings to neural models, as in Kok et al.

(2012). To minimise this misinterpretation, I will later distinguish the effect of repetition not only on overall classification performance, but also on within-class similarity and between-class similarity. While classification performance generally increases when within-class similarity increases and between-class similarity decreases, it is possible, for example, for both within- and between-class similarity to increase, but within-class similarity increase more, so that overall classification increases (more details in Chapter 4).

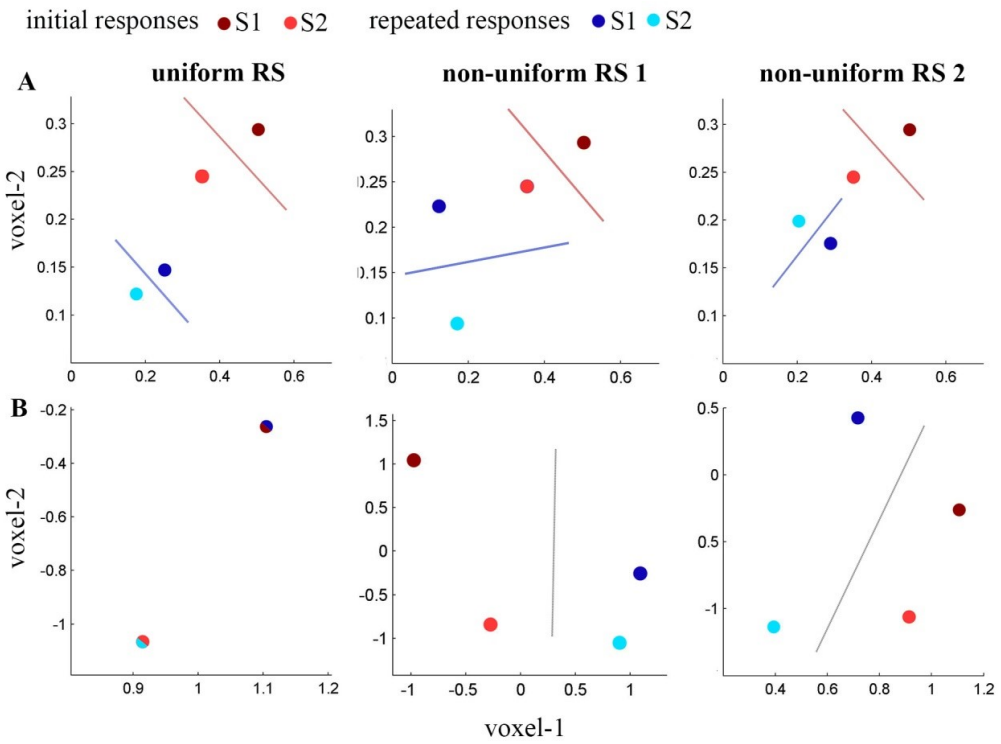


Figure 1.3 compares various effects of repetition on stimulus patterns across two voxels. Top panels show effect of repetition on the ability to classify two stimuli, S1 and S2 (red lines show possible decision boundary for initial presentations; blue lines are for repeated presentations). Bottom panels show same patterns in top panels, but after Z-scoring across the voxels (i.e, mean-correcting and scaling by standard deviation over voxels), and the grey lines now show possible decision boundaries for classifying initial versus repeated presentations, regardless the stimulus type. Note that the simulations involved more than 2 voxels, but are shown for just 2 voxels for simplicity.

Rather than examining how repetition affects the ability to classify two or more stimulus classes, MVPA can also be used to classify initial versus repeated

presentations, collapsing across stimulus class. This is illustrated in Figure 1.3-B. In order to prevent classification being based on the mean response across voxels (i.e, RS), or simply an overall scaling effect, the data in the bottom panels of Figure 1.3 have been Z-scored. Therefore any ability to classify initial versus repeated presentations must depend on changes in the relative responses across voxels (i.e. fine-grained pattern information). Thus a uniform scaling effect would not enable such classification (bottom left panel). Nonetheless, non-uniform scaling can enable such classification. More specifically, repetition must consistently produce a greater shift along one axis (voxel) than another. This is elaborated in Chapter 7.

Despite the potential "hyper-resolution" offered by MVPA, inferring the responses of neurons within an fMRI voxel will always be an inverse problem with no unique solution. In such situations, an alternative strategy is to simulate a number of candidate "forward models" that map from neurons/neurons (or neural populations) to voxels, and compare them in their relative ability to capture voxel-level fMRI data. Even if one can never know that the winning model is the true model (without other independent data, e.g, from single-cell recording), one can still make progress by determining the most likely model from a set of current theories. This is the approach taken in Chapters 5-7. While a small number of studies have tried to link their fMRI findings to neural adaptation models before (e.g, Andresen et al., 2009, Wiener et al., 2010, Kok et al., 2012; Hatfield et al., 2016), they failed to model the fact that neural adaptation often depends on the difference between a neuron's preference and the stimulus presented, a realisation that increases the number candidate models; nor did they consider a sufficient number of univariate and multivariate (MVPA) fMRI properties of stimulus repetition (which are necessary to further constrain likely models).

1.3.2 Addressing temporal limitations of fMRI

Single-cell recording also has the temporal resolution to separate responses locked to individual stimuli. Because fMRI measures the haemodynamic response that follows changes in neural activity, and this response can last 20-30s (Zarahn et al., 1997), measuring the response to individual stimuli (trials) requires a much longer time between stimuli (i.e, stimulus onset asynchrony, SOA). This makes fMRI experiments much less time-efficient. However, there is some evidence for linearity

in the BOLD response, i.e, that the response to stimuli close together in time can be predicted by simply summing the response to each stimulus alone. This allows linear deconvolution methods to estimate the responses to individual trials, even if those trials are only a few seconds apart (e.g, SOAs down to 2s), resulting in much more time-efficient experiments.

While there have been several methodological studies quantifying the efficiency of the General Linear Models (GLMs) used to deconvolve fMRI data (as a function of SOA and stimulus ordering), few of these studies have considered the effect of variability from trial to trial. This type of variability becomes especially important when it represents systematic differences between stimuli, and when that variability is expressed by systematic patterns across voxels (in MVPA). In Chapters 2-3, I use both simulations and real data to compare the ability of different GLMs to estimate fMRI responses as a function of trial variability and scan variability (scanner noise) and as a function of the coherency of those sources of variability across voxels.

1.4 Outline of thesis

Examining the effect of repeating stimuli on brain activity is an important research topic because of its potential link to perception, learning and memory. In this introductory chapter, I presented fMRI as a convenient, non-invasive way to examine repetition-related effects across the whole human brain. However, the BOLD signal measured by fMRI is far removed from the electrical activity recorded from single cells in animal studies of repetition effects. More specifically, relating fMRI to neuronal activity is challenging for several reasons, including the fact that 1) the BOLD response is slow, such that it is difficult to deconvolve responses to stimuli repeated in rapid succession efficiently for both univariate (from a single voxel or brain region) and multivariate (multiple voxels) analyses in the same design, and 2) the spatial resolution of the BOLD signal is limited, with the signal from a single voxel representing the summation a large number of neurons that may have quite different selectivities to different stimuli. Despite of these difficulties, there have been many claims about the neural mechanisms associated with fMRI repetition effects, using both univariate analysis and multivariate pattern analysis. However, none of these studies has adequately considered the temporal and spatial resolution limitations of fMRI. In this thesis, I attempt to tackle these limitations by combining

simulations and modelling with analyses of real fMRI data in order to infer repetition-related changes at the neural level. In Chapter 2, I focus on the GLMs commonly used to deconvolve single-trial BOLD estimates. Through simulations, I demonstrate that different types of GLMs are more efficient depending on the relative size of trial-variance versus scan-variance (a more extensive version of this work has been published in Abdulrahman & Henson, 2016). In Chapter 3, I use a publically-available fMRI dataset on face repetition effects to validate my simulation results, and identify the most efficient GLM for estimating single trial betas in that dataset (and use this type of GLM for subsequent chapters). In Chapter 4, I identify six important univariate and multivariate effects of repeating briefly-presented faces in a fusiform ROI (the fusiform face area, FFA). To link these properties to underlying neural adaptation mechanisms, in Chapter 5 I propose twelve neural adaptation models, all inspired by the findings of single-cell studies. Using a grid search across model parameters, I find that models that assume a non-uninform adaptation across the stimulus dimension fit the data better. Indeed, I show that only one of these models ("local scaling") can account for all six fMRI properties simultaneously. In Chapter 6, I use an independent dataset that involves different stimuli (oriented gratings), protocol (sustained adaptation) and ROI (V1), and again find that only a single "local scaling" model can simultaneously fit all six data properties (despite two of these properties differing from those of the first dataset). The results of Chapters 5 and 6 are also reported in a paper currently under review. Finally, in chapter 7, I classify initial from repeated trials (regardless of the stimulus type) and compare the results of the data to that of the local scaling model. In this case, it proves necessary to add coherent trial-variability across voxels to the model, in order to explain how the data show classification based on the Z-scored pattern exceeds classification based on the mean (returning to a theme in Chapter 2). The thesis therefore provides a better understanding of repetition-related changes in fMRI and their possible underlying neural mechanisms. It also illustrates the importance of simulations and formal modelling when trying to interpret fMRI data.

1.5 Chapter Summary:

In this chapter, I briefly reviewed single-cell recording and fMRI studies of repetition effects, and considered how various neural mechanisms might relate to voxel-level changes in fMRI. Then I outlined the challenges associated with interpretations of

repetition-related changes in fMRI and the difficulty of mapping back the results to the neural domain. More detailed introductions to relevant concepts are given at the start of subsequent chapters.

CHAPTER 2: OPTIMAL GLM MODEL TO ESTIMATE TRIAL SPECIFIC BETA

2.1 Introduction

Many fMRI experiments use rapid presentation of trials of different types (conditions). Because the time between trial onsets (SOA) is typically less than the duration of the BOLD impulse response (or haemodynamic response function, HRF), the fMRI responses to successive trials overlap. The majority of fMRI analyses use linear convolution models like the GLM to extract estimates of responses to individual trials (i.e, to deconvolve the fMRI response; Friston et al., 1998). The parameters of the GLM, reflecting responses to each trial or trial-type, are estimated by minimizing the squared error across scans (where scans are typically acquired with repetition time, or TR, of 1-2s) between the timeseries recorded in each voxel and the timeseries that is predicted, based on the known trial onsets, assumptions about the shape of the HRF and assumptions about noise in the fMRI data.

Many papers have considered how to optimize the design of an fMRI experiment, in order to maximize statistical efficiency for a particular contrast of trial-types (Friston et al., 1999; Dale et al., 1999; Josephs & Henson, 1999). However, these papers have tended to consider only the variability induced by the probability of occurrence of trials of each type and the (minimal) SOA, while assuming a fixed HRF and, most

relevant here, a fixed response to each trial of the same type. While some studies have considered other sources of variability, such as that in the HRF across participants (Aguirre et al., 1998; Neumann et al., 2003; Handwerker et al., 2004), few studies have considered different ways of modelling the variability in the amplitude of neural activity evoked from trial to trial (though see Josephs & Henson, 1999; Mumford et al., 2012). Such variability across trials might include systematic differences between the stimuli presented on each trial (Davis et al., 2014). This is the type of variability, when expressed differently across voxels, that is of interest to multi-voxel pattern analysis (MVPA), such as representational similarity analysis (RSA) (Mur et al., 2009). However, trial-to-trial-variability is also likely to include other components such as random fluctuations in attention to stimuli, or variations in endogenous (e.g., pre-stimulus) brain activity that modulates stimulus-evoked responses (Fox et al., 2006); variability that can occur even for replications of exactly the same stimulus across trials. This is the type of variability utilized by trial-based measures of functional connectivity between voxels for example (so-called "Beta-series" regression, Rissman et al., 2004).

2.2 Types of GLM

Provided the HRF is modelled with single (canonical) shape, standard efficiency analyses model all trials of the same type with a single regressor (Figure 2.1-A), or what we called Least Squares Unitary (LSU) in Abdulrahman & Henson (2016). If however one wants to allow for variability in the response across trials of the same type, then one has two further options. One could model each trial as a separate regressor in the GLM (Figure 2.1-B), which Mumford et al. (2012) called "Least-Squares All" (LSA), in terms of the GLM minimizing the squared error across *all* trials. Alternatively one could use a method originally proposed by Turner et al (2010) called "Least-Squares Separate" (LSS; Figure 2.1-C). This method actually estimates a separate GLM for each trial. Within each GLM, the trial of interest (target trial) is modelled as one regressor, and all the other (non-target) trials are collapsed into another regressor.[1] This approach has been promoted for designs with

[1] When there is more than one trial-type (condition), the regressor for the non-target trials can be split into separate regressors for non-targets of each condition. This enables differences in the mean response across conditions to be modelled. This approach is called "LSS-N" for N conditions, and generally found to be more sensitive than "LSS-1"

short SOAs, by virtue of providing better estimates when there is a high level of collinearity between BOLD responses to successive trials (Mumford et al., 2012).

Note that I only consider unbiased estimators here, e.g., do not consider here regularized LSA (Abdulrahman & Henson, 2016) such as "ridge regression" (Mumford et al., 2012), which can improve efficiency at the cost of biasing estimates towards zero. This means that the measures of design efficiency can focus on minimizing the variability in the GLM parameter estimates, ignoring any differences in mean or scaling from their true value.

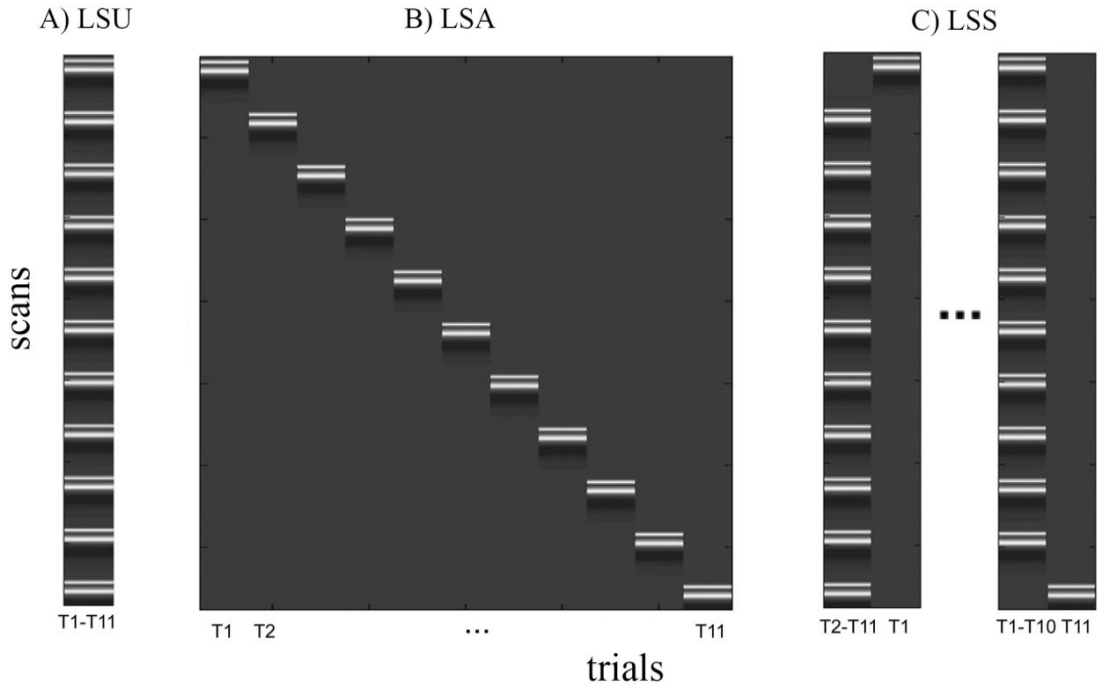


Figure 2.1: Example fMRI GLM design matrices for A) Least-Squares Unitary (LSU), B) Least-Squares All (LSA), C) Least-Squares Separate (LSS) with SOA = 32s. Tn = Trial number.

In the current study, I simulated the effects of different levels of trial-to-trial variability (trial-variability), as well as scan-to-scan-variability (i.e., scan-variability or scanner noise), on the ability to estimate responses to individual trials, across a range of SOAs (assuming that neural activity evoked by each trial was brief – i.e., less than 1s – and locked to the trial onset, so that it can be effectively modelled as a

(Abdulrahman & Henson, 2016) which collapses the regressors for all non-targets into one regressor (see Appendix 1). In the present thesis, "LSS" always refers to "LSS-N".

delta function). More specifically, I compared the relative efficiency of the different GLMs for two main questions: 1) estimating the BOLD response to each individual trial in a single voxel in the presence of a varying ratios of scan-variability to trial-variability, and 2) estimating the pattern of responses across voxels for each trial in the presence of either incoherent or coherent trial-/scan-variability across the voxels.

2.3 Simulations

I simulated fMRI timeseries for a fixed scanning duration of 45 minutes (typical for fMRI experiments), sampled every TR=1s. I modelled events by delta functions that were spaced with SOAs in steps of 1s from 2 to 32s, and convolved with SPM12's canonical HRF, scaled to have peak height of 1. To match the simulations of Mumford et al. (2012), the scaling of the delta-functions (true parameters) for the first trial-type (at a single voxel) was drawn from a Gaussian distribution with a population mean of 3 and standard deviation (SD) that was one of 0, 0.5, 0.8, 1.6 or 3 (trial-variability). Independent zero-mean Gaussian noise (scan-variability) was then added to each TR, with SD of 0.5, 0.8, 1.6 or 3, i.e, amplitude SNRs of 6, 3.8, 1.9 or 1 respectively. A second trial-type had a population mean of 5 (with same range of SDs for trial-variability and scan-variability).

For the simulations of two trial-types across two voxels, the same sample of parameter values was used for each voxel (coherent trial-variability) or different samples were drawn for each voxel (incoherent trial-variability, i.e, independent across voxels). The GLM parameters ("Betas") were estimated by ordinary least-squares fit of GLMs conforming to each of the GLMs in Figure 2.1. A final constant term was added to each GLM to remove the mean BOLD response (given that the absolute value of the BOLD signal is arbitrary). The precision of these parameter estimates was estimated by repeating the data generation and model fitting N=10,000 times, except when classification performance was examined with a linear Support Vector Machine (SVM), where N=1,000 for computational tractability. This precision can be defined in several ways, depending on the question, as detailed in the Results.

Transients at the start and end of the session were ignored by discarding the first and last 32s of data (32s was the length of the canonical HRF), and only modelling trials whose complete HRF could be estimated. Note also that I assumed linear summation

of all responses, e.g., no saturation of the neural or haemodynamic response for short SOAs that is likely to be significant for SOAs below 1-2s (Friston et al., 1998).

2.4 Results – Univariate analyses

2.4.1 Optimal SOA for estimating the average trial response in a single voxel

For this question, one wants the most precise (least variable) estimate of the mean response across trials (and does not care about the responses to individual trials). If one regards each trial as measuring the same "thing", but with a random (zero-mean) noise element, then the relevant measure is the precision of the population mean (PPM):

$$PPM = \frac{1}{std_{i=1..N}\left(\sum_{j=1}^{M}\frac{\hat{\beta}_{ij}}{M} - \beta\right)} = \frac{1}{std_{i=1..N}\left(\sum_{j=1}^{M}\frac{\hat{\beta}_{ij}}{M}\right)}$$

where $std_{i=1..N}$ is the standard deviation across N simulations and $\hat{\beta}_{ij}$ is the parameter estimate for the j-th of M trials in the i-th simulation. β is the true population mean, though as a constant, is irrelevant to PPM. Note that, because the least-square estimators are unbiased, the difference between the estimated and true population mean will tend to zero as the number of trials (M) tends to infinity. The PPM measure is relevant when each trial includes, for example, random variations in attention, or when each trial represents a stimulus drawn randomly from a larger population of stimuli, and differences between stimuli are unknown or uninteresting.

When there are two trial-types, $\hat{\beta}_{ij}$ can simply be recast as the difference between the parameter estimates for regressors of each trial-type (in example here, the true difference is 5-3=2). Because LSU cannot yield single trial betas, it is not of a primary focus in this thesis. However, for the sake of comparison I ran some further simulations comparing LSA and LSS models to LSU (see Appendix 2). In general, when the average Beta response is concerned, LSS's performance is very similar to LSU with a slight advantage for LSU in short SOAs when trial-variability high.

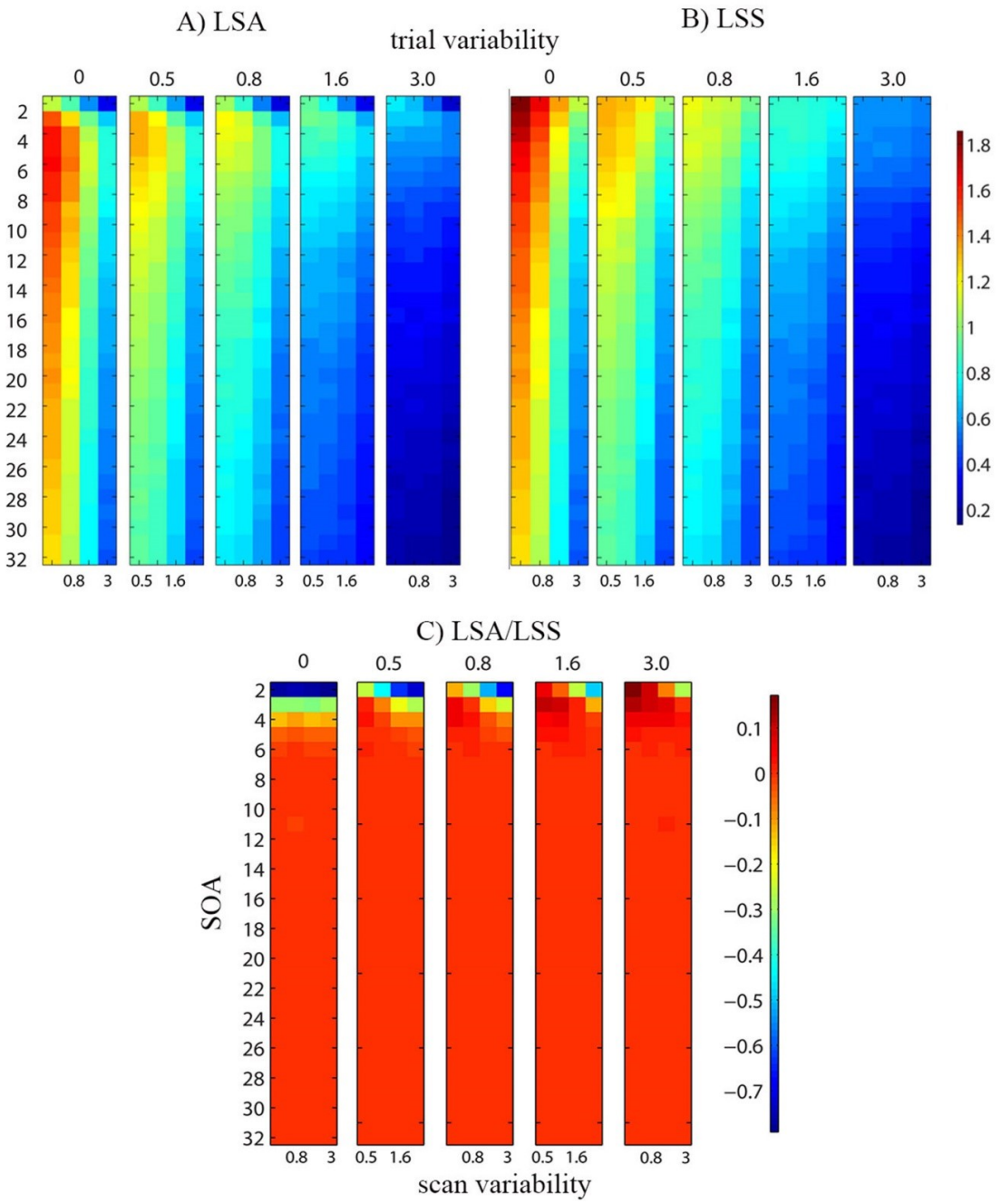


Figure 2.2: Precision of Population Mean (PPM) of the difference between two trial-types as a function of SOA (y-axes) and scan-variability (numbers on the bottom x-axes), for each degree of trial-variability (numbers on the top x-axes) for LSA (Panel A) and LSS (Panel B). Ratio of PPM for LSA relative to LSS (Panel C). The colour map for panel C has been log transformed to base 10 for visibility.

Figure 2.2 shows PPM plotted as a function of SOA, scan-variability and trial-variability for estimating this difference in a randomized design (where trials of each

type are randomized in order). LSS is shown in Figure 2.2-B (LSU is qualitatively similar; Appendix 2). When trial variability is zero, the results replicate those of previous efficiency analyses (e.g, Josephs & Henson, 1999), i.e, for estimating the difference between two randomly intermixed trial-types, the shortest SOA is optimal. This conclusion remains even if the trial variability is increased.

For LSA however, Figure 2.2-A shows a slightly different result, where the optimal SOA when trial-variability is zero is 4-6s. This only reduces to the minimum SOA (2s) as the scan-variability increases (i.e, the ratio of trial-variability to scan–variability decreases). Figure 2.2-C compares PPM directly for LSA relative to LSS models, for both models for a wide range of SOA, trial and scan variabilities. As can be seen, LSS is a better choice of GLM when the SOA is below about 5s, particularly when trial-variability is low. Only when the ratio of trial-variability to scan-variability is high (rightmost section of panel C) does LSA confer an advantage at such short SOAs.

2.4.2 Optimal GLM for estimating individual trial responses in a single voxel

For this question, one wants the most precise estimate of the response to each individual trial at each individual voxel, as necessary for example for trial-based connectivity estimation (Rissman et al., 2004), or for the voxel-wise feature selection sometimes used for dimension reduction in MVPA. In this case, a simple metric of efficiency is the Precision of Sample Correlation (PSC), defined as:

$$PSC = \sum_{i=1}^{N} \frac{cor_j(\hat{\beta}_{ij}, \beta_{ij})}{N}$$

where $\hat{\beta}_{ij}$ and β_{ij} are the estimated and true values respectively for the *j*-th of *M* trials in the *i*-th simulation, and $cor(x, y)$ is the sample (Pearson) correlation between x and y. Note that PSC is not defined when the trial-variability is zero (because β_{ij} is constant).

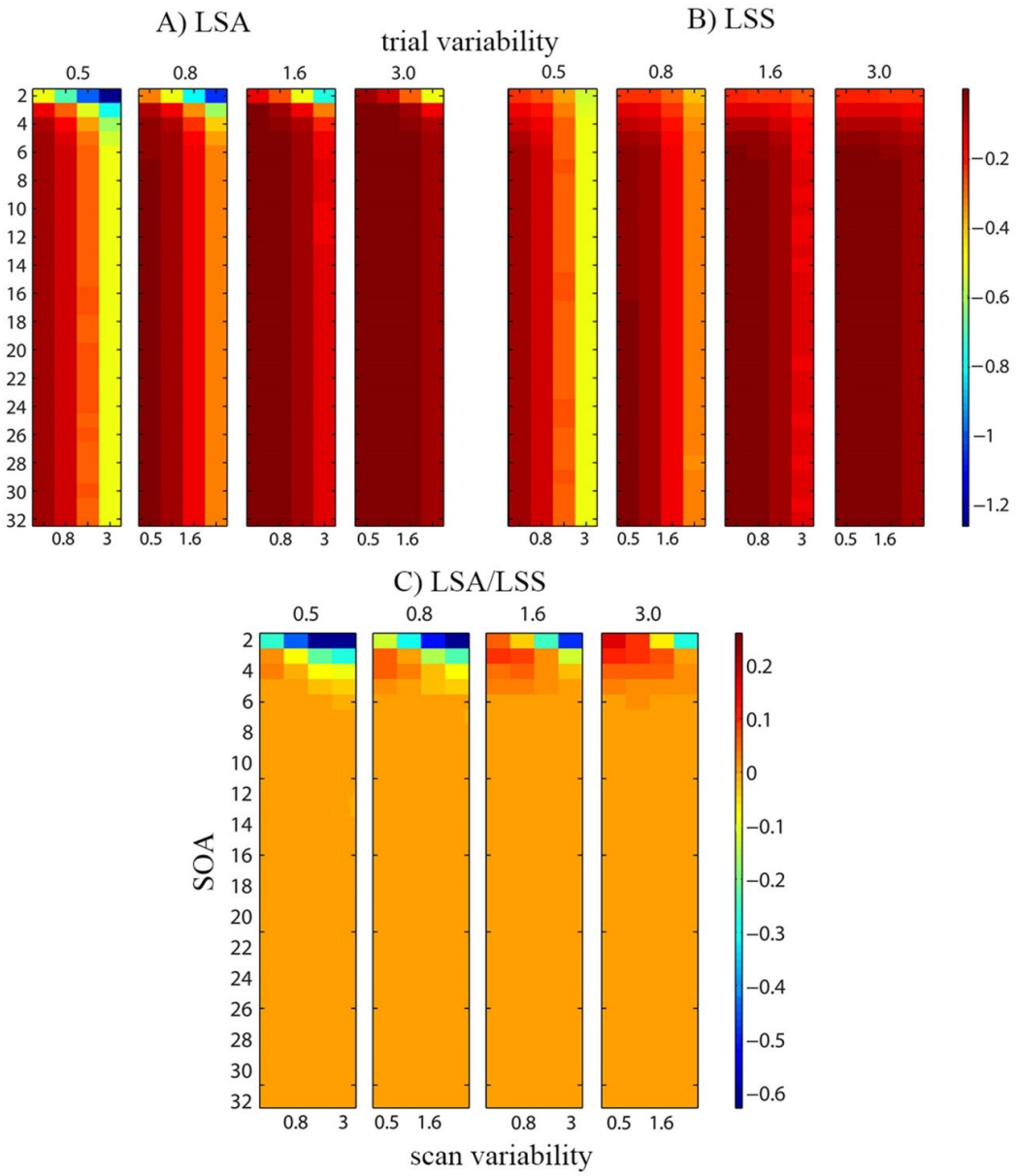


Figure 2.3: Log of precision of Sample Correlation (PSC) for two randomly intermixed trial-types for LSA (A) and LSS (B). Log of ratio of PSC in panel (C) for LSA relative to LSS. See Figure 2.2 legend for more details.

Figure 2.3-A and 2.3-B show that PSC increases as SOA increases for both LSA and LSS (LSU cannot be used here because it does not yield single trial estimates), with LSA doing particularly poorly for short SOAs when the ratio of trial-variability to scan-variability is low. Figure 2.3-C shows the ratio of PSC for LSA relative to LSS. In this case, for short SOAs, LSA is better when the ratio of trial-variabilty to scan-variability is high, but LSS is better when the ratio of trial-variabilty to scan-variability is low. It is worth considering the reason for this in a more detail.

The reason is exemplified in Figure 2.4, which shows examples of true and estimated parameters for LSA and LSS for a single trial-type when the SOA is 2s. The LSA estimates (in blue) fluctuate more rapidly across trials than do the LSS estimates (in red) – i.e., LSS forces temporal smoothness across estimates. When scan-variability is greater than trial-variability (top row), LSA "overfits" the scan noise (i.e, attributes some of the scan-variability to trial-variability). In this case, the "temporally regularized" LSS estimates are superior. However, when trial-variability is greater than scan-variability (bottom row), LSS is less able to track rapid changes in the trial responses, and LSA becomes a better model.

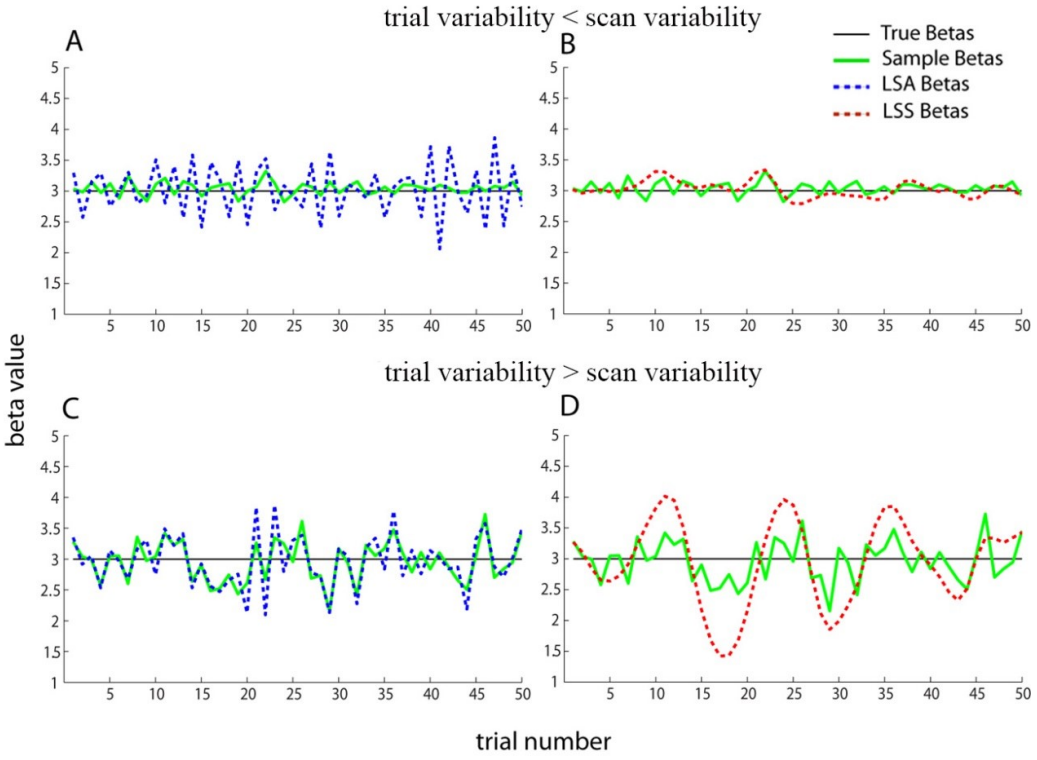


Figure 2.4: Example of sequence of parameter estimates ($\hat{\beta}_j$) for 50 trials of one stimulus class with SOA of 2s (true population mean B=3) when trial-variability (SD=0.1) is less than scan-variability (SD=0.3; top row) or trial-variability (SD=0.3) is greater than scan-variability (SD=0.1; bottom row), from LSA (left panels, in blue) and LSS (right panels, in red). Individual trial responses β_j are shown in green (identical in left and right plots).

2.5 Results – Multivariate analyses

2.5.1 Optimal SOA for estimating pattern of individual trial responses over voxels

For this question, one wants the most precise estimate of the relative pattern across voxels of the responses to each individual trial, as relevant to MVPA (Davis et al. 2014). For this question, our measure of efficiency was classification performance (CP) of a linear support-vector machine (SVM), which was fed the pattern for each trial across two voxels. Of course, different types of classifiers may produce different CP levels, but one would expect the qualitative effects of SOA, trial-variability and scan-variability to be the same.

In the case of multiple voxels, there may be spatial correlation in the trial variability and/or in the scan variability across the voxels. I start by assuming that the scan-variability (scan noise) and trial-variability are independent across voxels. Then I consider two more special cases where either trial-variability or scan-variability is correlated (coherent) across the voxels. For coherent trial-variability, the response for a given trial was identical across voxels, whereas for incoherent trial-variability, responses for each voxel were drawn independently from the same Gaussian distribution. Coherent trial-variability may be more likely (e.g, if levels of attention affect responses across all voxels in a region), though incoherent trial-variability might apply if voxels respond to different features of the same stimulus. In practice, there may be a non-perfect degree of spatial correlation across voxels in the trial-variability, but by considering the two extremes, one can interpolate to intermediate cases. I also considered the case of coherent scan-variability with incoherent trial-variability because in some cases it is likely for the scan noise to be correlated owing to, for example, artifacts remaining from motion of the whole head (not locked to the trials).

Figure 2.5 shows CP for incoherent trial variability and incoherent scan noise (top row), coherent trial variability and incoherent scan noise (middle row) and incoherent trial variability and coherent scan noise (bottom row), for LSA (left) and LSS (right). When scan variability is incoherent (i.e., comparing top and middle rows), the most noticeable effect of coherent relative to incoherent trial variability was to maintain CP as trial variability increased, while the most noticeable effect of LSS relative to LSA was to maintain CP as SOA decreased. On the other hand, when scan-variability is coherent and trial-variability is not, CP was maintained as the

scan-variability increased, and LSA was better able than LSS to maintain CP as SOA decreased. In short, making trial variability or scan noise coherent across voxels minimizes the effects of that type of variability on CP, because CP only cares about relative patterns across voxels.

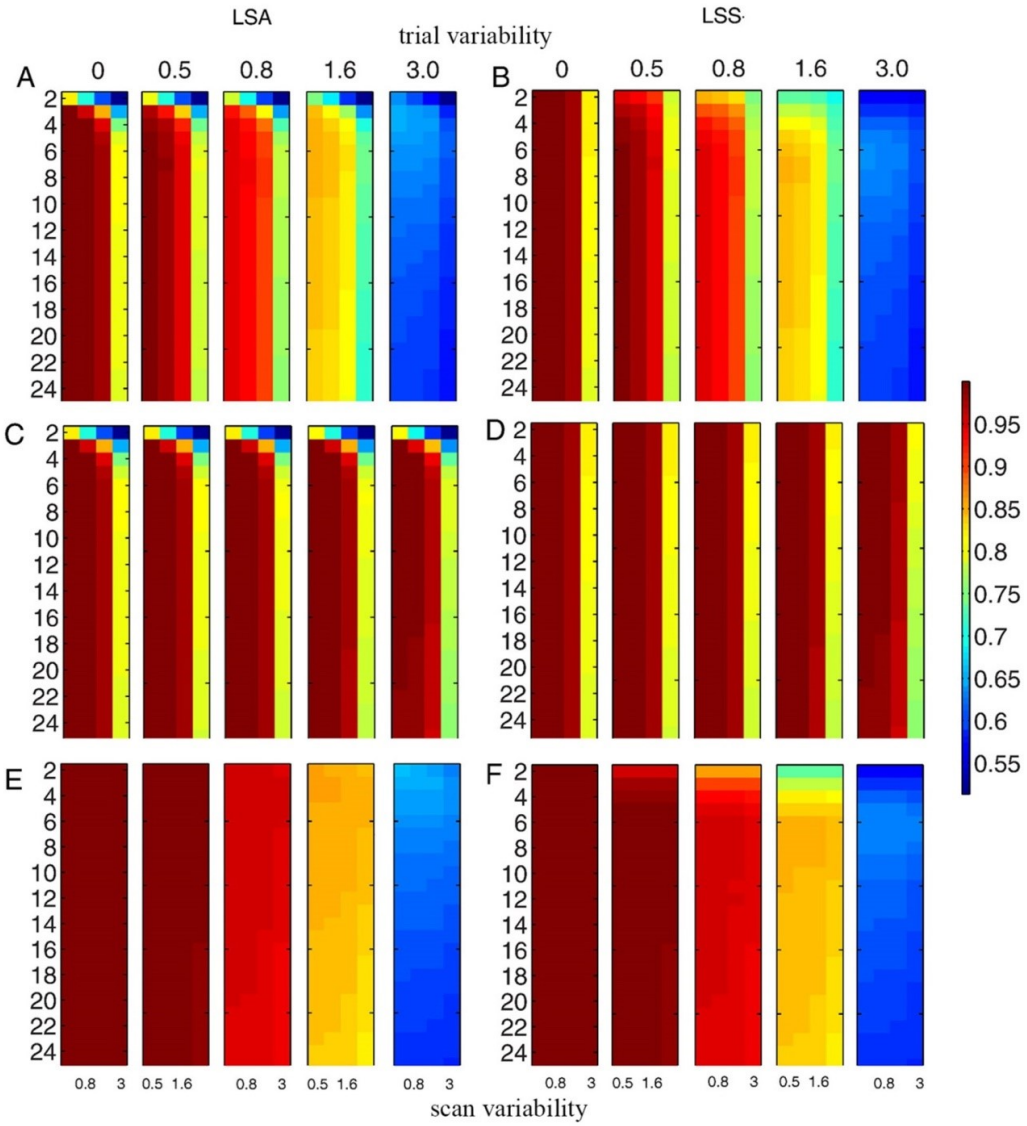


Figure 2.5. SVM classification performance for LSA (panels A + C + E) and LSS (panels B + D + F) for incoherent trial and scan variability (panels A + B), coherent trial-variability and incoherent scan variability (panels C + D) and incoherent trial variability and coherent scan-variability (panels E + F). Note colourbar is not log-transformed (raw accuracy, where 0.5 is chance and 1.0 is perfect classification). Note that coherent and incoherent cases are equivalent when trial-variability is zero (but LSA and LSS are not equivalent even when trial-variability is zero). See Figure 2.2 legend for more details.

When trial variability and scan noise are both incoherent (top row), the SOA has little effect for LSA and LSS when trial variability is low (as long as SOA is more than approximately 5 s in the case of LSA), but becomes optimal around 3–8 s as trial variability increases. With coherent trial variability and incoherent scan noise (middle row), SOA has little effect for low scan noise (again as long as SOA is not too short for LSA), but becomes optimal around 6–8 s for LSA, or 2 s for LSS, when scan noise is high. With incoherent trial variability and coherent scan noise (bottom row), the effect of SOA for LSA was minimal, but for LSS, the optimal SOA approached 6–7 s with increasing trial variability. The reason for these different sensitivities of LSA and LSS to coherent versus incoherent trial variability is explored in the next section.

2.5.2 Optimal GLM for estimating pattern of individual trial responses over voxels

Figure 2.6 shows the (log) ratio of CP for LSA relative to LSS for the three rows in Figure 2.5 above. Differences only emerge at short SOAs. For incoherent trial-variability and incoherent scan-variability (Figure 2.6-A), LSS is superior when the ratio of trial-variability to scan-variability is low, whereas LSA is superior when the ratio of trial-variability to scan-variability is high, much like in Figure 2.2-C (univariate analyses). For coherent trial-variability and incoherent scan-variability (Figure 2.6-B), on the other hand, LSS is as good as, or superior to LSA (for short SOAs), when coherent trial variability dominates across the voxels (i.e., the LSA:LSS ratio never exceeds 1, i.e. the log ratio never exceeds 0). For incoherent trial-variability and coherent scan-variability (Figure 2.6-C), LSA is as good as, or superior to LSS (for short SOAs), particularly when trial variability is high and scan noise low.

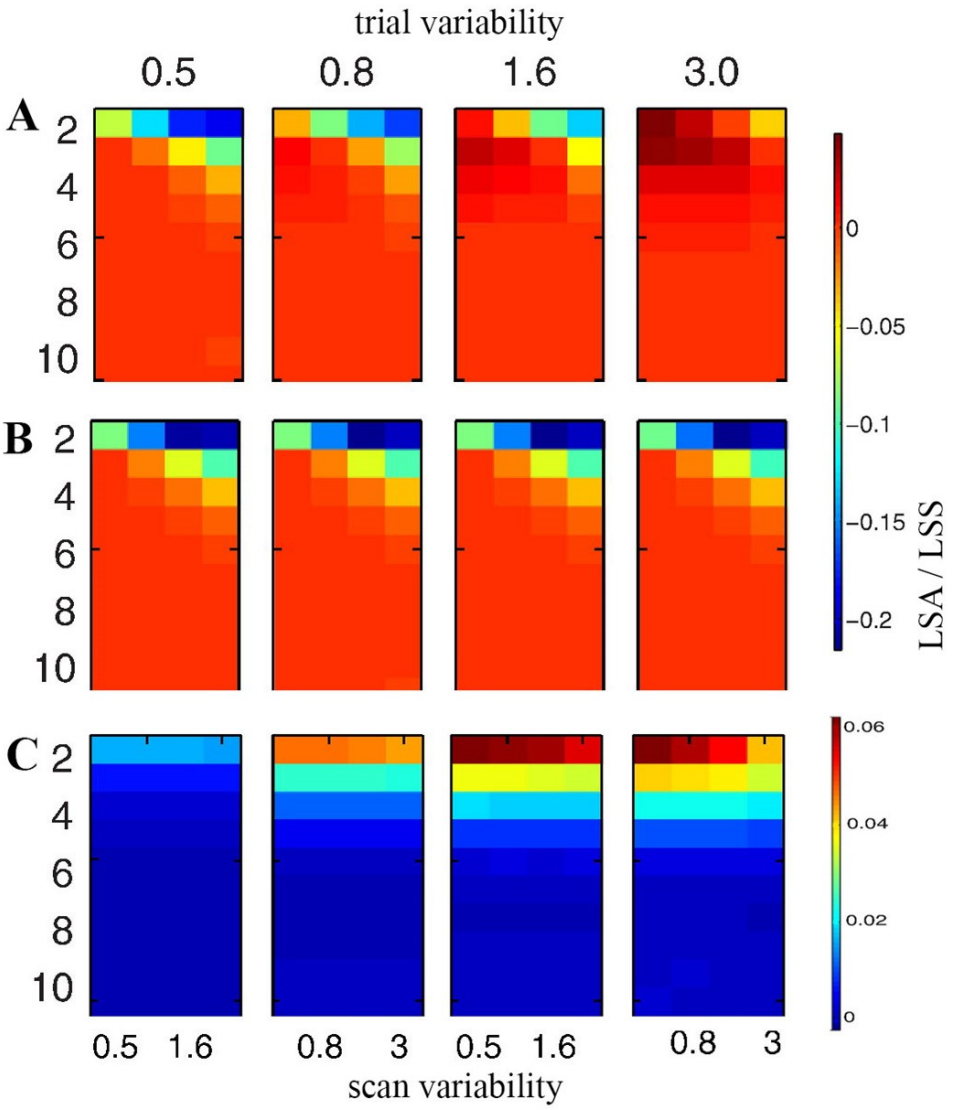


Figure 2.6, Log of ratio of LSA relative to LSS SVM classification performance (CP) for (A) incoherent trial variability and incoherent scan-variability, (B) coherent trial variability and incoherent scan variability and (C) incoherent trial-variability and coherent scan-variability. See Figure 2.2 legends for more detail. Note that only SOAs up to 10s are shown on the y-axes to clarify effects.

The reason for the interaction between LSA/LSS model and coherent/incoherent trial-variability and scan-variability (at short SOA) is exemplified in Figures 2.7 and 2.8. Figure 2.7 below is a special case where both trial- and scan-variability are coherent across the two simulated voxels, voxel1 and voxel2. Since multi-voxel classifiers take the relative BOLD response across the voxels, the simplest possible

classifier for the 2 voxels in Figure 2.7 will weight both voxels since they are equally informative for the two trial-types but in opposite directions. This difference between voxel1 and voxel2 is shown in the bottom panels of Figure 2.7. This simple linear output can separate both trial-types perfectly, despite the high level of trial and scan variability, because the variability in this case is coherent and MVPA only cares about the relation among the voxels (and thus variability cancels out). Thus both LSA and LSS are perfectly adequate in this case of coherent variability across voxels.

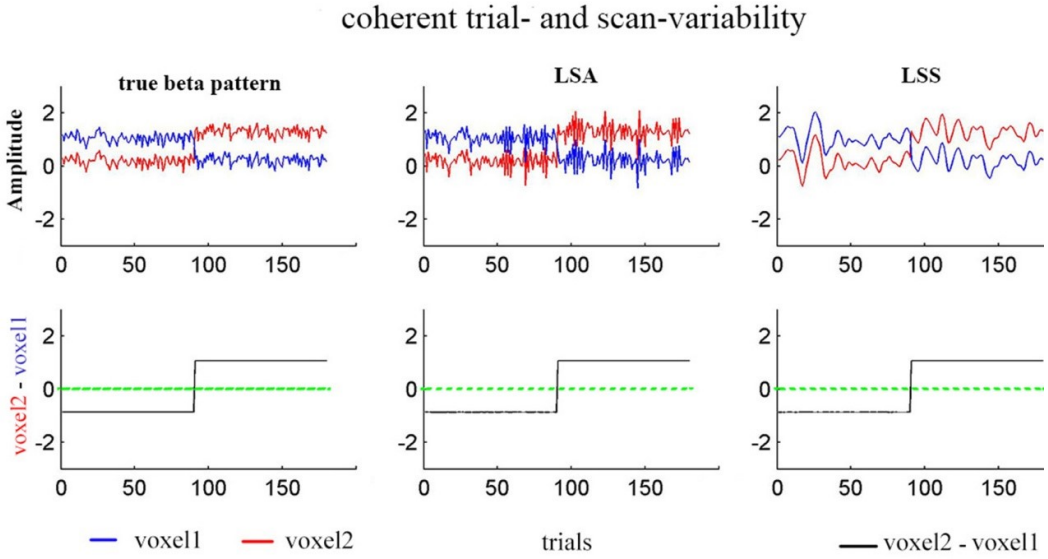


Figure 2.7. Illustration of coherent- trial and scan-variability across two voxels (SOA=2s and scan SD=0.2). Panels on top show parameters/estimates for 90 trials of each of two trial-types (trials 1-90 and 91-180 respectively) for each voxel (separate lines); bottom panels show difference between voxels for each trial (which determines CP). Left most Panels show true parameters (β_j), drawn from Gaussian with SD=0.3 and different means for each voxel. Middle and right Panels show corresponding parameter estimates ($\hat{\beta}_j$) from LSA and LSS models.

In real fMRI data however, complete coherence among the voxels is unlikely, so Figure 2.8 shows the same type of results as in Figure 2.7, but now for the three cases when trial- or scan-variability is incoherent, corresponding to the three rows in Figure 2.6.

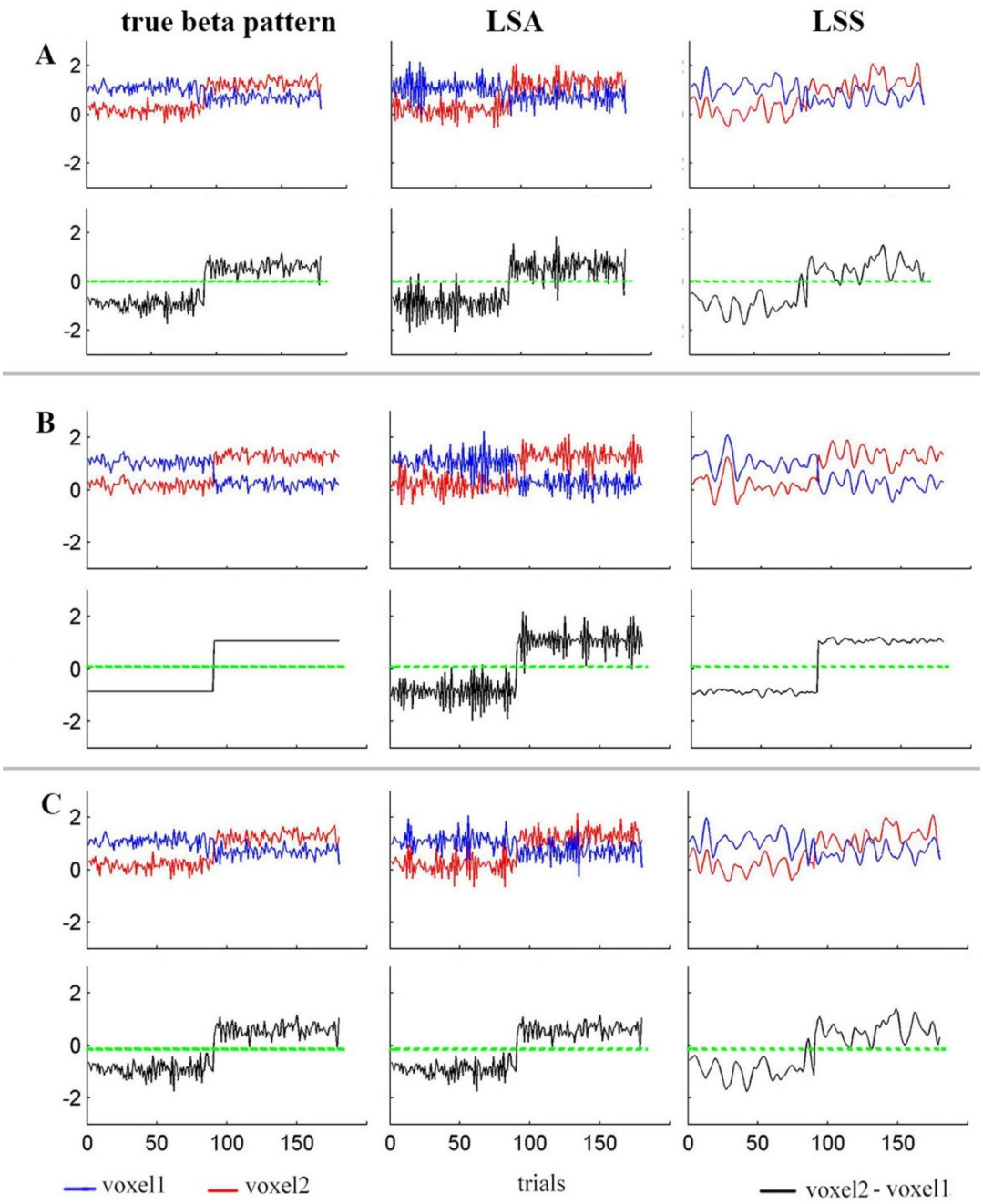


Figure 2.8, Illustration of CP performance in cases A) incoherent trial and scan variability across two voxels. B) coherent trial variability with incoherent scan variability, C) incoherent trial-variability with coherent scan variability. See Figure 2.7 legend and the text for details.

Panel A shows the case where both scan variability and trial variability are incoherent, so neither type of variability cancels out across the voxels. This means

that the relative performance of LSA to LSS performance depends on the ratio of scan-variability to trial-variability, similar to our findings for single voxel efficiency in Figure 2.6-A. Panel B shows the more interesting case of coherent trial-variability across the voxels, which cancel out when taking the difference between voxel1 and voxel2, leaving only the scan-variability, and hence LSS is always a better model regardless of the ratio of trial-variability to scan-variability. Panel C shows the complementary case where coherent scan noise cancels when taking the difference across the voxels, leaving only the trial variability, and hence LSA is always a better model.

2.7 Discussion

In this chapter, I compared the efficiency of two GLMs that are commonly used in the fMRI literature to estimate responses to single trials: LSA and LSS. The most obvious difference is that LSS produces a temporally smoother fit than LSA; this is due its lower degree of freedom, hence LSS is less flexible than LSA making it potentially more prone to under-fitting and less prone to over-fitting. As long as TR is less than SOA in an fMRI design (which is usually the case), the scan-to-scan variabilities will always be higher in frequency than the trial-to-trial variabilities, therefore, temporal smoothing is beneficial in case of high scan noise as it can attenuate the higher frequency noise by temporally smoothing it. However, this becomes problematic when there are high ratios of trial variability because it will become harder for the GLM to distinguish the scan variability from the trial variability especially at the lower SOAs.

Previous studies of fMRI design efficiency have given little consideration to the effect of trial-to-trial variability in the amplitude of responses. This variability might be random noise, such as uncontrollable fluctuations in a participant's attention, or systematic differences between the stimuli presented each trial. Through simulations, I calculated the optimal SOA and type of GLM. I summarise the main take-home messages, before considering other details of the simulations.

2.7.1 General Advice

There are two main messages for the fMRI experimenter:

1. If you care about the univariate responses to individual trials, for example for

functional connectivity using Beta-series regression (Rissman et al., 2004), and your SOA is short, then whether you should use the typical LSA model, or the LSS model, depends on the ratio of trial-variability to scan-variability: in particular, when scan-variability is higher than trial-variability, the LSS model will do better.

2. If you care about the pattern of responses to individual trials across voxels, for MVPA, then whether LSA or LSS is better depends on whether the trial-variability and scan noise is coherent across voxels. If trial variability is more coherent than scan noise, then LSS is better; whereas if scan noise is more coherent then trial variability, then LSA is better.

These results are summarised in Table 2.1.

Table 2.1 Optimal GLM model according to the noise level and types

Relative size of Variability **Coherency across Voxels**	more scan variability	more trial variability
Incoherent	LSS	LSA
Coherent	LSA	LSS

2.7.2 Unmodelled trial variability

Even if trial-to-trial variability is not of interest, the failure to model it can have implications for other analyses, since this source of variance will end up in the GLM residuals. For example, analyses that attempt to estimate functional connectivity independent of trial-evoked responses (e.g, Fair et al 2012) may end up with connectivity estimates that include unmodelled variations in trial-evoked responses, rather than the desired background / resting-state connectivity. Similarly, models that distinguish between item-effects and state-effects (e.g, Chawla et al, 1999) may end up incorrectly attributing to state differences unmodelled variations in item effects across trials. Failure to allow for trial-variability could also affect comparisons across groups, e.g, given evidence that trial-variability is higher in older adults (Baum and

Beauchamp, 2014).

Strictly speaking, unmodelled trial-variability invalidates LSU and LSS as GLMs for statistical inference within-participant (across-scans). LSA models overcome this problem, but at the cost of using more degrees of freedom in the model, hence reducing the statistical power for within-participant inference. In practice however, assuming trial-variability is random over time, the only adverse consequence of unmodelled variance will be to increase temporal autocorrelation in the error term (within the duration of the HRF), which can be captured by a sufficient order of auto-regressive noise models (Friston et al, 2002). Moreover, this unmodelled variance does not matter if one only cares about inference at the level of parameters (with LSA/LSS) or level of participants.

2.7.3 Estimating individual trials: LSS vs LSA

Since the introduction of LSS by Turner et al. (2010), it is becoming adopted in many MVPA studies. As mentioned above, LSS effectively imposes a form of regularization of parameter estimates over time, resulting in smoother "Beta series". This makes the estimates less prone to scan noise, so could be suitable for more accurate trial-based functional connectivity analyses too. However, as shown in Figure 2.4, this temporal smoothing also potentially obscures differences between nearby trials when the SOA is short (at which point LSA can become a better model). Thus for short SOA, the real value of LSS for functional connectivity analysis will depend on the ratio of trial-variability to scan-variability. This temporal smoothing does not matter so much for MVPA analyses however, if the trial-variability is coherent across voxels, because the resulting patterns across voxels become even more robust to (independent) scan noise across voxels, as shown in Figure 2.8-B.

However, when scan variability is more coherent across voxels than is trial variability, LSA is better than LSS, even when the ratio of scan noise to trial variability is high. This is because the coherent fluctuations of scan noise cancel each other across the voxels, leaving only trial variability, which can be modelled better by LSA than LSS, as shown in Figure 2.8-C. It is difficult to predict which type of variability will be more coherent across voxels in real fMRI data. One might expect trial variability to be more coherent across voxels within an ROI, if, for example, it

reflects global changes in attention (and the fMRI point-spread function / intrinsic smoothness is smaller than the ROI). This may explain why Mumford et al. (2012) showed an advantage of the LSS model in their data. However, Visser et al. (2016) found LSA to be better than LSS in their data. The latter may reflect cases where there is a high degree of coherency of scan noise across the voxels, for example owing to residual movement artifacts or other types of global scanner artifacts.

As mentioned in the previous sections, LSS attenuates the scan noise by smoothing the betas. Another possible approach is to simply first smooth the data using a low pass filter, and then use LSA to estimate the betas. This explicit way of smoothing will not work well especially for jittered SOA designs or randomized designs with more than one trial type because temporal smoothing smears and mixes different trial types together, as a result, they will become more similar to each other and classification performance will be reduced (see Appendix 3). More importantly, smoothing shrinks the mean beta toward the baseline and this shrinking becomes more obvious for longer SOAs. LSS deals with these issues better by including, additional to the trial of interest, all the other trial types in the form of N separate regressors (where N is the number of trial types) in its design matrix. These additional regressors allow the GLM to fit differences in the mean response across conditions, rather than assuming a homogeneous smoothing. 2.7.4 Caveats

In the present simulations, I have assumed temporally uncorrelated scan noise. In reality, scan noise is temporally auto-correlated, and the GLM is often generalized with an auto-regressive (AR) noise model (in conjunction with high-pass filter) to accommodate this (e.g., Friston et al., 2002). Regarding the spatial correlation in scan noise across voxels (for MVPA), this is usually dominated by haemodynamic factors like draining vessels and cardiac and respiratory signals, which can be estimated comparing residuals across voxels, or using external measurements. Future work could explore the impact on efficiency of such coloured noise sources (indeed, temporal and spatial covariance constraints could also be applied to the modelling of trial-variability in hierarchical models, Friston et al., 2002).

I have also not considered the class of regularized estimators for the GLM, such as ridge regression (Mumford et al, 2012). These tend to constrain some norm of the Beta estimates (such as the L2-norm in the case of ridge regression) – effectively

penalizing extreme estimates. This makes them more robust to scan-variability (noise), though at the expense of introducing bias (e.g, shrinking Beta estimates towards zero). Here I restricted efficiency analysis to unbiased estimators, but future work could extend efficiency comparisons (e.g, using cross-validation) to regularized estimators.

There are also more sophisticated modelling approaches than the common GLM, some of which have explicitly incorporated trial-variability, using maximum likelihood estimation of hierarchical models mentioned above (e.g., Brignell et al., 2015), or nonlinear optimization of model parameters (e.g Lu et al., 2005). Nonetheless, the general principles of efficiency, i.e, how best to estimate trial-level parameters, should be the same as outlined here.

2.8 Chapter Summary

In this chapter, I showed that SOA, scan-variability (scanner noise) and trial-variability all affect the efficiency of GLMs. When estimating responses to individual trials in a single voxel, LSU or LSS are generally a more efficient GLM than LSA when scan-variability dominates trial-variability, whereas LSA is a better GLM when trial-variability dominates. This difference between GLM efficiency is most noticeable when the SOA is short. When estimating the pattern of responses across voxels however, a second important factor is the degree of coherency of scan variability and trial variability across voxels: LSA benefits when scan variability is correlated across voxels, whereas LSS benefits when trial variability is correlated across voxels. Which is the better GLM therefore depends on both the ratio of scan variance to trial variance within voxels, and the ratio of scan covariance across voxels to trial covariance across voxels. Unfortunately these ratios are rarely known in advance for real data. However, one way to estimate efficiency in real data is to examine the stability of parameter estimates across different runs. Another way, when one expects an ROI to differ between two conditions, is to compare which GLM produces the biggest univariate and /or multivariate difference between conditions. These approaches were taken in Chapter 3.

CHAPTER 3: EMPIRICAL ANALYSES

3.1 Data description and preparation:

In Chapter 2, I used simulations to show that the optimal GLM model depends on the ratio of scan-variability to trial-variability and on the coherency across the voxels of these two types of variability. In this chapter, I compare these GLM designs using real fMRI data. For that purpose, I use an existing multimodal dataset provided by Wakeman and Henson (2015), which is freely available on "openfmri.org". In addition to its open access and prior use for methods development, this dataset has several benefits, such as 1) a large number of trials with short SOA (given that Chapter 2 showed that differences in GLM efficiencies are most noticeable at short SOAs); 2) nine independent runs (sessions) (useful for cross-validation and efficiency analyses) and 3) trials of some conditions are randomly intermixed, whereas trials from other conditions are temporally adjacent, allowing investigation of the effects of temporal smoothing from methods like LSS.

The dataset consists of 19 participants, aged 23-37, 8 female; I used a subset of 18 after removing one participant who had some scans missing from one of the fMRI runs. For each of the 9 runs, participants made left-right symmetry judgments to randomly presented images of 16 famous faces, 16 unfamiliar faces, and 16 scrambled faces. One-half of the stimuli repeated immediately, and the other half repeated after delays of 5-15 stimuli intervals. The stimuli were presented for a

random duration between 0.8-1.0s, with a random interval of 2.1-2.3s between stimuli, resulting in an SOA between 2.9-3.3s.

The experiment contained 9 conditions: 1) initial presentations of Famous faces (F_init), Unfamiliar faces (U_init), and Scrambled faces (S_init), 2) immediate repetitions of Famous faces (F_imm), Unfamiliar faces (U_imm), and Scrambled faces (S_imm) and 3) delayed repetitions of Famous faces (F_L), Unfamiliar faces (U_L), Scrambled faces (S_L). Note that one feature of this design is that immediate repetitions necessarily always follow initial presentations, so the regressors in a LSA model for trials from Initial and Immediate repetition conditions will be temporally correlated.

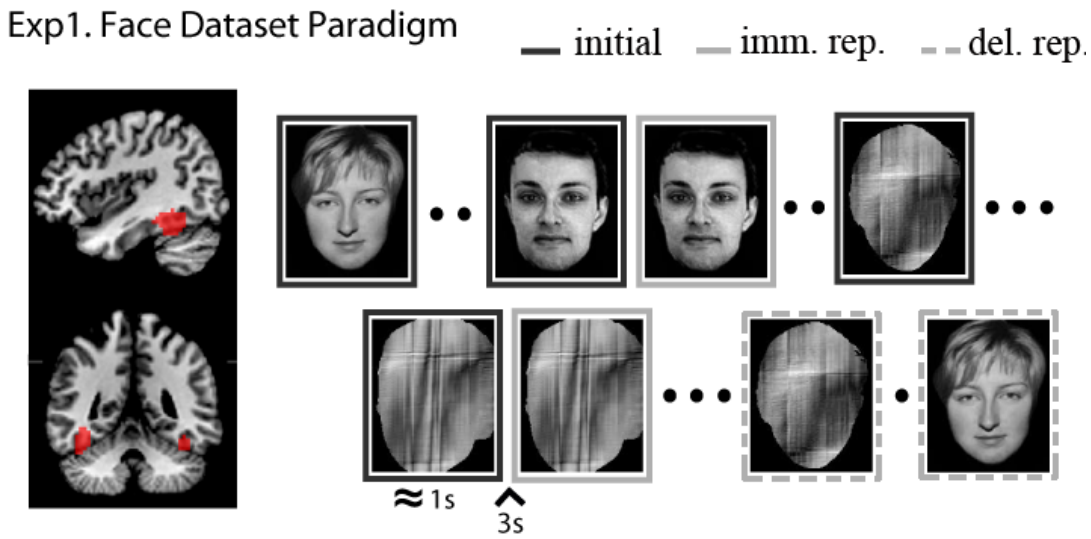


Figure 3.1: Showing the experimental design, with a thresholded Fusiform Face Area (FFA) mask from a group analyses (for more details about the experimental design, see Wakeman & Henson 2015). imm.rep. = immediate repetition, del.rep = delayed repetition.

3.1.1 fMRI acquisition:

The MRI data were acquired with a 3T Siemens Tim-Trio MRI scanner (Siemens, Erlangen, Germany). The fMRI data came from a (gradient) echo-planar imaging (EPI) sequence of 33, 3 mm-thick axial slices, with TR of 2000 ms, TE of 30 ms and flip angle of 78°. 210 volumes were acquired for each of the 9 runs. Slices were acquired in an interleaved fashion, with odd then even numbered slices (where slice 1 was the most inferior slice). The distance factor controlling the gap between slices

was adjusted for each participant to ensure whole cortex coverage, resulting in a range of voxel sizes of 3×3×3.75 mm to 3×3×4.05 mm across participants (for more details, see Wakeman & Henson et al. 2015). A T1-weighted structural image of 1x1x1mm resolution was also acquired using a MPRAGE sequence.

3.1.2 fMRI pre-processing:

The fMRI data were pre-processed using the SPM12 software package (www.fil.ion.ucl.ac.uk/spm) in Matlab 2012b (uk.mathworks.com). After removing the first two scans from each session to allow for T1 saturation effects, the functional data were corrected for the different slice times, realigned to correct for head motion, and coregistered with the structural image. The structural image was segmented and normalised to a standard MNI template, and the normalisation warps applied to the functional images. These were finally spatially smoothed using a Gaussian filter of 8 mm FWHM for mass univariate statistics. However, for all MVPA analyses, we used the unsmoothed data, in line with the previous literature (Haxby et al., 2001).

Then for the subsequent analyses, I compared separate GLMs that conformed to either LSU (when only average betas where concerned), or LSA or LSS models when estimating individual trials (the LSS model was an LSS-9 model; see footnote in Chapter 2). The regressors were created by convolving a delta function at the onset of each stimulus with a canonical HRF. All GLMs also included six motion regressors to capture residual (linear) movement artefacts, plus a constant term.

3.1.3 ROI analyses and contrasts of interest

To create the ROIs, a contrast was created for the LSU model, which averaged the mean betas for each condition across the 9 runs. A between-participant (2^{nd}-level) GLM was then estimated, using the 9 contrast maps for each of the 18 participants, together with participant effects (conforming to a repeated-measures ANOVA with a pooled error, Henson & Penny, 2003). This GLM was used to estimate the population mean betas, and the statistical significance of contrasts between these means were thresholded by controlling the Family-Wise Error (FWE) of $P<0.05$ across the whole-brain.

The FFA was defined as those suprathreshold voxels for the T-contrast of Faces > Scrambled (averaging across famous and unfamiliar and initial and repeated presentations) that fell within a cluster centred in left or right fusiform. These were combined to form a single bilateral FFA ROI containing 185 voxels (135 spatially continuous voxels for right FFA and 52 voxels for left FFA) (Figure 3.1 left panel). Note in this chapter I only cared to compare the efficiency of various GLMs in reproducing consistent results or patterns across independent fMRI runs, therefore, the difference in the BOLD response between F_init and S_init in a "Faces > Scrambled" contrast does not bias our efficiency analyses toward any specific GLM because the absolute value of this difference is not our primary focus (we do not know the true difference anyway). To put more concretely, we do not care what value a specific GLM produces for "F_init - S_init" but we care that an efficient GLM should reproduce the same value again when applied to similar but independent fMRI runs.

I focused on two pairwise comparisons of the 9 conditions: 1) the difference between "F_init - S_init", since these conditions were randomly intermixed (as assumed in Chapter 2), and 2) "F_init - F_imm", since these conditions were always adjacent, which may impact the relative efficiency of LSS versus LSA, given the temporal smoothing of LSA discussed in Chapter 2. Note that because only half of the repetitions were immediate, I randomly selected one half of the F_init estimates to match trial numbers with F_imm estimates (and so this contrast has half as many trials as the F_init – S_init contrast).

3.2 Evidence for trial-to-trial variability in real data

Before comparing the efficiencies of different models, I wanted to test whether there is evidence of significant trial-variability in these data. I therefore created a single GLM in which all 9 conditions were modelled both by an LSU partition and an LSA partition (see Figure 3.2-A). By using an F-test on the regressors from the LSA partition, one can test whether there is significance explained by allowing for variability across trials (relative to residual scan noise), over-and-above differences between conditions in the mean responses across trials.

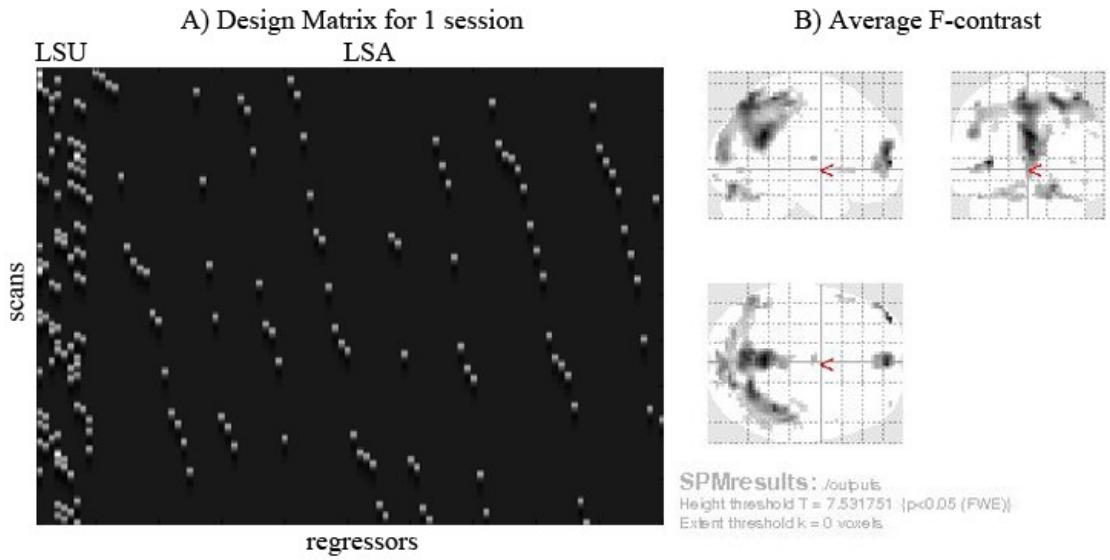


Figure 3.2 shows the custom design matrix [LSU LSA] and the resulted average F-contrast (log-transformed) for LSA showing extensive variabilities not captured by LSU.

Images of the F-value from the LSA partition, averaged across runs, were log-transformed (so an F-value of 1 became 0) and entered into a one-sample t-test across subjects to test for consistent effects across subjects (note this is not a conventional random effects analysis on a parameter, but a meta-analysis on a statistic). The one-tailed t-test for F-values consistently above 1 was then thresholded with FWE $p<0.05$ (Figure3.2-B). Several brain regions in lateral and medial parietal cortex, as well as medial prefrontal cortex, showed evidence of significant trial-variability. Interestingly, the FFA was not apparent at this threshold. However, the absence of FFA may reflect noisy individual voxels that did not pass the FWE correction. If the trial variabilities are coherent across the voxels, then averaging across the voxels in FFA will enhance sensitivity. Indeed, when the same analysis was repeated by averaging over voxels in the FFA mask, there was significant evidence of trial-variability in FFA too, $t(17)=12.12$, $p<.001$.

3.3 ROI - Efficiency Analyses

Given evidence that trial-to-trial variability exist in these data, I compared LSS and LSA models in their efficiency of modelling this variability within the FFA, starting with univariate measures of 1) mean activity over all voxels the ROI, and 2) bivariate

correlations between each voxel in the ROI. Note that in all the following statistical reports, correlation coefficient values were transformed using Fisher-z transformation and the standard deviations (SDs) were log transformed before running t-tests to compare the GLMs. Figure 3.3 simplifies the concepts behind some of the efficiency measures I used in this chapter. I will refer back to the relevant panels of this figure in each of the following sections.

3.3.1 Optimal GLM for Univariate Analyses

Estimating efficiency for real data is difficult because, unlike the simulations in our previous chapter, we do not know the true Beta of each condition. One way to estimate efficiency is to assume that the most efficient model is the one that produces the most stable beta estimates across the independent runs (assuming there is no true variability across runs, e.g, owing to participant fatigue).

3.3.1.1 Reproducibility of the average trial response across runs

I started with the mean estimate after averaging across all trials in each condition (like the PPM measure in Figure 2.2 of Chapter 2). I then averaged these condition means over trials, and calculated the difference between "F_init - S_init" as an example of randomised trials belonging to different stimulus classes, and "F_init - F_imm" as an example of adjacent trials belonging to the same stimulus class. I then estimated the standard deviation (SD) of this difference across runs (Figure 3.3-A), such that a lower SD indicates more stable estimates. Since SDs are skewed and always positive, I log-transformed the results before comparing the means.

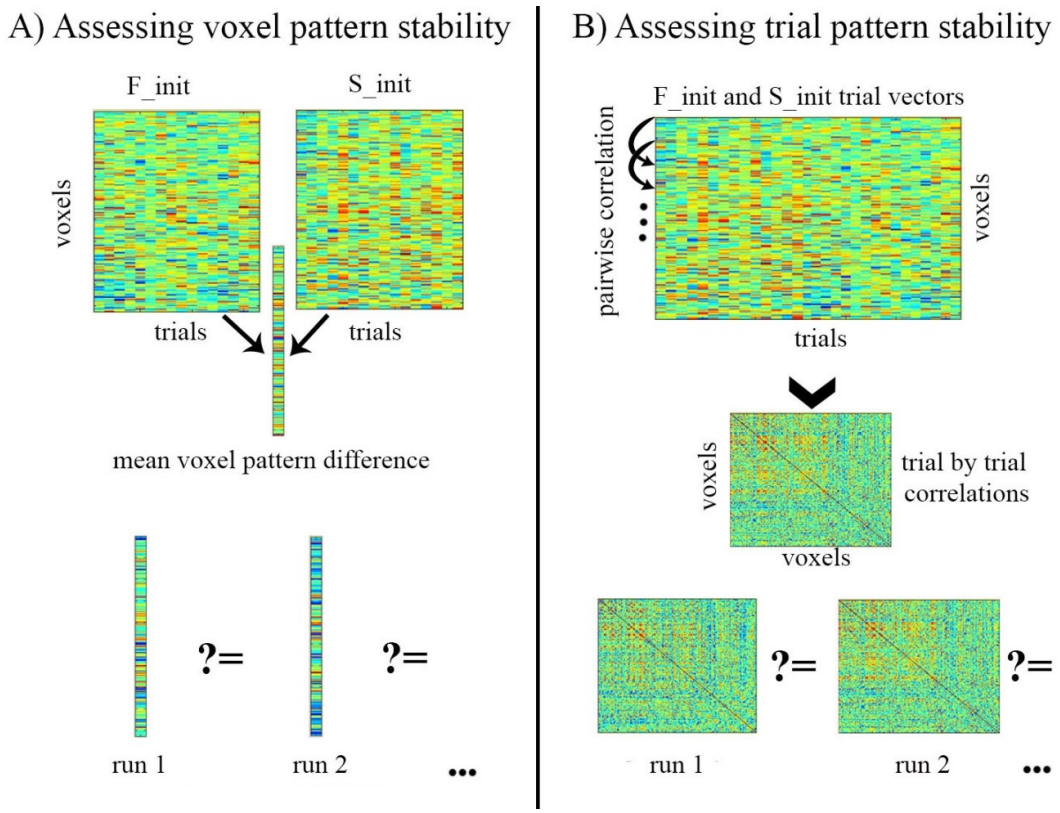


Figure 3.3 illustrates the efficiency measures through A) assessing voxel pattern stability across the runs, first, by measuring the mean pattern difference between the two stimulus types for each run, then these measures are compared across the independent runs to assess the stability of the voxel values across the runs by measuring either their SDs (univariate – section 3.3.1.1) or pattern similarity using correlation (multivariate – 3.3.2.1), B) assessing trial pattern stability across the runs, first, by correlating the trial vectors among all the voxels to measure the trial pattern coherency within FFA, then these coherency matrices are compared among the runs using SD (section 3.3.1.2).

Paired t-tests showed that SD was much lower for LSS than LSA for both "F_init-S_init", $t(17)=14.81$, $p<0.001$ (Figure 3.4-A) and "F_init-F_imm", $t(17)=16.70$, $p<0.001$ (Figure 3.4-B), suggesting that LSS is a more efficient model (and therefore that the ratio of trial-variability to scan-variability was relatively low). Because this analysis estimates the average Beta across the trials, I added the standard LSU for the sake of comparison. Similar to LSS, the SD of LSU was significantly lower than LSA ($t(17)=15.4$; $p<0.001$ for F_init - S_init; $t(17)=18.5$, $p<0.001$ for F_init-F_imm). LSS and LSU were numerically very similar for F_init - S_init, but showed a significant difference nonetheless, with LSU being better than LSS $t(17)=3.9$,

p=0.001 (consistent with simulations in Appendix 2). Interestingly, the opposite was found for F_init - F_imm, where LSS was better, $t(17)=4.9$, $p<0.001$, most likely reflecting the fact that F_init and F_imm trials were adjacent to each other. However, because LSU was not the primary focus, I did not explore this effect further.

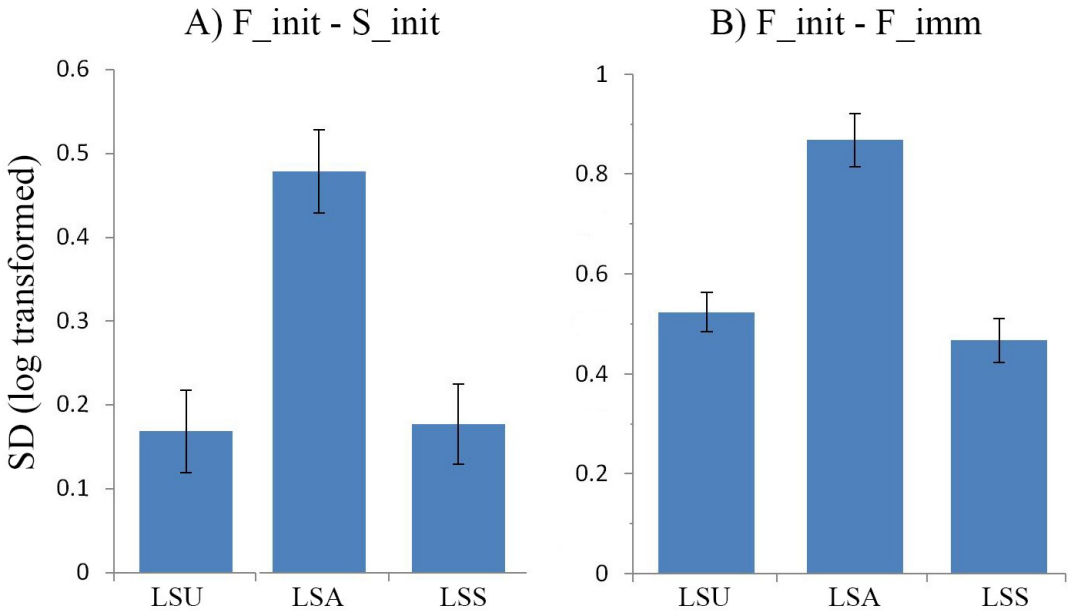


Figure 3.4 Shows the optimal GLM for the average trial responses using standard deviation of the beta estimates differences across the runs (lower is better). Error bars are 95% CI given between-participant variability, so can overlap even if within-participant differences are significant in a paired t-test.

3.3.1.2 Reproducibility of individual trial responses across runs (trial by trial correlations)

I then looked at the temporal correlation between single-trial estimates, i.e, the "Beta series" correlation sometimes used as a measure of functional connectivity in fMRI (Rissman et al., 2004, and considered in Figure 2.3 of Chapter 2). We do not know the true functional connectivity in these data. However, if we assume that the FFA is functionally homogeneous, then all voxels should show similar trial estimates, i.e, high correlations of Beta series between each voxel. As in the previous section, I constrained analyses to F_init and S_init trials and F_init and F_imm trials. The results are shown in Figure 3.5, where the mean correlation was higher under LSS

than LSA (t(17)=8.31 , p<0.001 for F_init and S_init trials; t(17)=7.89, p<0.001 for F_init and F_imm trials).

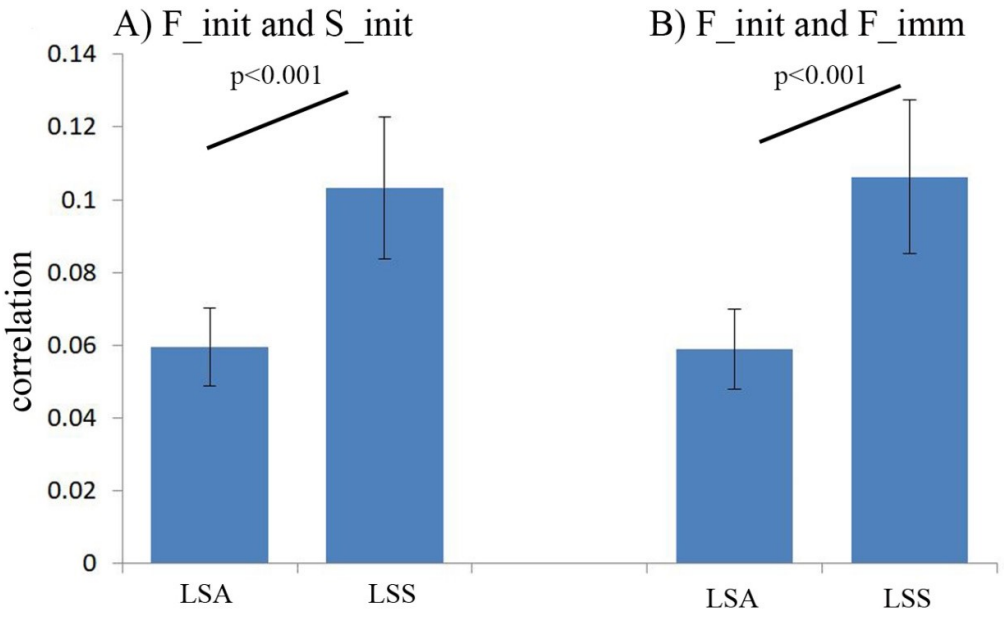


Figure 3.5: Shows the results of Beta series correlations in FFA for both GLMs.

However, one does not have to make this assumption of functional homogeneity within FFA, because one can also test how reproducible those voxel-to-voxel correlations are across runs (see Figure 3.3-B). Having calculated the 185-by-185 matrix of voxel correlations, I computed the standard deviation of each correlation across the 9 runs and averaged these SDs across all elements in the matrix to produce the values in Figure 3.6 (top panels). Paired t-tests showed that LSS produced higher SD than LSA for F_init and S_init trials, t(17)=5.90, p<0.001, i.e, less reproducible estimates. The same pattern was seen for F_init and F_imm, t(17)=8.83, p=0.001. It is possible that the lower SD of correlations across runs for LSA is a consequence of the lower mean correlations overall (Figure 3.5). I therefore also normalised the SD by the mean correlation for each subject (i.e, calculated the coefficient of variation; Figure 3.6 bottom panels), and now there was no longer any significant difference between LSS and LSA (t(17)=0.51, p=0.61 for F_init and S_init; t(17)=0.70, p=0.48 for F_init and F_imm). Thus the evidence for the better GLM for Beta-series correlation in these FFA data is moot.

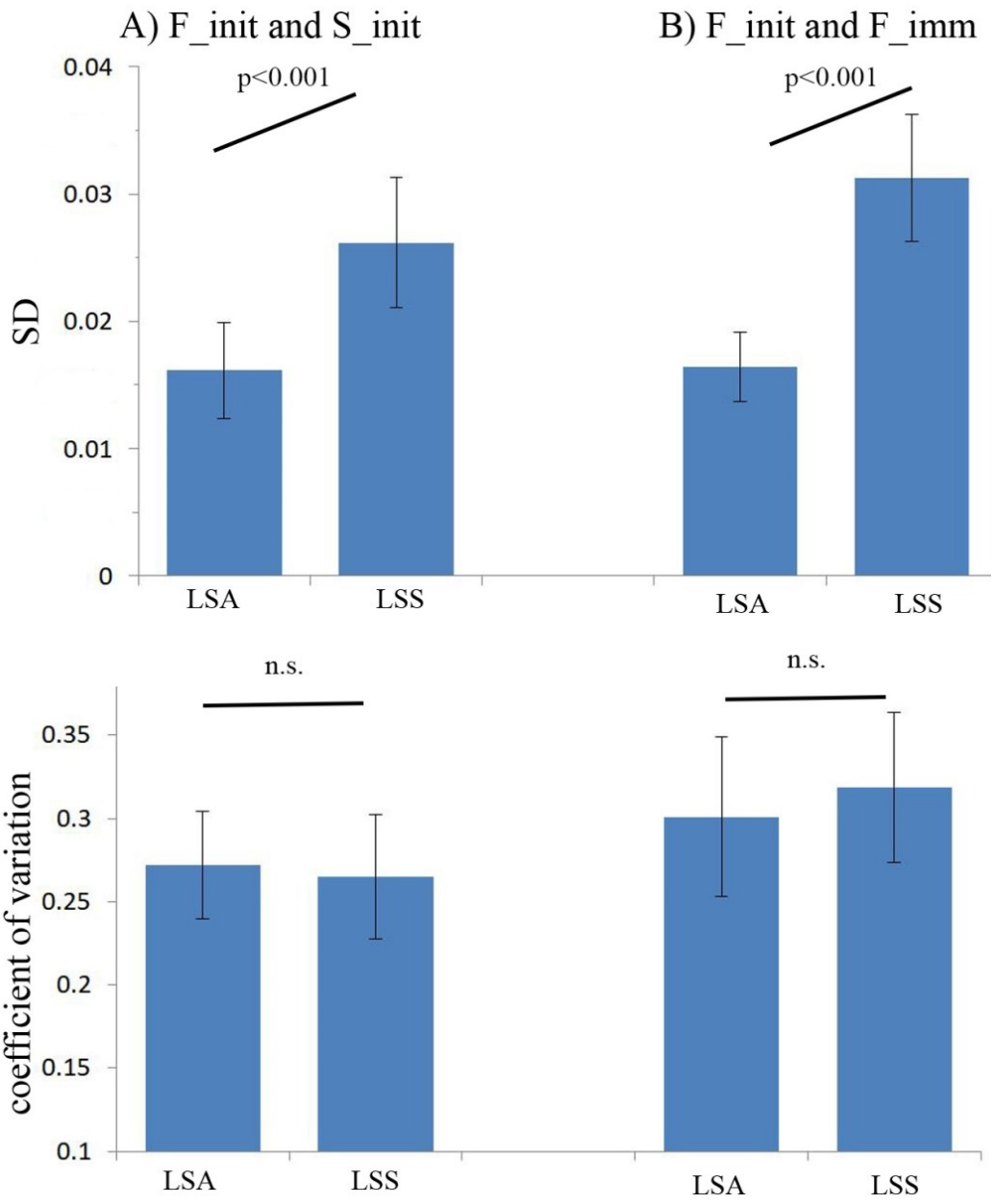


Figure 3.6 Upper panels show the SD of Beta series correlations between all pairs of voxels in FFA mask. Lower panels show the normalised SD by the mean correlations in Figure 3.5. (error bars = 95% CI).

3.3.2 Optimal GLM for Multivariate Analyses

3.3.2.1 Reproducibility of voxel patterns across runs

In the previous sections, I assessed the efficiency of univariate voxel estimates, either averaged across trials or for each trial separately. In this section, I assess the

efficiency of estimating spatial patterns across voxels, averaged across trials (stimuli) of the same type (because individual stimuli cannot be matched directly across runs). This assumes that each trial-type has a canonical pattern that should be identical across runs (e.g, that even though every face might produce a different pattern across FFA, they share an average pattern that is different from scrambled faces).

Like in Section 3.3.1.1 and Figure 3.3-A above, I estimated the mean voxel pattern difference for condition pairs, then instead of examining the stability of the individual voxel mean differences across the runs through the SD, I measured the similarity among the voxel pattern differences across the run pairs using correlation (higher similarity among the run pairs means higher reproducibility of the voxel pattern, regardless of the individual absolute voxel values). The average results for all the subjects are shown in Figure 3.7.

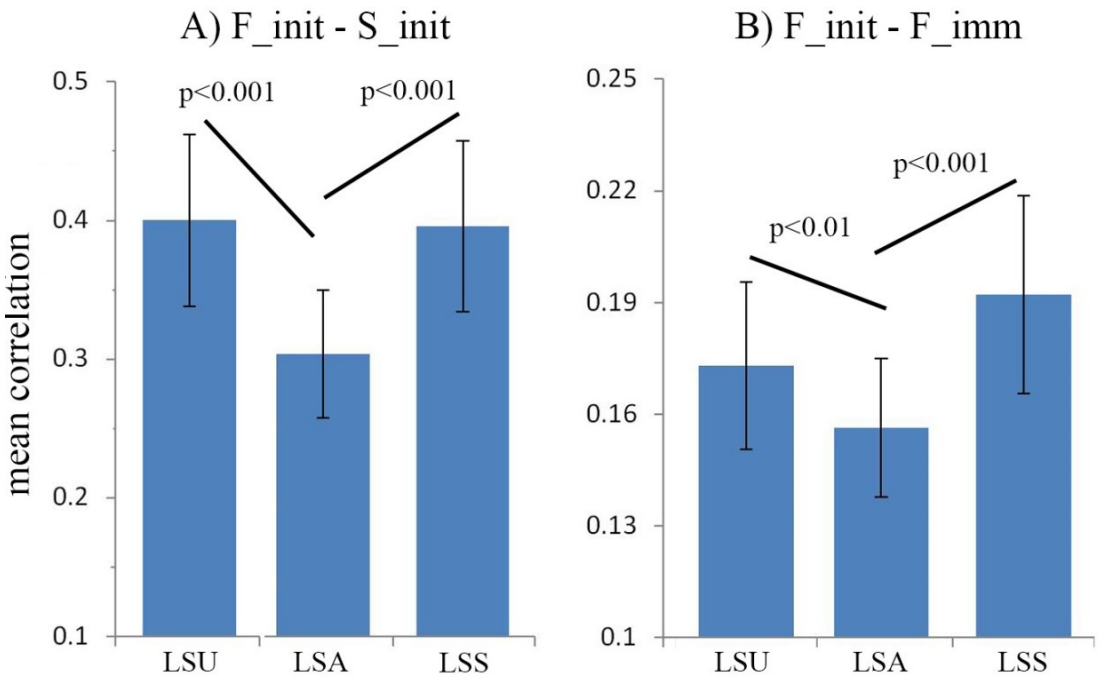


Figure 3.7: Compares the similarity of voxel pattern differences across the independent runs for each GLM (higher is better). Errors are CI 95%. Again note that error bars can be misleading for LSS and LSU as this is a paired t-test (see the text for the stats)

LSS showed higher correlations across the runs in both cases, $t(17)=6.88$, $p<0.001$ for F_init –S_init and $t(17)=5.45$, $p<0.001$ for F_init – F_imm. Since these analyses involve the average Beta values, I again added LSU for the sake of comparison. Again, like LSS, LSU produced greater similarity of the voxel pattern across the runs than LSA; however, compared to LSS, LSU was significantly higher for F_init – S_init, $t(17)=5.51$, $p<0.001$ (despite differences only in the 3rd decimal position). In contrast, LSS was significantly better than LSU for F_init – F_imm, $t(17)=3.28$, $p=0.004$.

3.3.2.2 The ability to discriminate different stimuli classes

Another way to estimate the efficiency of each GLM for pattern analysis is to compare the ability of their Beta estimates to separate the different trial classes, using separate estimates of each trial (rather than averaging over trials, as in the previous section). Here, I used binary classification using the same linear SVM as in the previous chapter (e.g., Figure 2.5). I used a linear SVM classifier provided from the Bioinformatics Toolbox in Matlab 2012b and leave-one-out cross validation across the 9 runs.

Figure 3.8 shows SVM classification results for each GLM. The accuracy was generally lower for F_init vs F_imm than for F_init vs S_init for both LSS and LSA. This may be because distinguishing initial versus repeated presentation of the same face is more difficult than distinguishing faces from scrambled faces, or because there were half as many trials for the F_init versus F_imm condition (see Section 3.1.3).

For F_init vs S_init classification (Figure 3.8 A), LSS performed significantly better than LSA $t(17)=8.21$, $p<0.001$. For F_init vs F_imm classification (Figure 3.8-B), there was no significant difference between the performance of LSA and LSS, $t(17)=0.11$, $p=0.92$.

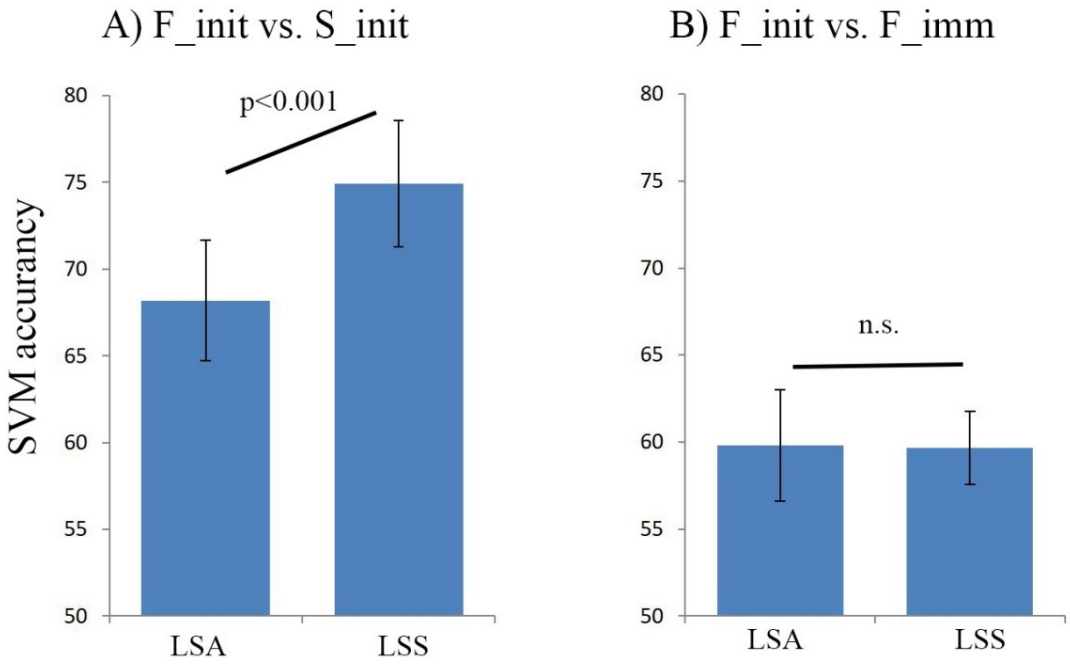


Figure 3.8: Compares the trial-wise classification accuracy among the GLMs using Linear SVM in FFA for A) F_init vs S_init and B) F_init vs F_imm. (Higher is better, error bars = 95% CI).

3.4 Whole brain - Efficiency Analyses

All our results so far were specific to FFA. While FFA is an important ROI for this dataset, the relative efficiency of GLMs may vary across brain regions, owing to different ratios of trial-variability to scan-variability. I therefore repeated the above SVM classification performance using a search-light procedure across the whole brain (excluding cerebellum and brain stem) (Kriegeskorte et al., 2006). A spherical searchlight of 268 voxels (radius of 4 voxels) was centred in turn on all 55221 voxels in the normalised brain images (voxels in spheres that extended outside the brain were excluded). The resulting SVM classification accuracy was smoothed with an isotropic Gaussian of 8 mm FWHM (to render the data in each voxel more Gaussian across participants), entered into one sample t-test across participants versus the chance level of 50%, and thresholded at FWE $p<0.05$.

Figure 3.9 shows the results for "F_init vs S_init". Large clusters in the visual cortex were seen for both GLMs. When comparing the GLMs directly with a paired t-test, LSS showed significantly higher classification than LSA in these regions (Figure 3.7

bottom panels). No voxels showed the opposite pattern of significantly better classification for LSA than LSS.

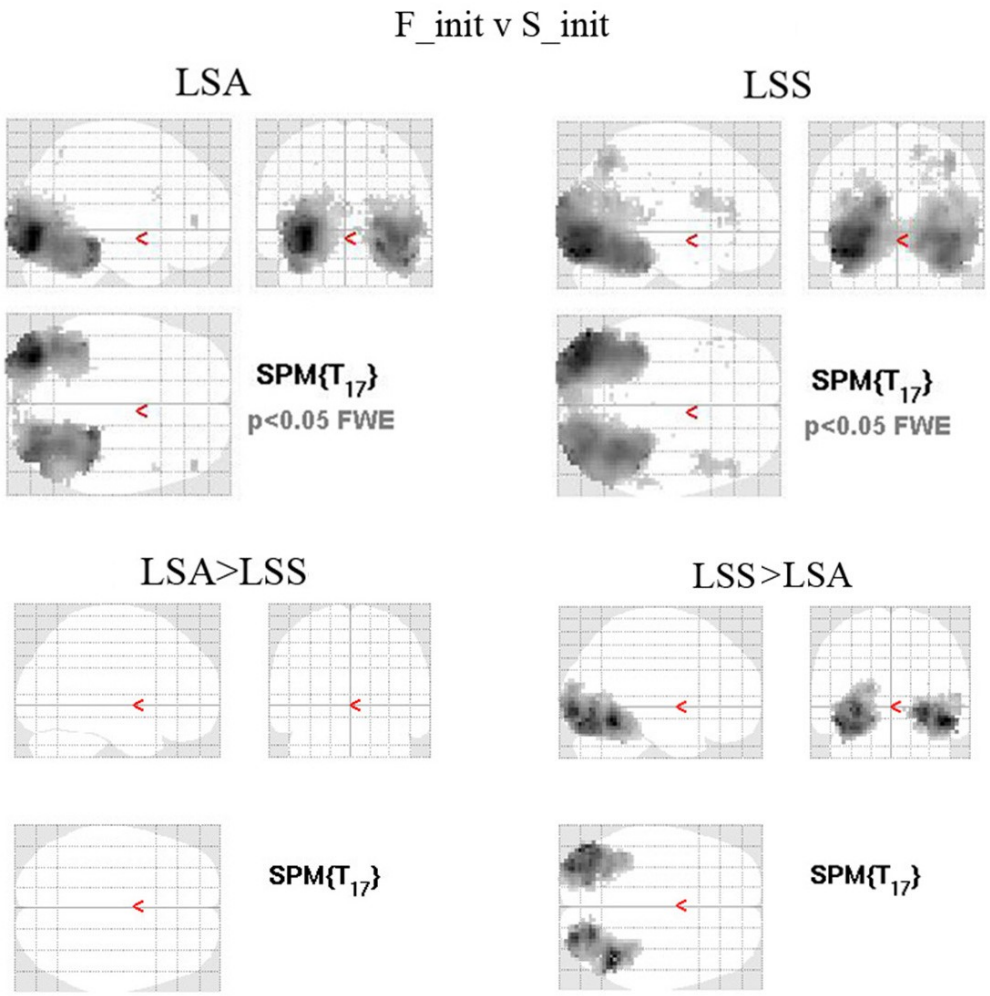


Figure 3.9: Searchlight analyses of CP across subjects (p<0.05 FWE) for F_init vs S_init. Upper panels show one-sample t-tests vs chance classification; lower panels show paired t-test for LSA vs LSS.

For "F_init vs F_imm", the classification performance was generally low, so only a few clusters survived the threshold of p<0.05 FWE. When I lowered the threshold of p<0.001 uncorrected (Figure 3.10 upper panels) for further exploration, several parietal and prefrontal clusters emerged (regions associated with repetition enhancement and possibly explicit memory for the repeat). Direct comparison at this threshold again showed some clusters in ventral temporal regions (Figure 3.10 bottom panels) for which LSS was again the better model, but these were not close to

regions of interest (and may be type I errors), so strong conclusions cannot be drawn.

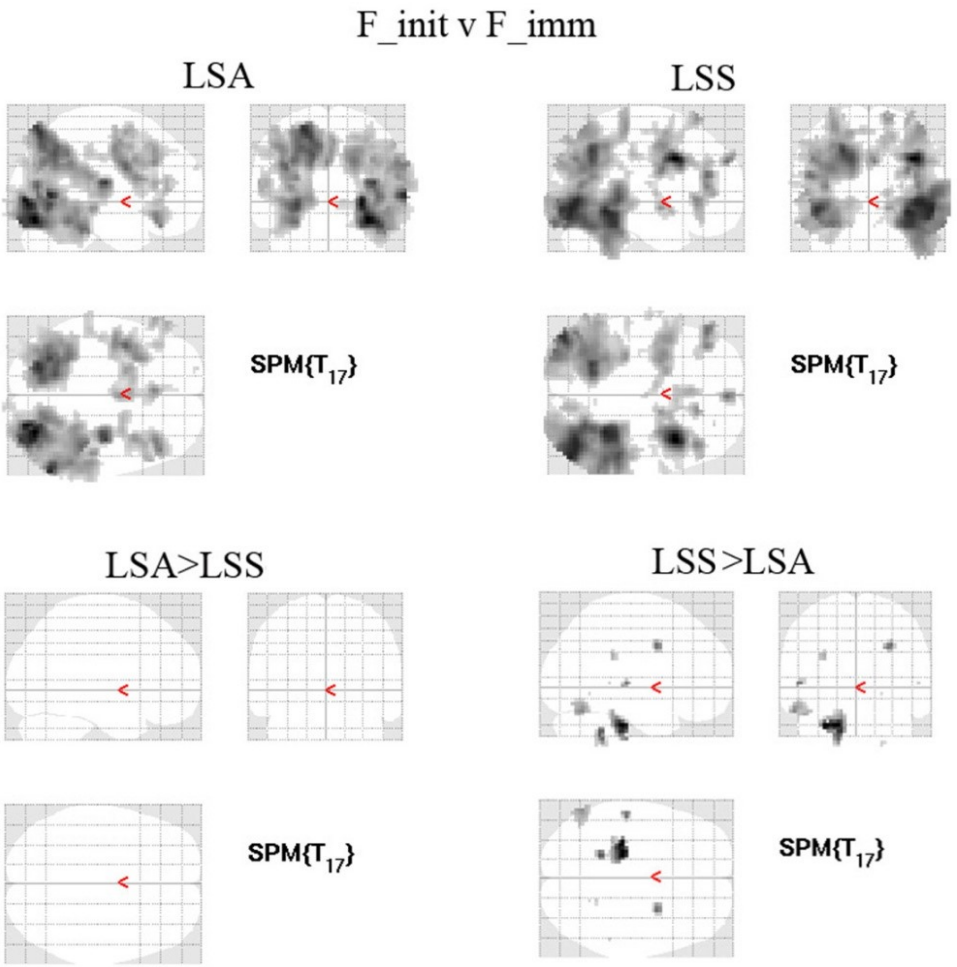


Figure 3.10 Searchlight analyses of CP across subjects ($p<0.001$ uncorrected) for F_init vs F_imm. See Figure 3.9 legend for more details.

3.5-General discussion

The first aim of this chapter was to demonstrate that significant trial-variability exists in real data, and hence this source of variability cannot be ignored as it has been in previous calculations of GLM efficiency for fMRI designs (e.g, Josephs & Henson, 1999; Friston et al, 1999; Dale, 1999). For this purpose, I used a well-known data that presented faces and scrambled faces at a short SOA (approximately 3s). Several brain regions showed evidence for such trial-variability, even when correcting for multiple comparisons across the brain. When focusing on the FFA, there was also evidence for trial variability within that ROI.

Given such trial-variability, the simulations in Chapter 2 showed how a GLM with high degrees of freedom (LSA) models the trial-variabilities better at the expense of over-fitting the scanner noise (scan-variability), whereas a GLM with fewer degrees of freedom (LSS) can be more robust the scan-variability at the expense of under-fitting trial-variability (by virtue of temporal smoothing). Here, I compared these LSA and LSS models on real data. Since we do not know the true parameter values in real data, I employed two main efficiency measures: 1) the stability of the trial estimates across the independent runs, 2) the ability to discriminate between various trial-classes. To calculate the stability of estimates across runs, I measured SD of 1) the average difference in response to two conditions, averaged across ROI voxels and trials, 2) correlations across voxels of individual trial estimates, and 3) correlations across conditions in the voxel patterns. For the ability to discriminate between trial classes, I used a binary SVM classifier. In all cases, I examined two contrasts: a) initial presentation of famous faces versus scrambled faces (F_init vs S_init), which were randomly intermixed (corresponding to the assumptions made in Chapter 2), and b) initial versus immediately repeated famous faces (F_init vs F_imm), which are temporally adjacent, and therefore be affected by the temporal smoothing employed by LSS.

In general terms, LSS appeared to be more efficient than LSA (like Mumford et al., 2012, but unlike Visser et al., 2016). This was certainly the case for the average difference in response to two conditions and the stability of their voxel pattern difference across the runs (Figures 3.4 and 3.7) and SVM classification of F_init vs S_init (Figure 3.8). Interestingly, this LSS advantage was not significant for classification of F_init vs F_imm, which may reflect the cost of temporal smoothing, but could also reflect the fewer number of trials and the more general difficulty of classifying initial versus repeated presentations of the same stimulus class (versus distinguishing two stimulus classes).

The estimates of efficiency based on Beta-series correlation was less clear cut: while LSS produced higher correlations between voxels on average, which might be appropriate if all voxels in the ROI respond the same way, it also produced higher SD across runs, i.e, less stable correlations. Nonetheless, this lower SD for LSA could be an artefact of the lower mean correlation for LSA, and when testing the

coefficient of variation, LSS and LSA no longer differed significantly.

The better performance of LSS in the univariate results, averaged across trials in each condition, suggests that scan-variability was higher than trial-variability in the FFA (see Figure 2.2 in Chapter 2). The better performance of LSS in the multivariate results then suggests that trial-variability may additionally be more coherent across voxels than scan-variability.

Finally, we used the searchlight analyses to see if our FFA results generalise to other brain regions, e.g, if the ratio of trial-(co)variance to scan-(co)variance differs across the brain. For the contrast of randomised trials (F_init vs S_init), LSS was significantly better in many posterior brain voxels, while no voxel showed an advantage for LSA that survived correction for the multiple comparisons. This suggests that most of the brain areas (those showed CP above the chance level) exhibit greater scan-variability than trial-variability in this dataset, and/or that the trial-variability is more coherent between nearby voxels than scan-variability.

One caveat with the above analyses is the assumption that the true parameters are stable across runs. It is possible that factors like fatigue affected trial-variability across successive runs (or even that scanner noise changed with time), and therefore affected the present results. However, without knowing the true values, it is difficult to test this assumption directly.

3.6- Chapter Summary

In this chapter, I used empirical analyses to compare the efficiency of GLMs using various efficiency measures to parallel the measures used for the simulations in Chapter 2. For the average trial response, LSS yielded comparable results to the standard LSU. For single trial responses, LSS provided better estimates than LSA in most cases, possibly because this dataset exhibits more scan-variability than trial-variability and/or more coherent trial-variability than scan-variability. For these reasons, I use LSS for the rest of the analyses of this dataset in this thesis, despite its greater computational demands.

CHAPTER 4: RS-RELATED CHANGES IN FMRI / EXPERIMENT 1

4.1 Introduction

As mentioned in Chapter 1, many fMRI studies have reported a reduction in the BOLD response for repeated versus initial presentations of stimuli (Grill-Spector et al., 2006; Henson, 2003). Most of these studies focus on ROIs that respond strongly to the stimuli of interest (e.g, identified in a separate localizer scan, when compared to various control stimuli) and average the BOLD response across all voxels within each ROI. Other studies adopt a mass univariate search across the whole brain, identifying clusters of voxels where RS is significant, and these clusters do tend to overlap with regions that respond strongly to the stimuli of interest (Avidan et al.,2002; Weiner et al., 2010). For instance, the Fusiform Face Area (FFA) shows strong RS for repeated faces (Henson et al., 2002; Fang et al., 2006), Lateral Occipital Cortex (LOC) exhibits strong RS for repeated objects (Pourtois et al., 2008; Hatfield et al., 2016), and V1 shows strong RS to repeated visual gratings (Sapountzis et al., 2010).

The fact that RS tends to be greatest in ROIs that respond most strongly to the stimulus being repeated (relative to another stimulus type) is easy to model in terms of simple multiplicative scaling since if $R_2 = cR_1$, where R_1, R_2 are responses to first

and second presentations respectively and $0 < c < 1$, then $RS = R_1 - R_2 = (1-c)R_1$, i.e., RS scales linearly with R_1.[2] The univariate effects of repetition on voxels' mean amplitude have been investigated thoroughly in fMRI studies. However, fewer studies have examined the multivariate effects of repetition on patterns across voxels. The majority of those fMRI studies that have used MVPA have tried to correlate the findings with behavior, or compared multivariate results with univariate results (Ward et al., 2013; Xue et al., 2010; Sapountzis et al., 2010; Rissman et al., 2010; Hatfield et al., 2016).

An fMRI study more relevant to present concerns modulated repetition as a function of expectation (Kok et al., 2012), and found that classification performance (CP) between two different visual gratings was inversely related to the amount of suppression, suggesting that the selectivity between the two visual gratings was increased by BOLD suppression, which they interpreted as evidence for the neuronal sharpening hypothesis (Chapter 1). However, CP depends on how BOLD suppression affects the patterns for stimuli from the same class relative the patterns for stimuli of other classes. However, to my knowledge, no study so far has unpacked CP thoroughly into within- versus between-class similarity. Kok et al. also ranked V1 voxels by their selectivity to one of two orientations of a visual grating, and found that the least selective voxels were actually suppressed more than the most selective voxels (note again that they manipulated expectancy rather than repetition – a point I return to in the Discussion – but this does not matter for the purpose of the present argument). In contrast, Weiner et al. (2010), who used short lag and long lag repetition paradigms, reported the opposite in face-selective ROIs, i.e., the amount of RS was greater for the preferred stimulus category compared to non-preferred stimulus category.

Weiner et al. (2010) also conducted more detailed modelling that simulated scaling or sharpening at the level of neurons within a voxel, and related the degree of RS to the size of the initial response and to the degree of selectivity at the level of voxels. They found that neither scaling nor sharpening was sufficient to explain the RS-

[2] Some studies define RS proportionally rather than additively, i.e, in terms of an adaptation ratio $RS = R_2 / R_1$ (e.g, Grill-Spector et al.,2006; Rossion et al., 2008), for which RS would be independent of initial response. However, the studies reviewed here do not use this definition.

selectivity relationship in all their ROIs and paradigms, so they concluded that multiple adaptation mechanisms exist. However, I will show in Chapter 5 that their simulations of neuronal scaling and neuronal sharpening were too simplistic, by virtue of assuming that all neurons are suppressed by the same amount, regardless of their selectivity for the stimulus in question. Yet we know from single-cell studies that the degree of neuronal suppression depends on the difference between the neuron's preferred stimulus and the repeated stimulus (e.g, Kar & Krekelberg., 2016; see also Chapter 1). Furthermore, I will show that it is not sufficient to consider one property of fMRI repetition effects (such as relating RS to selectivity): only by simultaneously considering multiple features of the data can the underlying neuronal models be distinguished. In the current chapter, I will formally define six repetition-related changes in fMRI signals across voxels within an ROI, and test their significance in the face dataset I described in the previous chapter.

4.2 Methods

I chose LSS single trial estimates for the current analyses because the previous chapter suggested that they were more stable than LSA estimates for the face dataset. For simplicity, I ignored the delayed repetition estimates and focused on the immediate repetitions to avoid potential interference effects from stimuli intervening between repetitions (such effects are beyond the scope of this thesis, though recently explored by Spigler and Stuart, (2017). Since initial trials were twice as frequent as immediate repeats, I randomly dropped half of the estimates for initial trials in each session. The total number of estimates across all runs was 49 for each of the 6 trial-types (F_init, U_init, S_init, F_imm, U_init, S_imm).

Because face trials are twice as frequent as scrambled trials, I matched the number of trials by running two separate analyses, one for famous faces versus scrambled faces and another for unfamiliar faces versus scrambled faces. Since the results of these two analyses were qualitatively similar, I only report the results of unfamiliar faces and scrambled faces in this chapter (interpretation of the results is also slightly easier since subjects have no pre-experimental associations for the unfamiliar faces).

I used the same FFA mask as in the previous chapter, and analysed six repetition-related changes as defined below:

1- Mean Amplitude Modulation (MAM): This is the typical ROI-based measure of RS, i.e, the univariate difference in BOLD response (B_{vspo}) to *p=1* (initial) or *p=2* (repeated) presentations, averaged across the $v=1\ldots$ voxels, $s=1\ldots$ trials (where N_s is the total number of trials per stimulus class) and $o=1\ldots$ stimulus classes:

$$MAM = \overline{\overline{\overline{B_{..1.}}}} - \overline{\overline{\overline{B_{..2.}}}} = \left(\sum_{v=1}^{N_v} \sum_{s=1}^{N_s} \sum_{o=1}^{2} B_{vs1o} - \sum_{v=1}^{N_v} \sum_{s=1}^{N_s} \sum_{o=1}^{2} B_{vs2o} \right) / N_v N_s 2$$

2-Within Class Correlations (WC): This is the mean pairwise correlation of patterns over voxels with the same stimulus class, averaged over the two classes, and then contrasted for initial versus repeated presentations:

$$WC = \left(\sum_{o=1}^{2} \sum_{i=1}^{Ns} \sum_{j>i}^{Ns} cor(B_{.i1o}, B_{.j1o}) - \sum_{o=1}^{2} \sum_{i=1}^{Ns} \sum_{j>i}^{Ns} cor(B_{.i2o}, B_{.j2o}) \right) / N_s(N_s - 1)$$

This captures how repetition makes patterns for the same class more or less similar.

3-Between Class Correlations (BC): This is the mean pairwise correlation of patterns over voxels for different classes, contrasted for initial versus repeated presentations:

$$BC = \left(\sum_{i=1}^{Ns} \sum_{j=1}^{Ns} cor(B_{.i11}, B_{.j12}) - \sum_{i=1}^{Ns} \sum_{j=1}^{Ns} cor(B_{.i21}, B_{.j22}) \right) / N_s N_s$$

This captures how repetition makes patterns of the opposite class more or less similar.

4-Classification Performance (CP): This refers to the ability to classify the two stimulus classes based on their patterns across voxels (MVPA). In its simplest form, it relates to the difference between Within- and Between-Class correlations (as in Carp et al, 2010):

$$CP = WC - BC$$

Note that this measure it is not redundant with the previous two features, since repetition might decrease both WC and BC, but decrease BC more, for example, such that CP increases. To confirm that the same results arise with more sophisticated MVPA methods, I replicated the pattern of significant results below using a SVM and leave-one-run-out cross-validation (similar to the previous chapter).

5- Amplitude Modulation by Selectivity (AMS): This is a further breakdown of the first feature above, where the degree of MAM is related to the degree of "selectivity" of each voxel (as in Kok et al, 2012). Thus voxels are first binned by the absolute t-value of the difference between mean response to each stimuli class (combining both initial and repeated presentations, to avoid regression-to-the-mean), and then the slope estimated of a linear regression of MAM against selectivity across the *b* bins:

$$AMS = slope(MAM_b, bin_b(|ttest(B_{v..1}, B_{v..2})|))$$

where $bin_b(t)$ bins voxels according to ascending values of *t*, and MAM_b is the amplitude after repetition averaged across all voxels in bin *b*. A positive slope indicates that adaptation suppresses the selective voxels more than the non-selective ones. Here I used 6 bins.

6- Amplitude Modulation by Amplitude (AMA): This is similar to feature 5 above, except that voxels were binned by amplitude (averaging across stimulus classes and presentations and runs), rather than by selectivity (as in Wiener et al, 2010):

$$AMA = slope(MAM_b, bin_b(\overline{\overline{\overline{B_{v...}}}}))$$

A positive slope means that repetition suppresses more responsive voxels more.

The values of these six metrics were calculated for each participant and the two-tailed p-value reported for a one-sample T-test across participants versus zero.

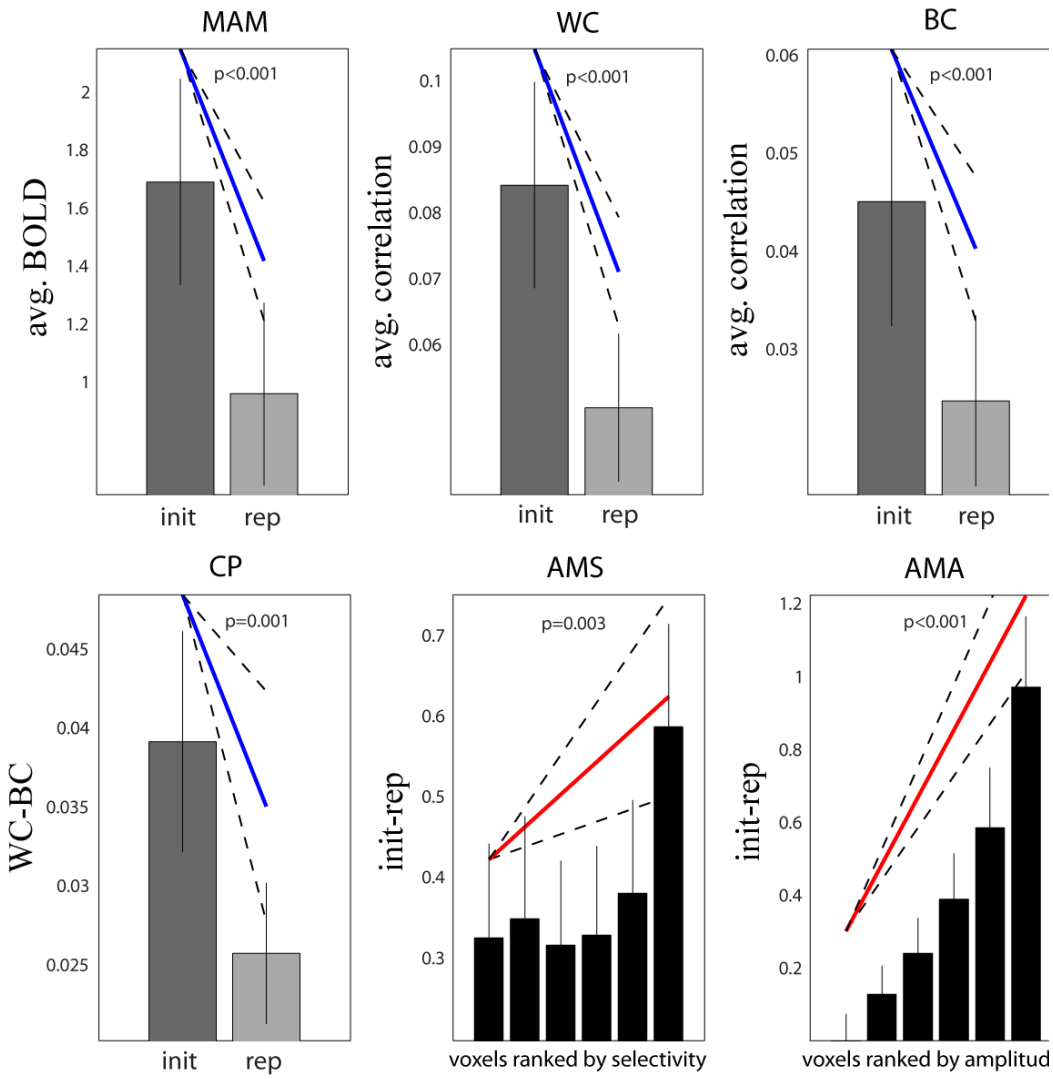


Figure 4.1: Shows the results of the six metrics in FFA. Bars reflect mean across participants for each condition (init = initial presentation; rep = repeated presentation), with error bars reflecting 95% confidence interval versus zero; diagonal line represents slope of linear contrast across conditions (red = positive; blue = negative) with dashed error margins reflecting 95% confidence interval of that slope (equivalent to pairwise difference when only two conditions) and p-value indicated above.

4.3 Results

The six metrics are shown in Figure 4.1. As expected, FFA showed a significant RS in response to repeated faces (**MAM**) (t(17)=-7.53, P<.001). Stimulus repetition also reduced both within-class (**WC**) and between-class (**BC**) correlations between trials

(WC, t(17)=-9.36, P<.001, and BC, t(17)=-6.07, P<.001 (all correlation coefficients were Fisher-z transformed for the t-tests), and the difference between within- and between-class correlations (**CP**) also decreased (t(17)=-3.84, P=0.0012), as confirmed by a SVM (t(17)=-3.06, P=0.007). Furthermore, linear regression showed that RS increased with voxel selectivity (**AMS**) (t(17)=+3.46, p=0.003), and also increased with mean amplitude (**AMA**), (t(17)=+9.26, P<.001.

4.4 Discussion

In this chapter, I introduced 6 metrics that characterise the univariate and multivariate effects of repetition, and how they relate to voxel selectivities and amplitudes. I then applied these metrics to immediate repetition of faces in the FFA ROI in a group of 18 participants. All 6 metrics showed significant effects for unfamiliar and scrambled faces (and also for famous faces, though data not shown). In the next chapter, I attempt to reproduce these 6 data features with a set of neural models.

The significant reduction in MAM was expected because RS to faces in FFA has been reported in many previous studies (though RS was not reported for unfamiliar faces in some studies, this was using delayed repetition, and with immediate repetition, both famous and unfamiliar faces show RS, Henson et al 2004). The reductions in both WC and BC indicate that individual trials become less similar within themselves and with the opposite trial class. One simple explanation for this finding is a decrease in the signal-to-noise ratio, owing to the lower BOLD response for repeated stimuli (and assuming additive noise). This is explored further in the next chapter.

The reduction in CP and AMS for faces in FFA contrasts with the findings of Kok et al., (2012) for visual gratings in V1. This cannot be because Kok et al., (2012) used LSA rather than LSS, because I obtained the same pattern of significant results as above when using LSA (the correlations and the classification performance were quantitatively lower, which most likely owes to less efficient estimates, as explained in Chapter 3). I also repeated the same analyses with spatially smoothed data (Gaussian filter, FWHM=8) to confirm that the pattern of results were not due the differences in the pre-processing steps. Another reason for why some of our data

patterns were different from Kok et al. (2012) could be that expectation-related effects are different from repetition-related effects (Kovacs & Vogels, 2012; Grotheer & Kovacs, 2016; Todorovic & de Lange, 2012). However, because immediate repetition occurred on 50% of trials, expectation effects are likely to be highly correlated with repetition effects in the present paradigm. Therefore, it seems most likely that the different CP and AMS outcomes for our data compared to those of Kok et al. (2012) reflect differences in the stimuli and ROI (i.e, faces and FFA versus gratings and V1). Indeed, in Chapter 6, I will show data from a paradigm closer to Kok et al's, where some of the data features (CP and AMS) show the opposite pattern to the face dataset used here, but in agreement with Kok et al. This would be consistent with claims from Weiner et al. (2010) that different adaptation mechanisms operate in different ROIs. However, as I will show in Chapters 5-6, it is in fact possible to reproduce both sets of results using the same neural mechanism, but with different parameter values (most specifically, the width of neural tuning curves).

Finally, for AMA, I found a positive correlation between the RS and the overall voxel response, similar to Weiner et al. (2010). Superficially, this supports a scaling model; however, the next chapter will demonstrate that other neural mechanisms can produce the same pattern (and indeed, it will be shown that no single data feature is sufficient on its own to identify the underlying neural model; only by considering all six together can useful inferences be drawn).

It is worth noting that the six metrics above are not entirely independent from each other: for instance, CP depends on both WC and BC, though as argued above, it is still theoretically possible to get 2^3=8 different patterns depending on the precise quantitative values of WC and BC. Finally, I do not claim that these are the only six metrics worth measuring - future studies might identify more - but the next Chapter will show that, taken together, they are sufficient to rule out a large number of neural models.

4.5 Chapter Summary

In this chapter, I identified six repetition-related fMRI metrics and reported their values for immediate face repetition in FFA. Many of these metrics were inspired by previous work (Weiner et al., 2010; Kok et al., 2012), but have never been considered all together. The importance of considering all of them will be demonstrated by simulations in the next chapter.

CHAPTER 5: MODELLING FACE REPETITION EFFECTS IN FFA

5.1 Introduction:

In this chapter, I revisit the repetition-related data features identified in the previous chapter and compare several neural adaptation models in their ability to fit all six features. Face stimuli have a lot of features and it is hard to know how they are represented in FFA precisely. For simplicity, I assumed that intact faces and scrambled faces differ along a single dimension (Figure 5.1), though in reality they are likely to differ along multiple dimensions. Assuming that neural populations have a unimodal tuning-curve along a relevant stimulus dimension, at least four basic neural mechanisms of adaptation have been suggested: 1) scaling, where neural populations reduce their firing rate in proportion to their initial firing rate, i.e, their tuning curves are suppressed (Ringach et al, 2002; Swindale et al., 2003; Grill-Spector et al., 2006; Kar and Krekelberg, 2016), 2) sharpening, where the width of neural tuning curves decreases (e.g., Kar and Krekelberg, 2016), 3) repulsive shifting, where the peaks of tuning curves shift away from the adaptor (Dargoi et al., 2000) and 4) attractive shifting, where the peaks shift towards the adaptor (Bachatene et al., 2015). These four mechanisms can be further parametrized according to

whether the "domain" of adaptation is 1) global, affecting all neural populations regardless of their preferred stimulus (as in Weiner et al, 2012), 2) local, where adaptation is greater for neural populations whose preferences are closer to the adaptor, and 3) remote, where adaptation is greater for neural populations whose preferences are further from the adaptor. This results in a space of 12 possible models, as shown in Figure 5.1.

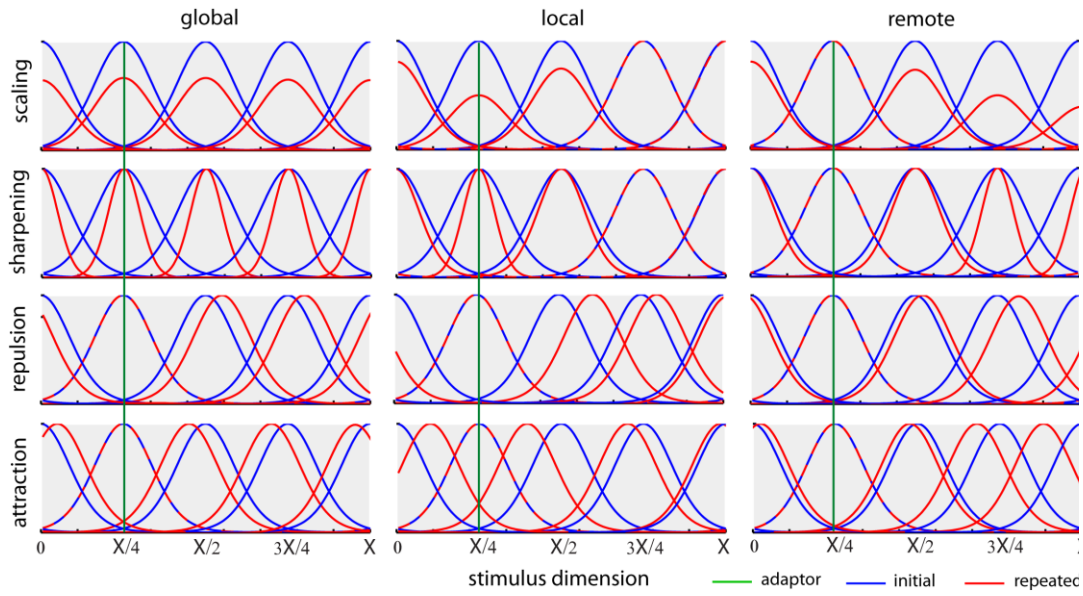


Figure 5.1 – Example tuning curves along a stimulus dimension (ranging from 0 to X), both before (blue) and after (red) repetition of a single stimulus (with value X/4, shown by green line) according to the twelve different neural models of repetition suppression, created by crossing four mechanisms (rows) with three domains (columns). For illustrative purposes, only five neural populations are shown, equally-spaced along the stimulus dimension.

5.1.1 Simulating neural responses

Neural responses were modelled by a Gaussian distribution, which has been shown to provide a good fit to the neural tuning curves (Swindale et al., 1998; Rinach et al., 2002). The firing rate, $f_i(j)$, for the i-th neural population in response to the first presentation of a stimulus, stimulus j, was defined as:

$$f_i(j) = G(x_j, \mu_i, \sigma)$$

where G is the Gaussian distribution, x_j is the value of stimulus j on the stimulus dimension, which ranged from 0… , μ_i is the stimulus preference of the i-th neural population and σ is the dispersion of the tuning curves.[3] Since we do not know the true distance between classes (e.g, faces and scrambled faces) along the stimulus dimension, we arbitrarily set this distance to be $X/2$ by assuming that faces correspond to $x_j = X/4$ and scrambled faces to $x_j = 3X/4$, and varied the neural tuning widths (since increasing the tuning width is equivalent to decreasing the distance between the two classes). The preferred stimulus for each tuning curve (μ_i) was selected randomly from a uniform distribution.

The extent of adaptation was expressed through the variable c, which is a function, h, of the difference between the neural preference and stimulus value, i.e., $c(i, j) = h(\mu_i - x_j)$, which varied between 0 and 1. According to the four basic neural mechanisms of adaptation, the firing rate in response to the second presentation of stimulus j is:

I. Scaling models: $$f_i(j) = G(x_j, \mu_i, \sigma) \times c(i, j)$$

II. Sharpening models: $$f_i(j) = G(x_j, \mu_i, c(i, j) \times \sigma)$$

III. Repulsive Shifting models: $$f_i(j) = G(x_j, \mu_i + c'(i, j) \times \frac{X}{2}, \sigma)$$

IV. Attractive Shifting models: $$f_i(j) = G(x_j, \mu_i - c'(i, j) \times \frac{X}{2}, \sigma)$$

where for shifting models, the adaptation factor additionally depended on 1) the sign of the difference between neural preference and stimulus value:

$$c'(i, j) = \text{sign}(\mu_i - x_j) \times c(i, j)$$

[3] Although "X" can be any value, in this experiment we chose "X" to be equal to "π" to ease comparison with the oriented gratings experiment in the next chapter

and 2) was set to 0 when $i = j$, i.e, $c'(i = j) = 0$. The second property meant that no shift occurred when the stimulus matched the neuron's preference, even for repulsive shifting. The latter is actually found empirically (Dragoi et al., 2000; Dragoi et al., 2001) and corresponds to a quadratic effect of the difference between μ_i and x_j on the size of c'. Rather than parametrize this quadratic function further, by setting $c'(i = j) = 0$, we are effectively limiting its form to the level of discretization of stimulus values ($X/8$ here).

The value of $c(i, j)$, i.e, nature of the function *h* above, determined the range over which neural adaptation applied, which was modelled in three ways, controlled by two free parameters: *a*, controlling the maximal adaptation, and *b*, controlling how rapidly adaptation changes (linearly) with distance between neural preference and stimulus orientation:

A. Global adaptation: $c_G(i, j) = a$

$0 < a < 1 = \text{constant}$

B. Local adaptation: $c_L(i, j) = \min\left[1, a + \left|\frac{(x_i - x_j)}{b}\right|(1 - a)\right]$

$0 < b < \frac{X}{2} = \text{constant}$

C. Remote adaptation: $c_R(i, j) = \max\left[a, 1 - \left|\frac{(x_i - x_j)}{b}\right|(1 - a)\right]$

$0 < b < \frac{X}{2} = \text{constant}$

The parameter *b* represents a linear slope (Figure 5.2). The min/max operation will ensure that adaptation values are constrained between *a* and 1. This nonlinearity (modelled as piecewise linear) is important for the visual grating data presented in Chapter 6, to break the symmetry of results after adapting to two orthogonal classes (see Chapter 6 for details).

Global adaptation is a special case of Local adaptation when $\lim\{b->\infty\}$,

$c = \min(1 , a + 0)$ → $c = a$.

Global adaptation is a special case of Remote adaptation when $\lim\{b->0\}$

$c = \max(a , 1 - \infty)$ → $c = a$,

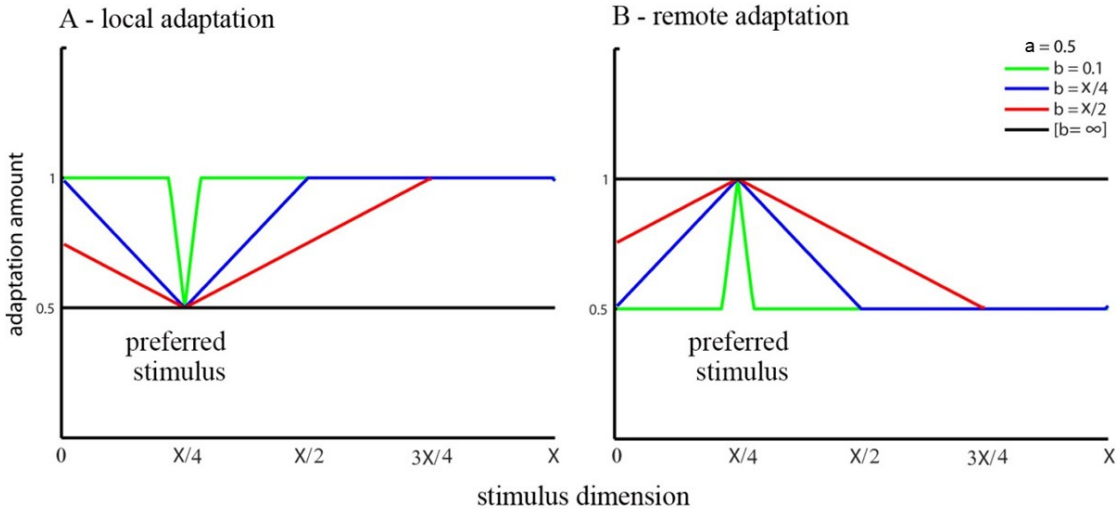


Figure 5.2: Distance functions, showing how amount of adaptation depends on distance between stimulus class (x-axis) and neural preference (here X/4). The *a* parameter is fixed to 0.5, while the *b* parameter is shown from 0.1 to ∞, though note that in our simulations, *b* only ranged from 0.1 to X/2.

Note that the peak firing rate could never increase after adaptation (i.e, the tuning curves do not represent probability distributions, where sharpening for example would also produce an increase in firing of the most selective neuronal populations). To be clear, the width of the tuning-curves was kept fixed for the scaling and shifting models, and the height of the tuning-curves was kept fixed for the sharpening models and shifting models.

In the present simulation, there were 9 runs, each containing 2 alternating trial types, X/4 and 3X/4, reflecting faces and scrambled faces, each repeated 6 times per

session[4]. Given the fact that each face and scrambled face stimuli was unique in the fMRI experiment, I assumed adaptation wore-off after each repetition by resetting the adaptation factor, *c*, to 1.

5.1.2 Simulating voxel responses

Although a single voxel contains a large number of neurons, that does not mean all of them are functionally distinct or non-overlapping. In fact, many neurons are clustered into functionally similar groups or columns (Blasdel, 1992). In this study, each voxel was assumed to contain *N* functionally-distinct neural populations, whose preferences were randomly selected from a uniform distribution (see below). Assuming that the BOLD response is proportional to the neural firing rate (Rees et al, 2000; Heeger et al., 200), the voxel response was simply the average firing rate of each population within that voxel. The number of neurons per voxel, *N*, does not have a qualitative effect on the simulation results. However, it does have a quantitative effect: When *N* is large; the majority of the voxels would be similar to each other in their response, with very weak overall voxel biases towards particular stimuli class. If *N* is small however, the voxels have stronger biases, and the quantitative differences among the models become more evident. Here I used *N=8.*

I then simulated $N_v = 200$ voxels, and added a small amount of independent noise to each voxel, drawn from a Gaussian distribution with standard deviation of 0.1 (which represented an SNR of 10, since the peak value of $f_i(j)$ for initial presentations was scaled to 1).

To generate voxels that vary in their selectivity and activity, the value of μ_i was sampled randomly from 9 possible values in steps of $X/8$ from $0...X$. These values therefore included two neural populations that responded optimally to one of the stimuli (Faces and Scrambled Faces at $x_j = X/4$ and $x_j = 3X/4$ respectively), three non-selective neurons (at $x_j = 0, X/2, X$), and four partially-selective neurons in between. However, to explain why FFA responds more overall to faces than

[4] The real fMRI data contained more trials per session (see chapter 4), however, for computational reasons I simulated a smaller number of trials. Since noise levels can be varied in the simulations, excess trials were not needed (as confirmed by sanity checks).

scrambled faces, I sampled twice more often from X/4 than any other value (which represents neurons selective to faces[5]). Since only a small number of populations were randomly at each voxel, this in turn generated a variety of voxel selectivities. Note that it is important to distinguish "sharpening" at the level of the neuronal population (i.e, individual tuning curves) from "sharpening" at the level of the voxel (i.e, mean response over all neuronal populations within that voxel) as mentioned in Chapter 1. Local sharpening of neural populations does not necessarily cause much "sharpening" of the stimulus representation at the level of voxels. Rather, it is the neuronal mechanism of remote *scaling* that causes the most marked sharpening of the representation within a voxel – i.e, the "drop out" of non-selective neurons from the initial presentation, in the sense proposed by Wiggs & Martin (1998).

5.2 Simulation results

Each model had 3 free parameters: a, b and σ (except for the Global models where there was no b parameter). I explored the predictions of each of the 12 models for each of the 6 data features in a grid search covering a wide range of values for the three parameters. The a values ranged from 0.1 to 0.9 in steps of 0.1 to cover a wide range of maximal adaptation, while b values ranged from 0.1 to X/2, in steps of 0.2 (where $X = \pi$). For σ, values ranged from 0.1 to 1 in steps of 0.2, and then from 2 to 12 in steps of 3 to cover a wide range of tuning widths. For each model, I ran 50 simulations for each of the 8000 unique combinations of these three parameters (or 800 for Global models with just two parameters). For each parameter combination, I calculated the 99% confidence interval across the 50 simulations for the value associated with each of the 6 data features, and tested whether this was above, below, or overlapped zero. Figure 5.3 summarises the results of the grid analyses. The colours in each circle summarise the possible trends after repetition, i.e, whether the models could explain an increase, decrease or no effect, or some combination of these (with different parameter combinations).

The first thing to note from Figure 5.3 is that no single data feature was sufficient to identify the underlying neural model, illustrating the difficulty of inferring from fMRI data at the level of voxels to mechanisms at the level of neurons (i.e., no value

[5] For interest, I also simulated equal sampling from all tuning-curves (as done for the visual gratings in the next chapter), and this did not affect the qualitative results.

in any row in Figure 5.3 is unique to one of the twelve models). Note that this conclusion holds regardless of the empirical value of the data features observed in the present experiment (leftmost column). This conclusion is important because some of these features, such as the increase in CP after repetition, have been assumed to support sharpening models (Kok et al., 2012), yet Figure 5.3 shows that several other non-sharpening models can produce an increase in CP. The same goes with the negative slope in AMS, which was also thought to support sharpening models (Kok et al., 2012).

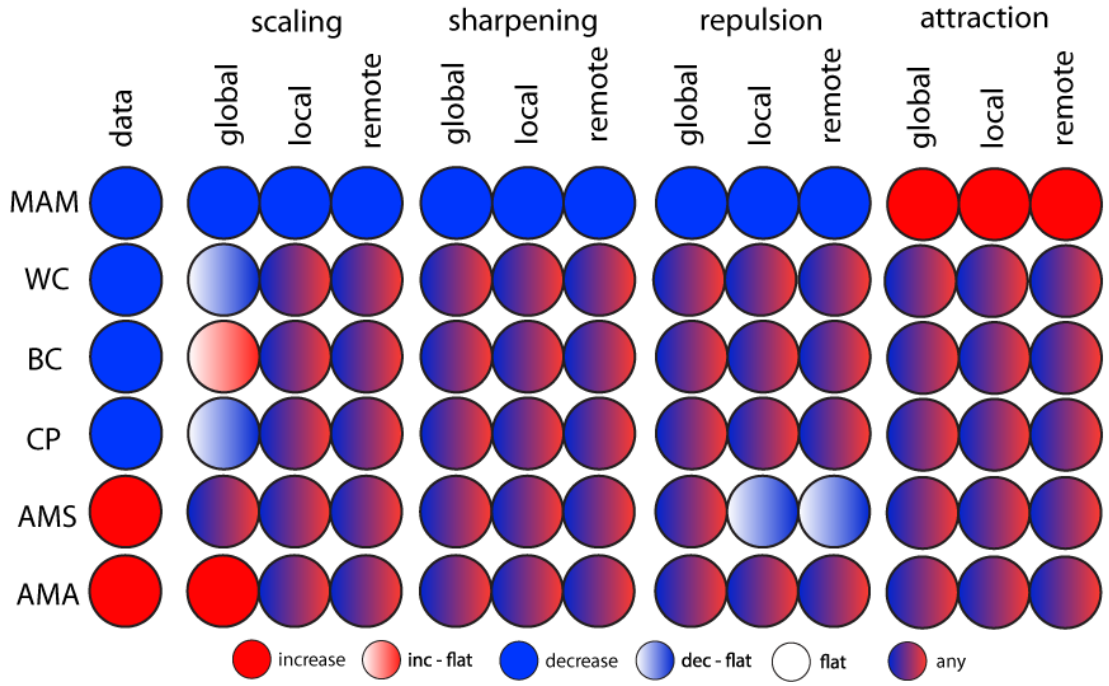


Figure 5.3: Simulation results for each of the 12 models (columns) for each of the 6 data features (rows). Each coloured circle represents either no effect (flat; white), a decrease (blue), an increase (red), or some combination of these, when the model parameters a, b, and σ could take any value (within the grid search) for any data features (i.e parameters were not constrained to be same across data patterns; cf Figure 5.4).

The second thing to note is that some of the neural models cannot produce at least one of the data features observed in the present experiments (whatever their parameter settings within the large range explored here). This can be seen by comparing the leftmost column with the remaining twelve columns. This means that,

by considering a range of consequences of repetition (both univariate and multivariate), one can at least rule out some neural models. Nonetheless, with unconstrained parameters, there were still six models that could fit the data (local and remote scaling, all three sharpening models and global repulsion). Interestingly however, when simulations were constrained to have the same parameter values for all 6 data features, only one model – local scaling – survived. This can be seen from Figure 5.4, in which green and red colors now show whether a model could fit the data feature observed empirically (when parameter values were selected that explained the maximum number of features): only the column for the local scaling model has green colors for all data features.

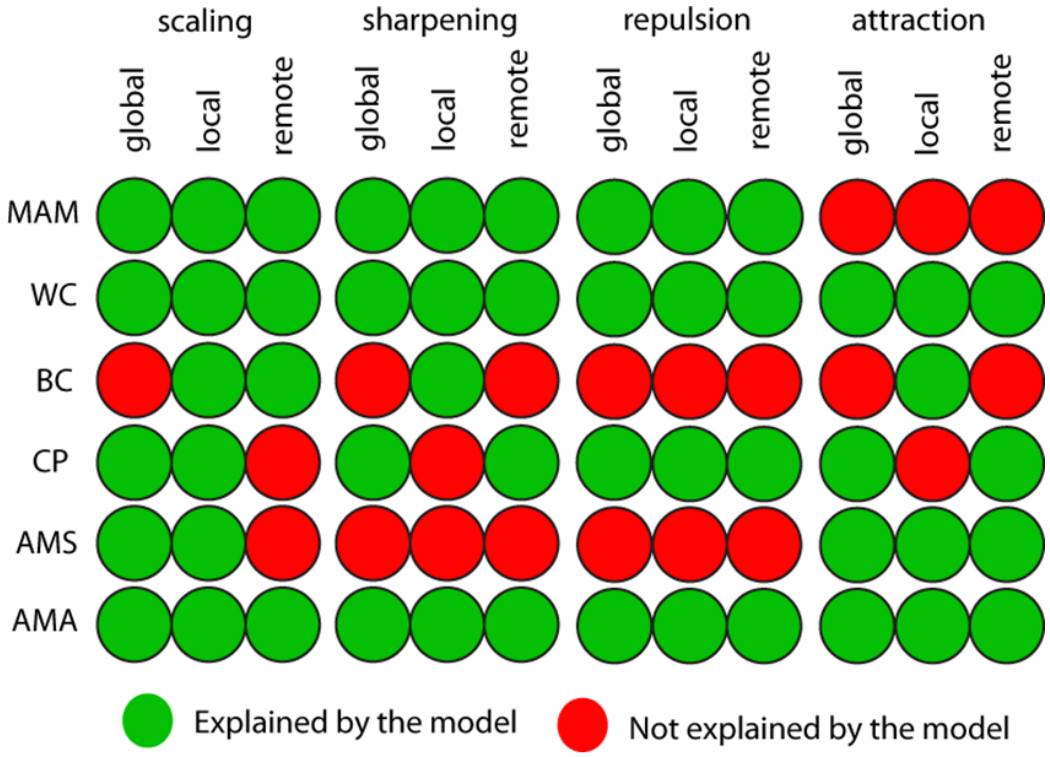


Figure 5.4: The maximum number of data properties explained by each model when parameters are constrained to be equal across all data properties. Note that, for some models that can explain only 4 or 5 data properties with the same parameter values, there may be different subsets of the same number of data properties that can be explained (i.e., this figure only shows one such subset).

The winning parameters for the local scaling model were in the following ranges: a (0.6-0.8), b (0.1-1.0) and σ (0.1-0.4) (see Appendix 6 for the complete list of the

winning parameters). Figure 5.5 shows the simulation results for one of the wining parameter combinations for local scaling. The qualitative results are similar to those in real fMRI data in Figure 4.1. Note that we are only modelling the repetition effects, i.e the difference between initial and repeated conditions. The BC values for each condition alone are negative, unlike the positive values in the data. This positive offset in correlations could have many causes, one of which is modelled in Appendix 5.

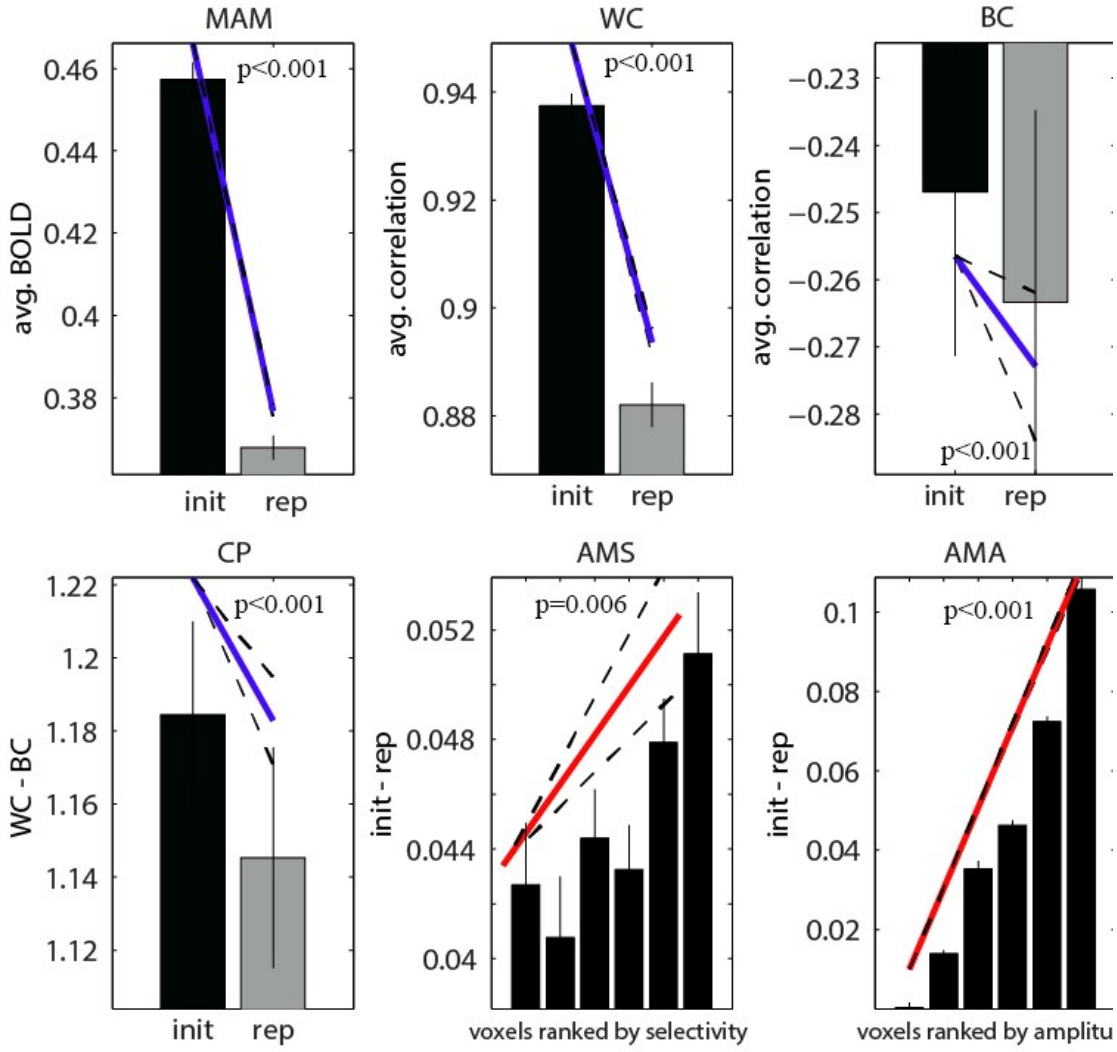


Figure 5.5: Predicted data features from averaging 18 simulation runs using the winning model (local scaling), with parameters: *a=0.7, b=0.3, σ=0.2* (cf. data in Figure 4.1) to explain the behaviour of the average participant (fixed effect). Apart from the correlations, the Y-axes have arbitrary units. All trends were significant p<0.001; error bars are CI 95%.

5.3 Discussion

Many previous studies (Grill-Spector et al., 2006; Wiggs & Martin, 1998; Henson & Rugg 2003) proposed sharpening or scaling models to explain the reduced BOLD response for repeated stimuli. However, the few simulation studies that have attempted to model repetition-related effects in fMRI did not model the effects of the distance between the adapter and the neural preference, which is clearly an important factor in single-cell studies (Kar and Krekelberg, 2016; Dragoi et al, 2000; Bachatene et al., 2015). Here, I added this additional assumption, and expanded the model space to 12 different models, in order to encompass shifting models as well. By searching over the parameter space of each model, the simulation results show that no single data feature uniquely identifies a model, and in principle, any of these mechanisms (except attraction models) can explain the basic effect of repetition on the mean univariate fMRI response across voxels. However, by considering a range of features of fMRI repetition effects (univariate and multivariate), and formally modelling a range of potential neural mechanisms, various hypothetical neural mechanisms can be distinguished. Indeed, the results show that local scaling of neuronal firing is the only model, of the twelve considered here, that can simultaneously explain six fMRI features of face repetition in FFA in the present experiment.

Scaling models assume a reduction in the firing rate. Global scaling refers to a uniform reduction in the firing rate for all neurons, regardless of the distance of the stimulus from their tuning curves, while local scaling refers to a greater decrease in the firing rate for neurons whose preference is closer to the stimulus (Kar & Krekelberg, 2016). Remote scaling proposes the opposite, ie. a greater decrease for non-optimal neurons (Ringach et al., 2002). Importantly, both local and remote scaling at the neural level can improve CP after repetition, by increasing the voxel-to-voxel variance (Davis et al. 2014). This is shown in Figure 5.3, where local and remote scaling can increase CP (or decrease CP, depending on parameter values). Indeed, it is somewhat counter-intuitive that both can increase CP, since local scaling is the "opposite" of remote scaling. In fact, our grid search showed that all models except global scaling can potentially explain the increase in CP reported by Kok et al, and therefore this finding of sharpening at the voxel level does not imply

sharpening at the neural level.

Occam's razor principle dictates that the definition of a best fitting model should also account for its complexity. In that case, the global scaling model is simpler than the local scaling model because it is controlled by one rather than three free parameters. However, our grid search covered a wide range of parameter values (Figure 5.3) and showed that global scaling can never produce a decreasing trend for BC, suggesting that the additional model complexity of local models was necessary to achieve the qualitative fitting for this criterion. It is possible that the various local models also had different model complexities (even though they had the same number of free parameters), when model complexity takes into account correlations between the effects of the parameters and/or the functional form of the equations (Myung et al., 2009). However, the same argument applies when comparing local scaling with the other local and remote models, because Figure 5.4 showed that only local scaling was able to qualitatively fit all six data features, whereas the other local models were at best able to qualitatively fit only 4 data features. Therefore, even if the local scaling model were inherently the most complex model, the approach in this thesis to exclude models based on their ability to achieve a qualitative fit to data features necessitate this additional model complexity.

Although the local scaling model was the only model capable of simultaneously fitting all six data features with the same set of parameter values, it is possible that these findings could be explained by combinations of mechanisms (e.g. global scaling and local sharpening), or by neuronal mechanisms beyond the twelve considered here. Nonetheless, local scaling remains the most likely current explanation, in terms of parsimony. Our finding that local scaling best explains fMRI repetition suppression does not question previous findings of stimulus repetition effects on single-cell recordings (Ringach et al, 2002; Swindale et al., 2003; Kar and Krekelberg, 2016; Dargoi et al., 2000; Bachatene et al., 2015). As alluded to above, it is possible that multiple mechanisms operate in parallel, but in different neural populations or cortical layers, and that the dominance of the local scaling model is simply due the greatest proportion of neurons exhibiting local scaling. However, I did not consider combinations of models further because an important aim was to understand the capabilities of each of the 12 adaptation models on their own.

Note that local scaling of neuronal tuning curves could itself arise from multiple potential mechanisms within the context of a neuronal circuit, such as synaptic depression of bottom-up inputs, or recurrent inhibition by top-down inputs. For example, the hypothesis of predictive coding (see Chapter 1), which has been used to explain other aspects of repetition suppression (Garrido et al., 2009; Henson, 2003), would also result in maximal suppression of neurons that are most selective for the (repeated) stimulus – i.e., local scaling.

5.4 Chapter Summary

In this chapter, I tested the predictions of 12 neural adaptation models to simulate the six repetition-related fMRI features for faces and scrambled faces in FFA described in the previous chapter, and only the local scaling model was able to fit all six with the same parameter values. Thus a reduction in the firing rate of neurons that is proportional to how closely an adapting stimulus matches their preferred stimulus appears to be the most parsimonious explanation of the FFA response to repeated faces. However, an obvious question is whether this is true for other stimulus types and other paradigms. In the next chapter, I will test the same models on another dataset that differs in stimuli, paradigm and ROI.

CHAPTER 6: MODEL VALIDATION WITH A DIFFERENT DATASET / EXPERIMENT 2

6.1 Introduction

In the previous two chapters, I identified six effects of repeating faces and scrambled faces within the FFA, and showed how only one neural model – local scaling – could simultaneously fit all six. However, these findings may not generalize to other types of stimulus, paradigm, or brain region (ROI). Indeed, Wiener et al. (2010) found contrasting results between medial and lateral visual ROIs for one of their repetition effects. Moreover,Kok et al. (2012) found that perceptual expectation of visual gratings (on the basis of an auditory cue) reduces BOLD response in V1 (a phenomenon known as "expectation suppression", ES) while at the same time increasing the classification performance (CP). They also found that the amount of BOLD suppression was inversely related to voxel selectivity (AMS). These results are opposite to our findings in Exp.1. Although Kok et al. (2012) focused on ES, our proposed twelve adaptation models seem applicable to ES as well, especially given that "scaling" versus "sharpening" is an ongoing debate in ES studies too. For

example, Blank & Davis (2016) found the best support for their data results from a scaling model, while Kok et al. (2012) attributed their findings above to neural sharpening (though without performing simulations to validate their claim). Furthermore, our grid search results in Figure 5.3 show that a local scaling model also has the potential to explain Kok et al's findings in V1 without resorting to neural sharpening.

It is worth noting that ES and RS might be related and share the same underlying cause. For example, Summerfield et al. (2008) found that unexpected repetitions have lower RS than expected repetitions; hence, RS could reflect an increased expectation to a repeated stimulus due to its recent presentation. However, there are also claims that BOLD suppression due to expectation is mechanistically different from BOLD suppression due to repetition (Todorovic & de Lange 2012; Kovacs and Vogels, 2014; Grotheer and Kovacs, 2015). Thus, I wanted to see if local scaling could explain all six features of repetition using the same stimuli (orthogonally-oriented gratings) and ROI (V1) as Kok et al. (2012), but used data from a more conventional grating-adaptation paradigm, provided by Dr Arjen Alink in an experiment previously reported in Alink et al. (2013, 2015). Not only did this paradigm differ in stimuli and ROI from Chapter 4, but it also used more sustained epochs of each stimulus class, rather than brief events, and a paradigm in which the stimulus classes alternated, so repetition of one class was separated by an epoch of the other class (e.g. S1-S2-S1), rather than being repeated immediately as in Chapter 4 (e.g, S1-S1…S2-S2).

6.2 Design, fMRI acquisition, and Preporcessing

Eighteen healthy volunteers (13 female, age range 20–39) with normal or corrected-to-normal vision took part in the experiment. The gratings were oriented 45° ($\pi/4$ radians) or 135° ($3\pi/4$ radians) from the vertical, with a spatial frequency of 1.25 cycles per visual degree (a frequency that strongly drives V1, Henriksson et al., 2008). These stimuli were presented during 2 runs of 8 minutes, with each run divided into 4 subruns, and each subrun containing 6 blocks, with the orientation presented in each block alternating (Figure 6.1). Each block lasted 14s and contained 28 phase-randomized gratings of one orientation, presented at a frequency of 2 Hz. The stimulus duration was 250 ms, followed by an interstimulus interval (ISI) of 250

ms, during which a central dot was present, surrounded by a ring that determined the task (see below). The spatial phase was drawn randomly from a uniform distribution between 0 and 2π. Stimulus blocks were separated by 2s fixation periods and subruns by 24s fixation periods. In addition, each participant participated in a 12-minute run for retinotopic mapping. A description of the stimuli employed and the procedure used to define individual regions of interest (ROIs) for the primary visual cortex can be found in Alink et al. (2013).

Participants were instructed to continuously fixate on a central dot (diameter: 0.06◦ visual angle). The dot was surrounded by a black ring (diameter: 0.20◦, line width: 0.03◦), which had a tiny gap (0.03◦) either on the left or right side. The gap switched sides at random at an average rate of once per 3s (with a minimum inter-switch time of 1s). The participant's task was to continuously report the side of the gap by keeping the left button pressed with the right index finger whenever the gap was on the left side, and keeping the right button pressed with the right middle finger whenever the gap was on the right side. The task served to enforce fixation and to draw attention away from the stimuli.

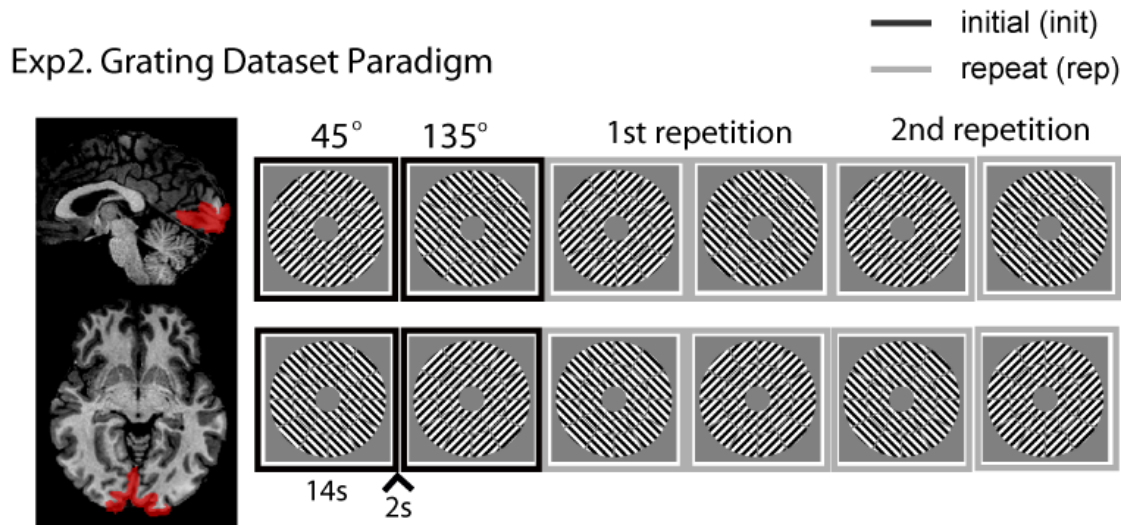


Figure 6.1, showing the experimental design, with the V1 mask from Alink et al (2013).

Functional and anatomical MRI data were acquired on the same scanner as Experiment 1 (see above). During each stimulus run, 252 volumes were acquired containing 31 transverse slices covering the occipital lobe as well as inferior parietal,

inferior frontal, and superior temporal regions for each subject using an EPI sequence (TR=2000ms, TE=30ms, flip angle=77◦, voxel size: 2.0mm isotropic, field of view: 205mm; interleaved acquisition, GRAPPA acceleration factor: 2). The same EPI sequence was employed for a retinotopic mapping run, during which we acquired 360 volumes. We also obtained a high-resolution (1mm isotropic) T1-weighted anatomical image using a Siemens MPRAGE sequence.

Functional and anatomical MRI data were pre-processed using the Brainvoyager QX software package (Brain Innovation, v2.4). After discarding the first two EPI images for each run to allow for T1 saturation effects, the functional data were corrected for the different slice times and for head motion, detrended for linear drift, and temporally high-pass filtered to 2 cycles per run. The data were aligned with the anatomical image and transformed into Talairach space. After automatic correction for spatial inhomogeneities of the anatomical image, an inflated cortex was reconstructed for each subject. All ROIs were defined in each individual subject's cortex reconstruction and projected back into voxel space.

For simplicity, I contrasted first versus third presentations of each orientation within a subrun, to maximize repetition effects. There are likely to be effects induced by intermediate presentations of the other orientation, which are modelled fully in the models described below. Any effects of the order of specific orientations were controlled by counterbalancing (i.e., averaging over sub-runs that started with 45° and sub-runs that started with 135°). The number of voxels in the V1 ROIs varied from 775-1598 across participants (M=1100, SD=220). The number of trials (replications) for each stimulus class and presentation (N_s) corresponded to 32 (4 in each of 8 sub-runs).

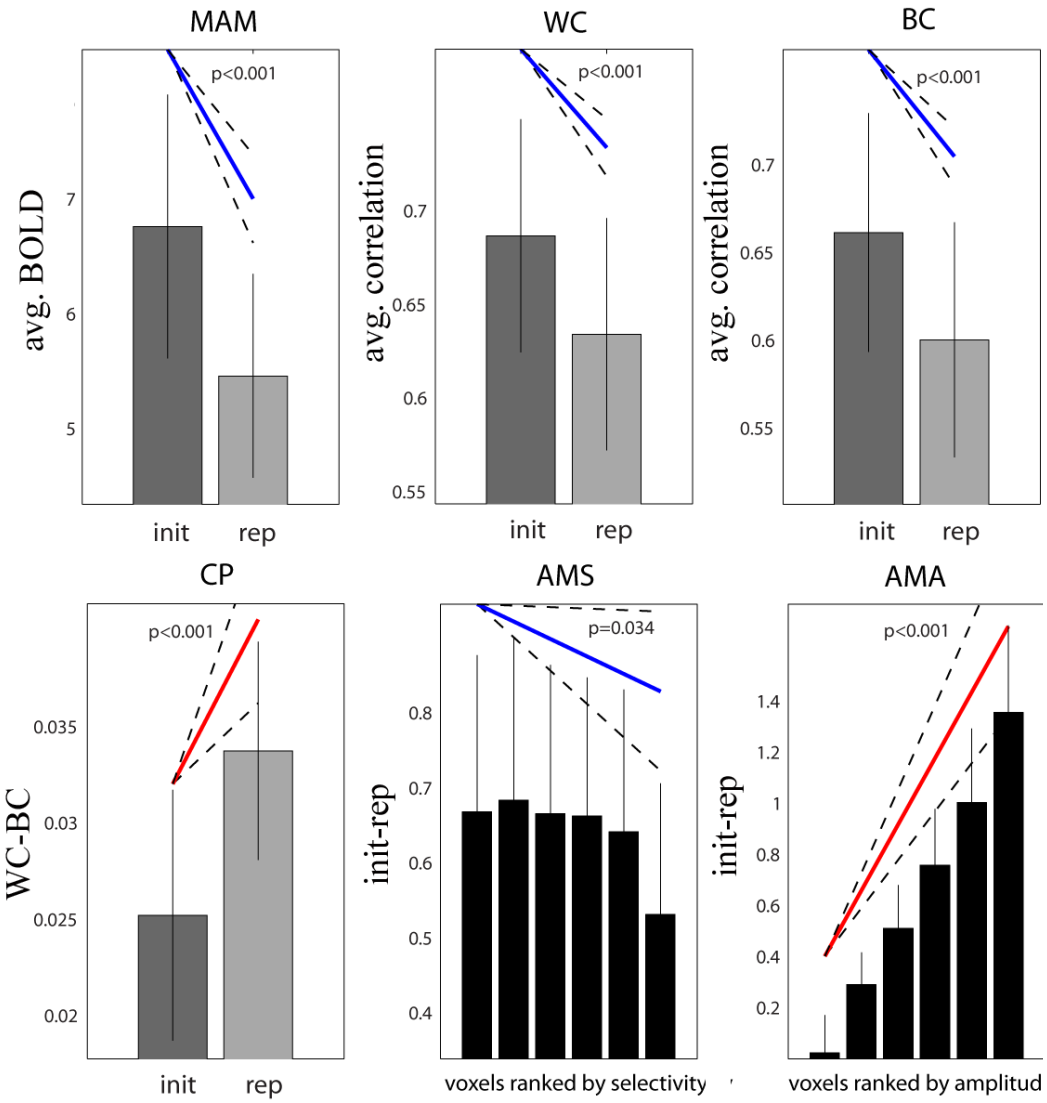


Figure 6.2: Results of the six data features in V1. Bars reflect mean across participants of each condition (init = initial presentation; rep = repeated presentation), with error bars reflecting 95% confidence interval versus zero; diagonal line represents slope of linear contrast across conditions (red = positive; blue = negative) with dashed error margins reflecting 95% confidence interval of that slope (equivalent to pairwise difference when only two conditions) and p-value indicated above.

6.3 Data Results

The six data features are shown in Figure 6.2. As expected, there was significant RS (**MAM**) (t(17)=-7.13, P<.001). Stimulus repetition also reduced both within-class

(**WC**) and between-class (**BC**) correlations between trials (WC, t(17)=-6.62, P<.001, and BC, t(17)=-7.02, P<.001; after Fisher-transforming the correlation coefficients), similar to FFA in Chapter 4. However, the difference between within versus between class correlations (**CP**) increased with repetition (t(17)=+4.15, P<.001 (as confirmed by support-vector machines t(17)=+3.09, P=.006). In other words, repetition improved the ability to classify stimuli according to their two classes, contrary to the FFA results in Chapter 4, but consistent with the previous V1 results by Kok et al. (2012). Furthermore, while linear regression showed that RS decreased with mean amplitude (**AMA**) as for FFA (t(17)=+7.83, P<.001), its dependence on voxel selectivity (**AMS**) showed the opposite pattern of decreasing with selectivity (t(17)=-2.31 p=0.034). Thus, there are both commonalities and differences in the effects of repetition across this grating experiment and the previous face experiment.

6.4 Simulations

The modelling needed to address the main differences between the two datasets in ROI, stimulus type, and experimental paradigm. One difference in the ROIs is that we know that FFA responds more to faces than scrambled faces, which I accounted for by biased sampling towards faces in the previous chapter, whereas V1 typically responds equally (when averaging over all voxels in the ROI) for 45◦ and 135◦ gratings, and therefore I sampled V1 neurons uniformly across the stimulus dimension.

For the stimulus type, I circularised the stimulus dimension between 0 and $X = \pi$, given the symmetry of orientations across the vertical, i.e, that an orientation of 0 is equivalent to an orientation of π (180◦), since gratings are not directional. The tuning functions were therefore represented by von Mises rather than Gaussian distributions; distributions that have been shown to approximate real neural tuning curves in V1 to a reasonable extent (Swindale et al., 1998):

$$f_i(j) = VM(2x_j; 2\mu_i, 1/\sigma)$$

where VM is the classic von Mises distribution. I also needed to circularize the distance function between neural preference and stimulus value, according to:

$$d(i,j) = \min\left(\left|c(i,j)\right|, X - \left|c(i,j)\right|\right) \qquad c(i,j) = \mu_i - x_j$$

where *i, j* denote the neural preference and stimulus value respectively, as in Chapter 5. This ensures that a difference of 180◦ between preference and stimulus corresponds to $d(0,\pi) = 0$. Figures 6.3 and 6.4 show the effects of this circularisation (cf. Figures 5.1 and 5.2).

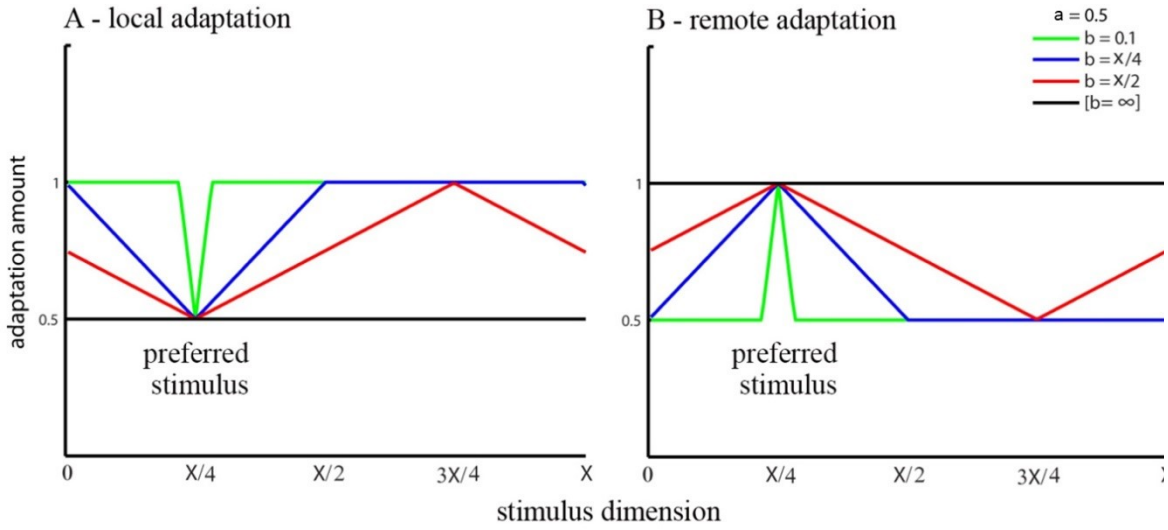


Figure 6.3: Distance functions on circular stimulus dimension, showing how amount of adaptation depends on distance between stimulus class (x-axis) and neural preference (here X/4). The *a* parameter is fixed to 0.5, while the *b* parameter is shown from 0.1 to ∞, though note that in our simulations, *b* only ranged from 0.1 to X/2.

In terms of the stimulation protocol, unlike the face paradigm, the visual gratings were not unique and each presentation (blocks of a given orientation) is a repetition. The blocks within each sub-run were ordered either:

$\left[\pi/4, 3\pi/4, \pi/4, 3\pi/4, \pi/4, 3\pi/4\right]$ or $\left[3\pi/4, \pi/4, 3\pi/4, \pi/4, 3\pi/4, \pi/4\right]$. I therefore applied the adaptation factor in a multiplicative fashion, i.e, for the *p*-th presentation of a stimulus (*p>1*):

$$c(i,j,p) = \prod_{q=2}^{q=p} c(i,j)$$

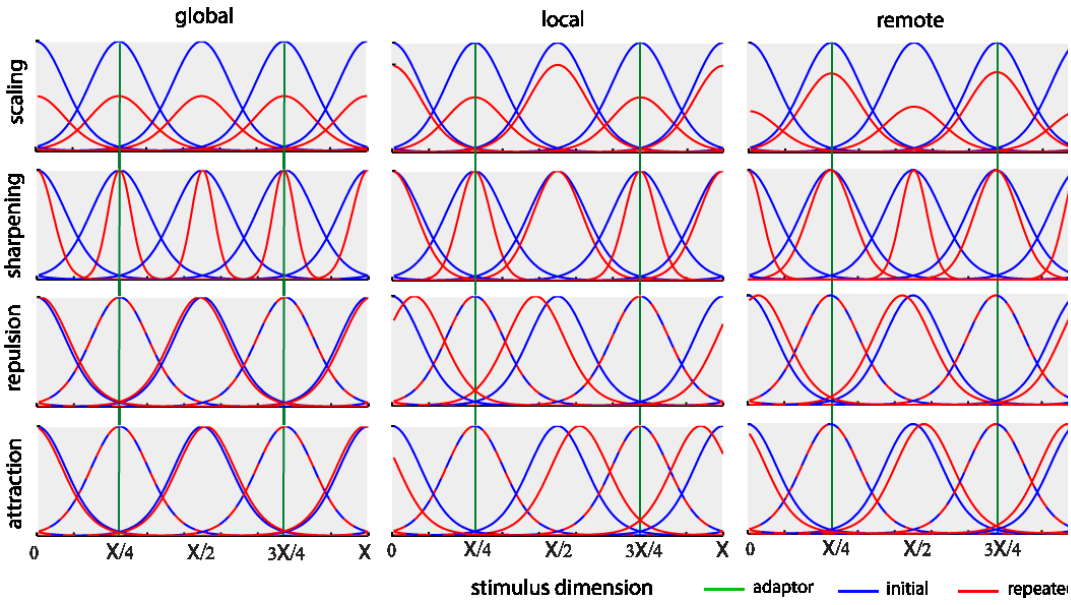


Figure 6.4: Example tuning curves before (blue) and after (red) after adaptation to both orientations (X/4 and 3X/4), according to the twelve different neural models of adaptation. For illustrative purposes, we only show 4 neural populations equally-spaced along the stimulus dimension; in the simulations below, we sampled the population preferences randomly from a uniform distribution across 8 possible equally-spaced orientations (see Chapter 5 for details).

Given the gap between sub-runs, I assumed that adaptation wore-off between runs, by resetting c to 1 at the start of each sub-run. This assumption is supported by the activity profile across the whole experiment shown in Appendix 4. Note that the pattern of simulation results below remained unchanged if we additionally simulated adaptation effects for every trial within each block (thereby affecting mean response to the first block too). Note also that in the grating protocol, unlike for the S1-S1 face protocol, adaptation to the stimulus class (S1) in one block affects the response to the other stimulus type (S2) in subsequent blocks. This is shown in Figure 6.4, in which the tuning curves for the "repeated" condition are after one presentation of both S1 and S2 blocks. Because S1 and S2 are orthogonal, the adaptation effects are symmetrical (around $\pi/2$). Indeed, this is the reason for introducing the nonlinear (piecewise linear) distance function in Chapter 5: if the distance function were purely linear, then the response of neurons whose preference is half-way between the S1 and S2 could never exceed that of neurons whose preference matches either one. Without this nonlinearity, none of the models were able reproduce the results in this

paradigm. Apart from the above changes, the same models and grid-searches were run as in Chapter 5.

6.5 Simulation Results

The grid search results are summarised in Figures 6.5 and 6.6. The first thing to notice is that results in which parameter values were unconstrained look slightly different from those in Figure 5.3, which must owe to the differences in paradigm. For example, attraction models can now produce RS, unlike in the previous paradigm where they could only produce an increase in BOLD response after repetition. This is because attraction to one stimulus (S1) results in repulsion for the other, subsequent stimulus (S2), hence potentially causing RS for the latter. Nonetheless, the same more general point arises as in Chapter 5, i.e, no single data feature alone is sufficient to identify a single model.

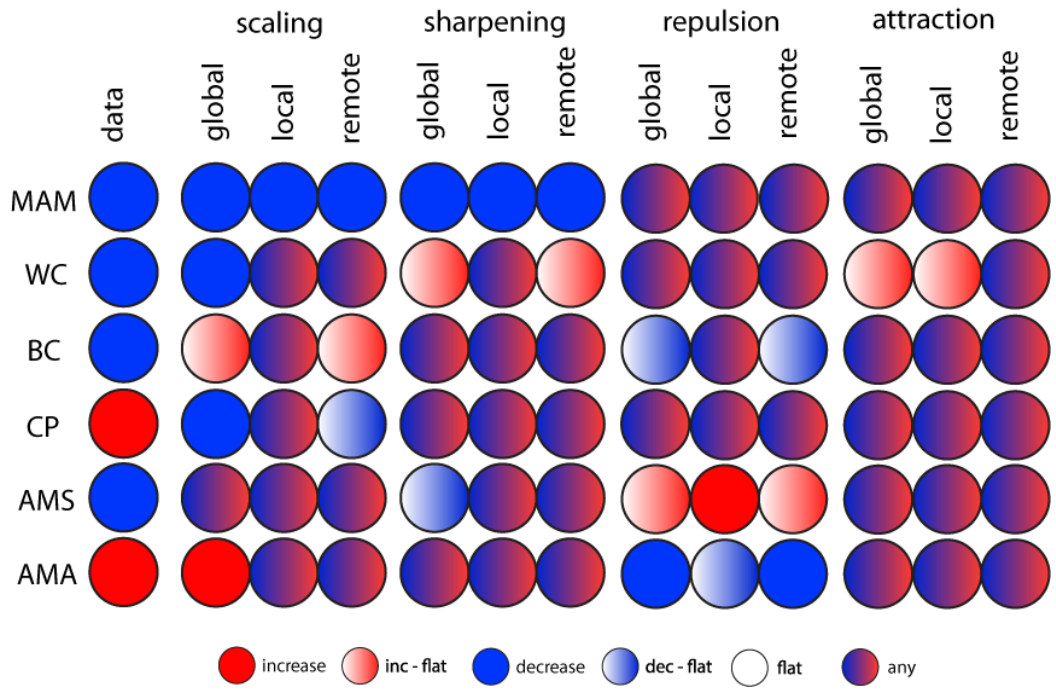


Figure 6.5: Simulation results for each of the 12 models (columns) for each of the 6 data features (rows). Each coloured circle represents either no effect (flat; white), a decrease (blue), an increase (red), or some combination of these, when the model parameters *a*, *b*, and σ could take any value (within the grid search) for any data features (i.e parameters were not constrained to be same across data patterns; cf below).

The most important result is apparent in Figure 6.6, which shows that, when the parameter values are constrained across all six data features, it is again only local scaling that can reproduce all those features. The winning parameters for the local scaling model were in the following ranges: a (0.7-0.9), b (0.1-0.8) and σ (0.4-1.0).

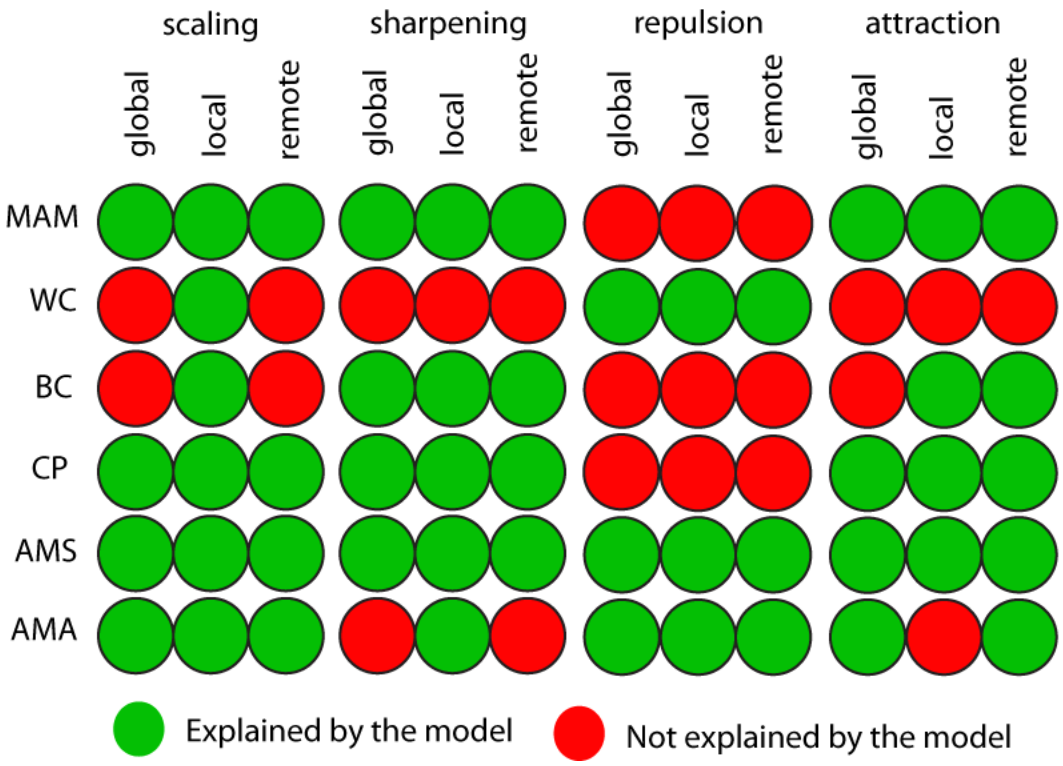


Figure 6.6: The maximum number of data properties explained by each model when parameters are constrained to be equal across all data properties. Note that, for some models that can explain only 4 or 5 data properties with the same parameter values, there may be different subsets of the same number of data properties that can be explained (i.e., this figure only shows one such subset).

For confirmation, Figure 6.7 shows one set of parameter values for the local scaling model, which reproduce the same qualitative patterns as the data in Figure 6.2 (for the complete list of possible parameter values, see Appendix 6). Again, the lack of quantitative fit for the BC correlations (being negative rather than positive) can be explained by some simple additions to the model, as shown in Appendix 5; the

important aspects are the direction of repetition effects.

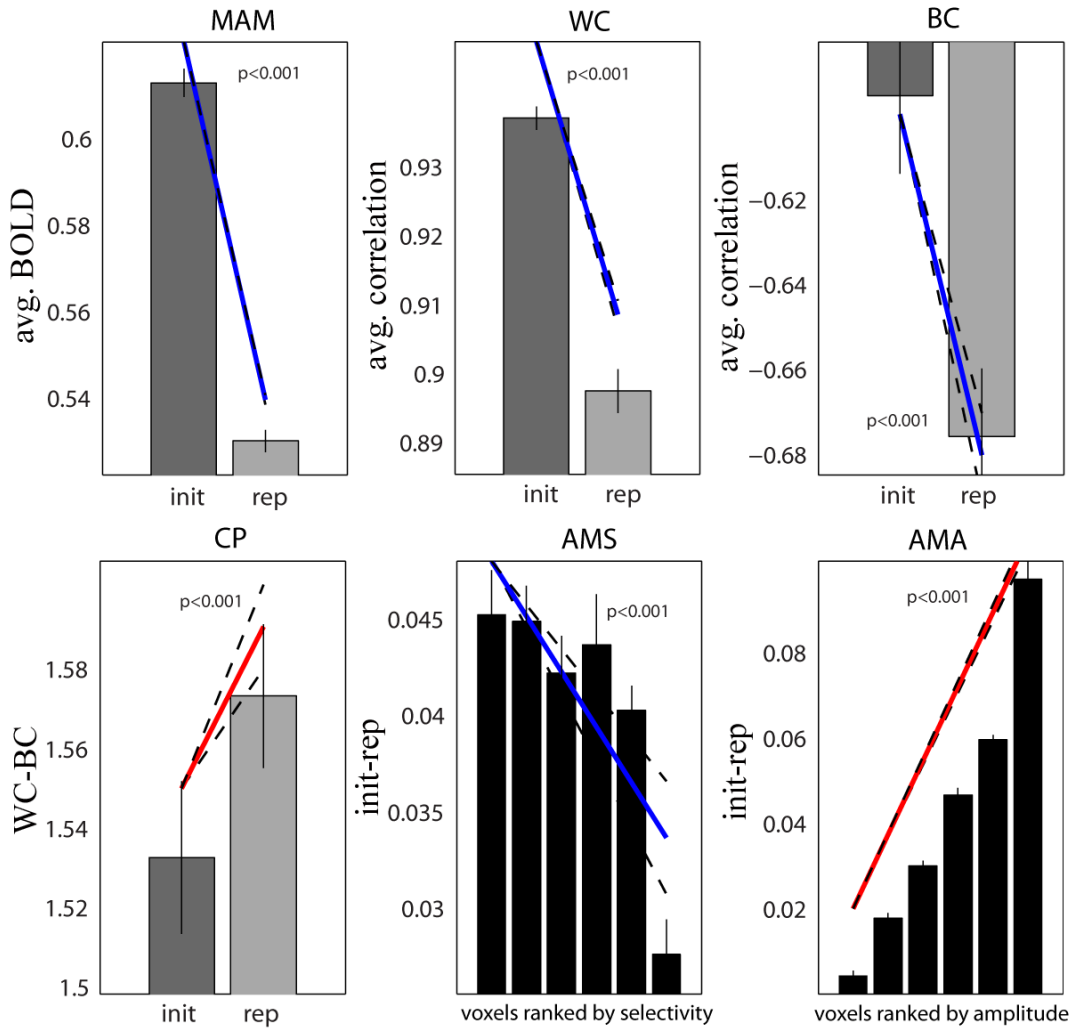


Figure 6.7 Predicted data features from averaging 18 simulation runs using the winning model (local scaling), with parameters: *a=0.8, b=0.4, σ=0.4* (cf. data in Figure 6.2) to explain the behaviour of the average participant (fixed effect). Apart from the correlations, the Y-axes have arbitrary units. All trends were significant p<0.001; error bars are CI 95%.

6.6 Discussion

The repetition-related changes in V1 for blocked, alternating presentation of visual gratings revealed a somewhat different pattern to those in FFA for the face paradigm in Chapter 4. While four of the data features (MAM, WC, BC and AMA) were similar across experiments, CP and AMS showed the opposite pattern across

experiments. Indeed, the V1 results for CP and AMS agreed with those of Kok et al., (2012), suggesting that, at least for these two features, repetition and expectation effects are similar.

Although the trial sequences in Exp.2 were arranged in a predictable manner, it is still possible that there are psychological reasons why the subjects' expectations of the upcoming trial types were different at the start of each subrun than that at its end (Appendix 4). Nevertheless, whatever the correct explanation of the data, our modelling strategy and conclusions from the simulations would still be valid, since we have simulated the exact trial sequence and paradigm of the fMRI data.

More importantly, our modelling shows that, despite empirical differences across the two experiments (Exp.1 & Exp.2), it is still only the local scaling model that can simultaneously fit all six data features. This not only reinforces the likelihood of local scaling being a common mechanism across the brain, but also demonstrates the flexibility of the local scaling model (depending on parameter values and paradigm details).

This raises the question of what makes CP and AMS differ so drastically across the two experiments: is it the circular stimuli, the stimulation protocol, or differences in model parameters (a, b, and σ)? From parameter-free results in Figures 5.3 and 6.5, we can conclude that experimental paradigm alone has an impact on the model predictions. Likewise, from the parameter-constrained Figures 5.4 and 6.6, we can conclude that the model parameters *a, b,* and σ also impact model predictions. However, the single most important parameter that determines the direction of both CP and AMS in the local scaling model is the σ (sigma) parameter (initial tuning width of neural tuning curves). For both experiments, when σ is less than approximately 0.3 (with average values of a and b), repetition reduces CP and increases AMS, and vice versa when σ is more than 0.3. (For the full set of the winning parameters see Appendix 6.)

The effect of σ on AMS is difficult to intuit, but it is important to clarify because it reveals an important factor that is easy to miss when translating brain activity from voxel space to neural space. I will illustrate it further because it may resolve some of the misinterpretations of fMRI results in literature (Kok et al., 2012; Summerfield &

de Lange 2014) and also some of the diverging fMRI results in terms of RS amount and voxel selectivity in different ROIs (Weiner et al., 2010; Utzerath et al., 2017; Krekelberg et al., 2006).

Figures 5.2 and 6.7 (AMS criterion) only show the relationship between selectivity and RS. More insight can be gained by plotting the BOLD amplitude for initial and repeat trials separately as a function of selectivity. Figure 6.8-A & B shows results for both datasets (but with more bins for greater resolution) for Experiment 1 (face data, left panel) and Experiment 2 (gratings data, right panel), while Figure 6.8-C & D shows corresponding results from the local scaling model, with parameters $\sigma=0.2$, $a=0.7$, $b=0.2$ for Experiment 1 and $\sigma=0.4$, $a=0.8$, $b=0.2$ for Experiment 2.

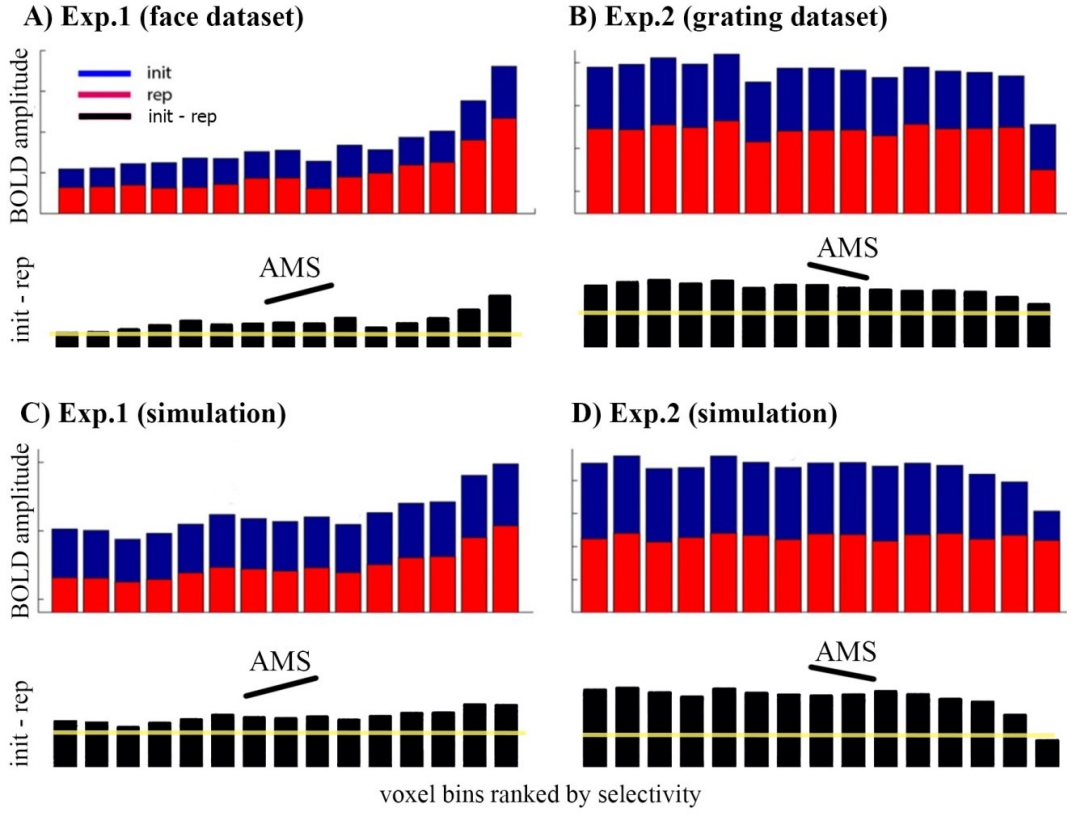


Figure 6.8: Voxel bins are ranked by their absolute t-values for the initial responses (init), the repeated responses (rep), and their differences which represent AMS (black bars). A & B panels show the results of fMRI datasets while C & D panels show the simulated results using the local scaling model and the following fitting parameters $\sigma=0.2$ $a=0.7$ $b=0.2$ (Exp.1) and $\sigma=0.4$, $a=0.8$, $b=0.2$ (Exp. 2).

It is clear that the increasing AMS profile in Experiment 1 is accompanied by increased overall response to both initial and repeated trials as selectivity increases, whereas the decreasing AMS profile in Experiment 2 is accompanied by decreased overall response to initial and repeat trials.

To understand better, I plotted heat maps of the neural tuning curves (Figure 6.9 and Figure 6.10) within two voxel types that vary in the number of neural populations that are selective to S1 and S2, as a function of narrow tuning ($\sigma=0.2$ as in Experiment 1) or broad tuning ($\sigma=0.4$ as in Experiment 2) values of σ. By chance, some voxels, such as voxel-1, will have a large number of neural populations selective for S1, while others like voxel-2 will have populations only partially selective for S1. For initial presentations and S1 and S2 (e.g. in gratings paradigm), voxels like voxel-1 will have higher selectivities (absolute difference in response to S1 vs S2) than voxels like voxel-2. However, after repetition of S1 and S2, the relative pattern of selectivities across voxel-1 and voxel-2 depends on σ.

When σ is small (Figure 6.9), the effect of local scaling from S1 and S2 is to dampen many neuronal populations in voxel-1 but fewer in voxel-2. Nonetheless, because the populations remaining less suppressed in voxel-2 are not selective for S1 or S2, the rank ordering of selectivity across voxel-1and voxel-2 remains the same as before adaptation (and so therefore does the average selectivity across initial and repeat trials, which corresponds to the x-axis in Figure 6.8). The response to initial and repeat trials (averaging across S1 and S2) is highest for voxel-1 (3.0 and 1.8) and lower for voxel-2 (0.685 and 0.44), and their difference (RS) is also greater for voxel-1 (1.2) than voxel-2 (0.245). This increase in overall response and increase in RS with increased selectivity is plotted at the bottom left of Figure 6.9 and resembles the data for the faces in Experiment 1 (left panels in Figure 6.8).

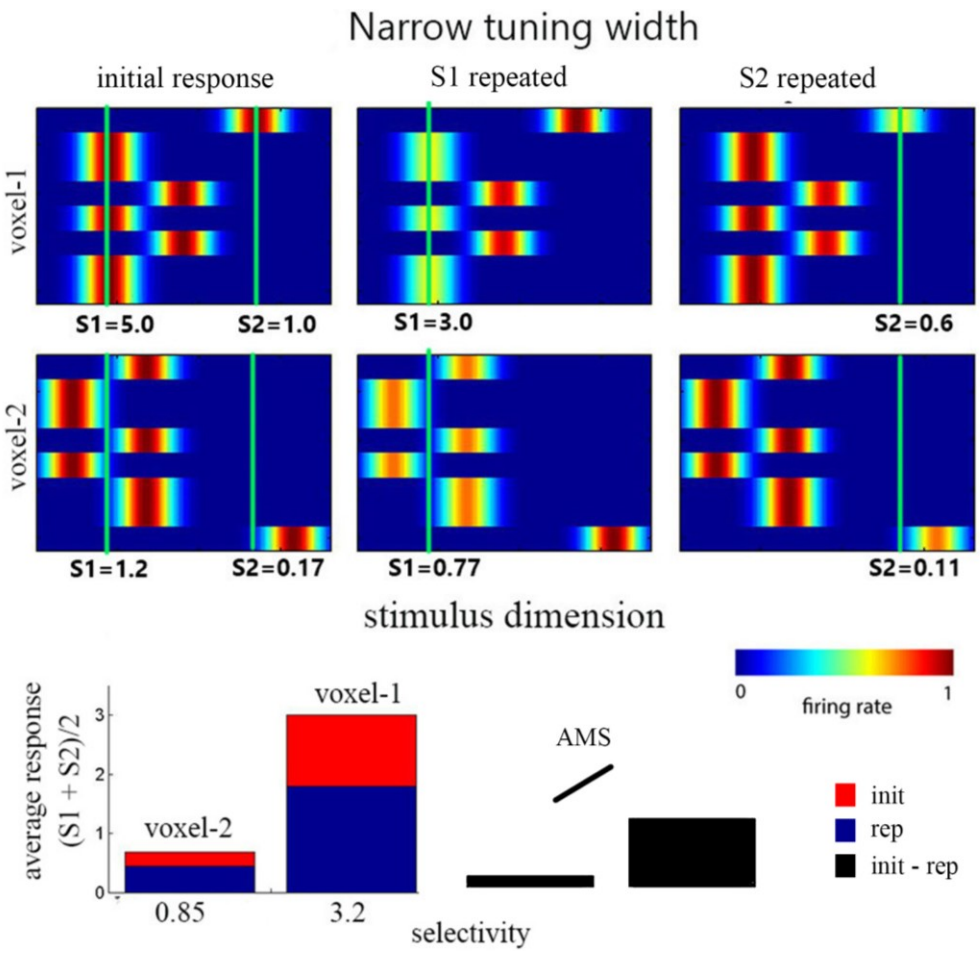


Figure 6.9: Example of two voxel types each possessing different underlying neural populations with narrow tuning widths (σ=*0.2*). Numbers at the bottom of each voxel represent the summed neuronal responses for S1 or S2 at the green line. Bars at the bottom are ranked by selectivity and shows the mean response to S1 & S2 in both initial (init) and the repeated response (rep). The black bars are the difference (init-rep) (see the text for details)

In case of the large σ (Figure 6.10), there is more overlap between the tuning curves (Figure 6.10). The effect of local scaling from S1 and S2 is still to dampen more neuronal populations in voxel-1 than voxel-2. Nonetheless, even though more tuning curves are suppressed in voxel-2, the amount of the suppression is smaller because they are further from the adapting stimulus. In addition, their tuning curves still overlap more with one stimulus (S1) than the other (S2). This means that the selectivity of voxel-2 voxels can actually surpass that of voxel-1 after adaptation, and in the above example the average selectivity across initial and repeat trials (measured by absolute difference in response to S1 vs S2) is higher for voxel-2 (3.21) than

voxel-1 (3.17). However, the response in voxel-2 to initial and repeat trials (averaging across S1 and S2) is lower (2.48 and 1.90) than in voxel-1 (3.3 and 2.03), as is the amount of RS, i.e, lower in voxel-2 (0.58) than voxel-1 (1.27). This decrease in overall response and decrease in RS with increased selectivity is plotted at the bottom right of Figure 6.10 and resembles the data for the gratings in Experiment 2.

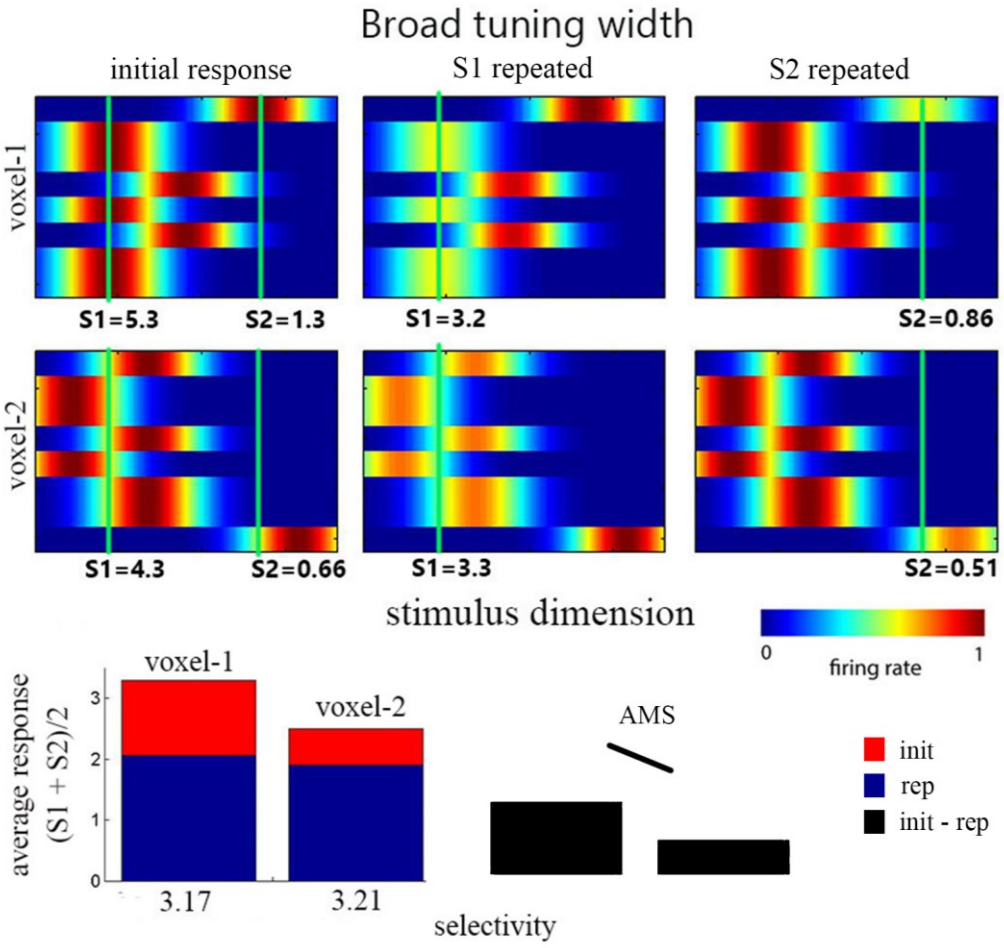


Figure 6.10. Shows the same voxels in Figure 6.10 but with broad tuning widths (σ=*0.4*). (see Figure 6.9 legends and the text for details)

The distributions in voxel-1 and voxel-2 showcase the differential effect of local scaling and σ parameter on the overall voxels' selectivity. Other neural distributions within the voxels might produce slightly different results but the general principle remains similar, which is that local scaling pushes the voxels with a high proportion of selective neurons backward in the overall voxel selectivity ranking, while voxels with a large proportion of partially selective neurons are pushed forward in the

overall voxel selectivity ranking, and this effect scales with the neural tuning widths. This demonstrates the difficulty of mapping voxel's overall selectivity to the ratio of selective and non-selective neurons within each voxel, especially in repetition paradigms.

To explain the difference in CP trend across the two experiments, it is important to note that CP depends on the difference between WC and BC. Since both experiments exhibit a repetition-related reduction in both WC and BC, a repetition-related increase in CP arises when BC is reduced more than WC (and conversely, a repetition-related decrease in CP arises when BC is reduced less than WC). When σ is low, a greater number of voxels have a selectivity for one stimulus (by chance), and so when these are suppressed, there is a considerable decrease in WC after repetition (because highly tuned voxels are suppressed more), but a less considerable decrease in BC after repetition. On the other hand, when σ is large, there a fewer selective voxels and hence these are suppressed less, and there is less reduction in WC after repetition and a more prominent reduction in BC (compare BC after repetition across Figures 5.5 and 6.7). Therefore CP depends on σ, again explaining the difference across the two experiments.

The above considerations assume that tuning curves in V1 for Experiment 2 are broader than the tuning curves in FFA for Experiment 1. I am not aware of any studies that have compared the neural tuning widths in V1 and FFA, but such comparisons will depend on the nature of the stimulus dimension. For instance, in FFA, there is unlikely to be a stimulus dimension that ranges from faces to scrambled faces, with various semi-scrambled faces in-between, because we are not typically exposed to degrees of face scrambling in everyday environments (more likely are dimensions related to the gender of the face, for example, as suggested by principal component analyses of face images, e.g. Burton et al., 2016). Indeed, while orientation may be represented along one (circular) dimension, it is likely that faces are represented along more than one dimension, and it is not clear how to compare tuning curves with different dimensionalities. Future studies using more controlled stimulus dimensions (e.g, with continuous sampling, e.g. face morphs) may be able to more directly study the effects of repetition on voxel/neural selectivities.

6.7 Chapter summary

In this chapter, I validated the modelling results in Chapter 5 with an independent dataset that shows a different pattern of fMRI results to those in Chapter 5, and found that local scaling model can still fit the qualitative results. I showed that these differences in the RS-related fMRI findings can be attributed to the width (overlap) between the neural tuning curves. In the next chapter, rather than considering how classification of two stimulus types is affected by repetition, I consider how well initial and repeated presentations can be classified, regardless of stimulus type. It turns out that this classification performance cannot be explained by the current local scaling model, and requires additional assumptions about the correlated trial-to-trial variability discussed in Chapter 2.

CHAPTER 7: REPETITION EFFECTS ON THE VOXEL PATTERN

7.1 Introduction

Whereas the previous chapters focused on classification performance for two stimulus classes (faces vs scrambled faces or gratings of 45^0 vs 135^0) as a function of repetition, this chapter focuses on classification of initial versus repeated presentations (regardless of stimulus type). The previous chapter showed that a uniform change in the neural firing rate across the stimulus dimension cannot explain all the repetition-related findings in fMRI; instead one needs a non-uniform scaling that is inversely proportional to the distance from the adaptor. In this chapter, I explore adaptation across the spatial dimension, i.e voxels.

Because the tuning-curves are randomly distributed within each voxel, the adaptation from local scaling will be non-uniform across voxels as well. This will lead to a change in the distributed voxel pattern of the repeated presentation from that of the initial presentation, even if the initial selectivity were similar across the voxels. To

illustrate this, imagine two voxels, voxel-1 and voxel-2, each with the same initial response to stimulus S1, but with different underlying distributions of neural preferences (Figure 7.1).

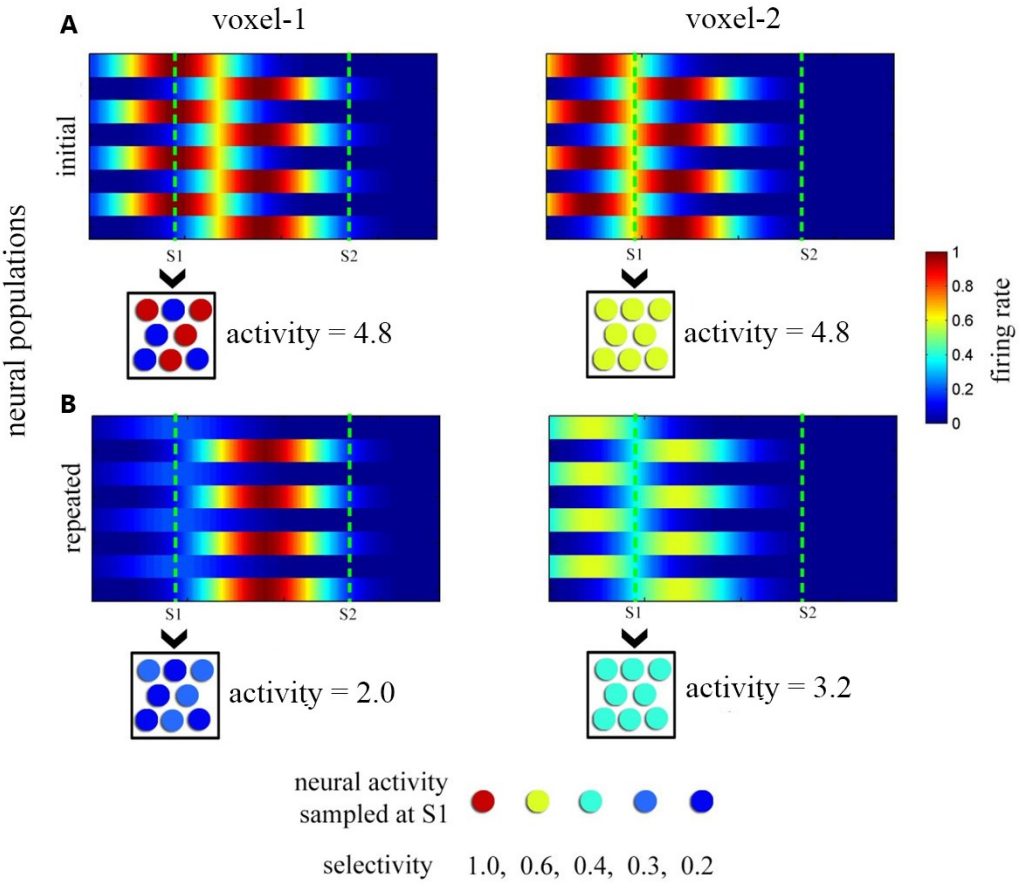


Figure 7.1: Illustrates the effect of local scaling on two different voxels (voxel-1 and voxel-2) with different underling distributions of neural preferences. The large squares show the neural populations within each voxel in the initial phase (Panel-A) and after adaptation with a local scaling model (Panel-B). The small squares show the total voxel activities after summing the firing rate for all the neural populations at S1.

Voxel-1 has 8 neural populations, 4 highly selective to S1, firing with rate of 1, and 4 highly non-selective toward S1, firing with rate 0.2. Therefore, the summed activity for voxel-1 is 4.8. Voxel-2 also has a summed activity of 4.8, but because of 8 partially-selective neural populations, with firing rate to S1 of 0.6. Thus voxel-1 and voxel-2 respond equally to initial presentation of S1. However, after local scaling, their responses differ (even with the same adaptation parameters a, b, and σ). This is

because the different initial distribution in each voxel results in a different adaptation factor, *c*, for each voxel. Specifically, voxel-1 will respond less than voxel-2 because its highly selective neurons are suppressed more. However, the question is whether these changes across voxels are systematic enough to enable a pattern classifier distinguish between the initial and the repeated presentations.

Rissman et al. (2010) used MVPA to classify novel faces from repeated faces and concluded that the main driver of classification was the difference in mean ROI response (i.e, the univariate effect of RS). However, it is more interesting if a classifier could distinguish novel from repeated stimuli even when the mean across voxels is removed from each voxel and furthermore, even when the voxel responses are re-scaled to have the same SD, because this would indicate that repetition produces a systematic change in patterns beyond just a scaling of the patterns for initial presentations. An example of this is shown in Figure 7.2, where this systematic change allows a classifier to insert a hyperplane between the two categories (obviously the example is chosen is an extreme, and reality there would be some overlap between trials in a higher dimensional voxel space). Figure 7.2 suggests that local scaling indeed can produce such systematic changes because, unlike global scaling, it produces a change in the voxel pattern that cannot be restored after Z-scoring (see Figure 1.3 in Chapter 1 for comparing uniform and non-uniform RS effects).

The previous chapters concluded that local scaling model provides the best qualitative fit to the effects of repetition on classifying stimulus types, and since the simulations in Figure 7.2 suggest that local scaling may also be able to classify initial versus repeated presentations, even after Z-scoring, I decided to test this in real data. In particular, I compared classification of the initial versus repeat in terms of both 1) the univariate mean across voxels (henceforth "Pattern Mean", PM) and 2) the multivariate pattern across voxels, after removing the univariate RS effect and scaling (henceforth "Z-scored Pattern", ZP).

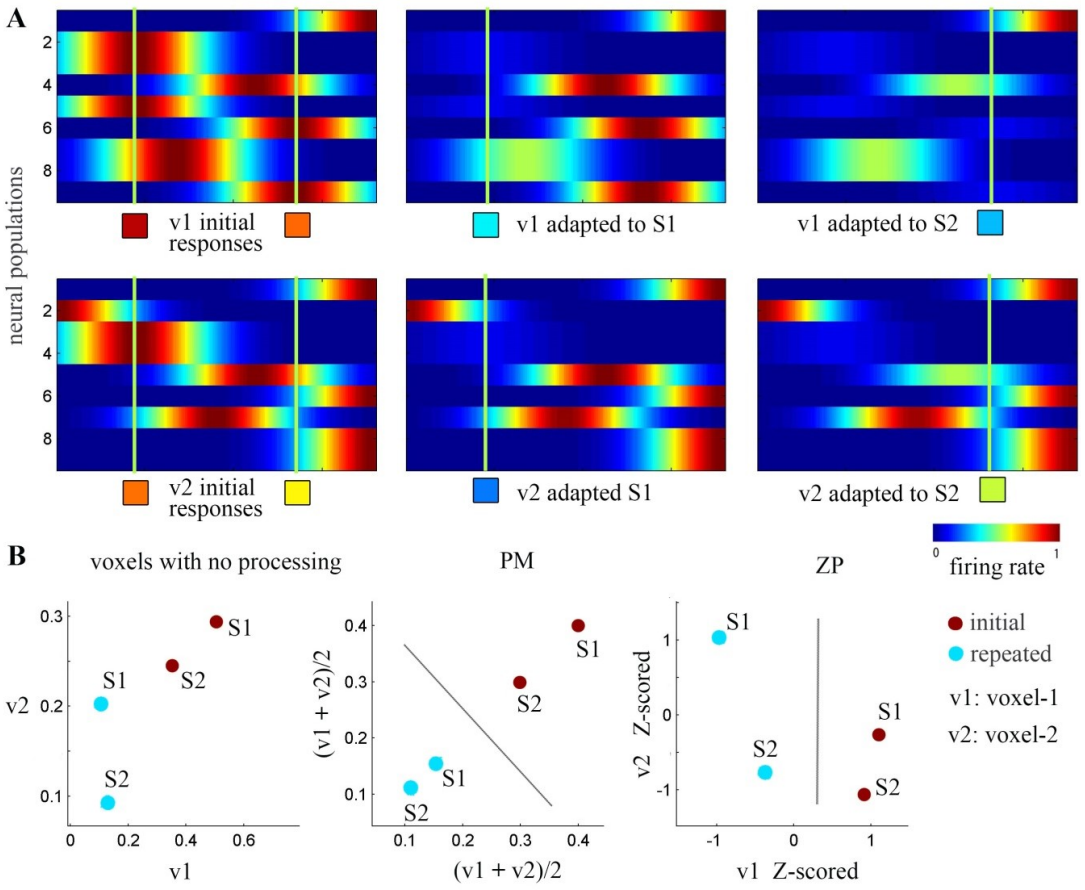


Figure 7.2 illustrates the effect of local scaling model on Pattern Mean (PM) and Z-scored Pattern (ZP). Panel A) shows the effect of local scaling on two different voxels: voxel-1 (top row), and voxel-2 (bottom row). Each voxel has a different underlying distribution of neural preferences. (Note that the neurons in the right most panels have been adapted to both S1 and S2, though in this case the results were similar even if adaptation was released for S1). The small squares show the average voxel response coloured by their relative activity (hotter means more active). Panel B) plots the voxel responses for each stimulus before and after adaptation (left panel), then after averaging across the voxels to illustrate PM (middle panel) and after Z-scoring across the voxels to illustrate ZP (right panel). The solid lines are possible decision lines between the initial and the repeated presentations.

7.2 Methods

I used a SVM with leave-one-run-out (like in previous chapters) to classify initial versus repeat presentations of faces, collapsing across famous and scrambled faces (using data from Chapter 4; henceforth "Experiment 1") and initial versus repeat presentations of gratings, collapsing across orientation (using data from Chapter 6; henceforth "Experiment 2").

I started by examining CP based on the mean across voxels within an ROI for each trial (i.e PM). This indicates how well trials can be classified in terms of initial versus repeats just in terms of overall activation (i.e, based on consistency of RS across stimuli). To then examine how much more information is present in voxel patterns, I compared CP based on PM with CP based on the pattern after Z-scoring across voxel (i.e. ZP).

I started with data from the FFA ROI for Experiment 1 and the V1 ROI for Experiment 2. Then to compare CP for initial versus repeat trials in these data with the local scaling model, I used the same simulations as in Chapters 5-6 (with 200 voxels per ROI). The only difference is that I increased the amount of Gaussian noise added to each voxel from SD=0.1 to SD=0.3, in order to reduce CP to closer to empirical levels, given the same number of trials as in the experiments (note that this noise is independent across voxels and trials)[6]. For the parameter of the local scaling model, I used the wining values reported in Appendix 6.

[6] In additional analyses, I varied the stimulus pattern for each class by sampling the neural activity from a Gaussian distribution centred at each stimulus type. This produced a variation in the stimulus pattern that was qualitatively similar to that of simply adding Gaussian noise.

7.3 Results of the data and Model prediction

The results of the ROI analyses for both Experiments are summarized in Figure 7.3 (panels A & C). Firstly, the data and the model showed that classification of initial versus repeated presentations can be above chance, whether based on PM or on ZP. Furthermore, ZP-classification yielded better performance than PM-classification in both datasets ($t(17)=4.65$, $p<0.001$ for the face dataset, $t(17)=4.01$, $p<0.001$ for the grating dataset). However, for the simulations, the results were the opposite: for all the wining parameter values, the local scaling model always produced better PM-based classification than ZP-based classification (Panels B & D). Note that absolute levels of CP cannot be compared, because these can always be adjusted in the model simply by changing the amount of independent voxel noise.

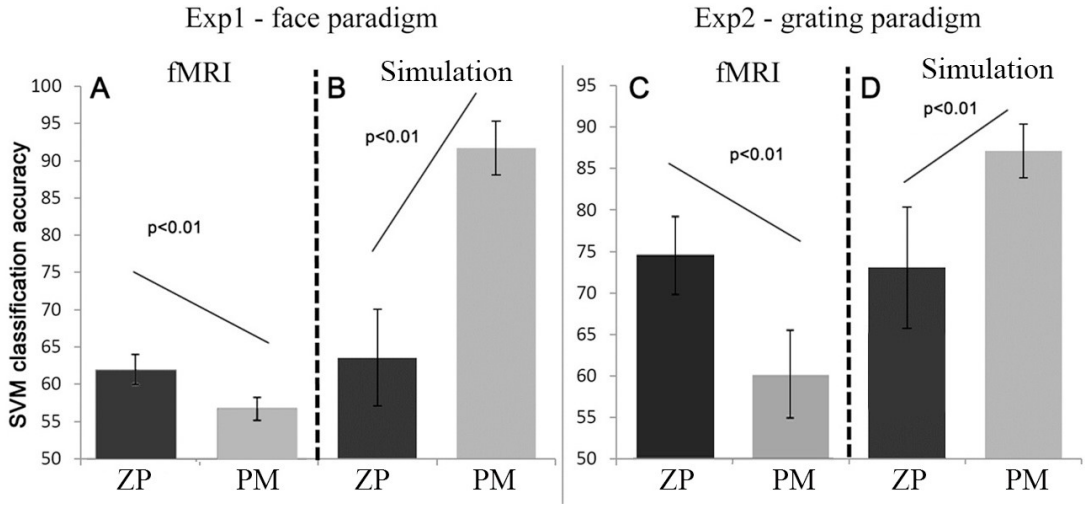


Figure 7.3: Compares the CP of ZP and PM in fMRI and simulation for both experiments. Parameter values used for simulations are the same as those reported previously in Figures 5.5 and 6.7.

7.4 Discussion and Model revision

In both experimental datasets, classification of initial versus repeat trials was better when using the Z-scored pattern information (ZP) than when using the mean across voxels (PM). Yet the local scaling model (using parameter values that were optimal for each experiment) predicted the opposite, i.e, better classification for PM than ZP. This disparity could be due to two reasons: 1) the model produces a lower ZP than it should; or 2) the model produces a higher PM than it should.

For the model to produce a higher ZP, repetition must increase the variance across voxels (Davis et al, 2014), i.e apart from the difference in mean response, the pattern for repeat trials need to become less similar to initial trials.

In Figure 7.2, I already explained one way in which local scaling changes the pattern from that of the initial voxel pattern. Even though the adaptation amount, *a*, is the same across voxels, the final adaptation amount, *c*, varies across voxels because they have different underlying tuning-curves (and hence *a* is applied non-uniformly across the stimulus dimension). One way to increase the variance after repetition is to draw *a* for each voxel based on a Gaussian distribution centered on the winning parameter value (but bounded in the range 0 to 1). This will exaggerate the resulting dissimilarity between initial and repeated voxel patterns and potentially increase ZP.

Figure 7.4 shows the model results for ZP and PM for several values for the SD of *a* and the SD of the noise. In both paradigms, CP for ZP increases proportionally to the variance of *a*. However, ZP CP never reaches PM CP, even with when the variance in *a* (SD=0.2) is high enough to cause CP to reach ceiling. Adding more Gaussian noise to prevent this still does not change the relative pattern of CP for ZP versus PM. Increasing the variance of *a* further still impairs the model's ability to fit the other 6 data features from the previous chapters. In fact no model among the 12 models survives when *a* SD > 0.2 in both paradigms.

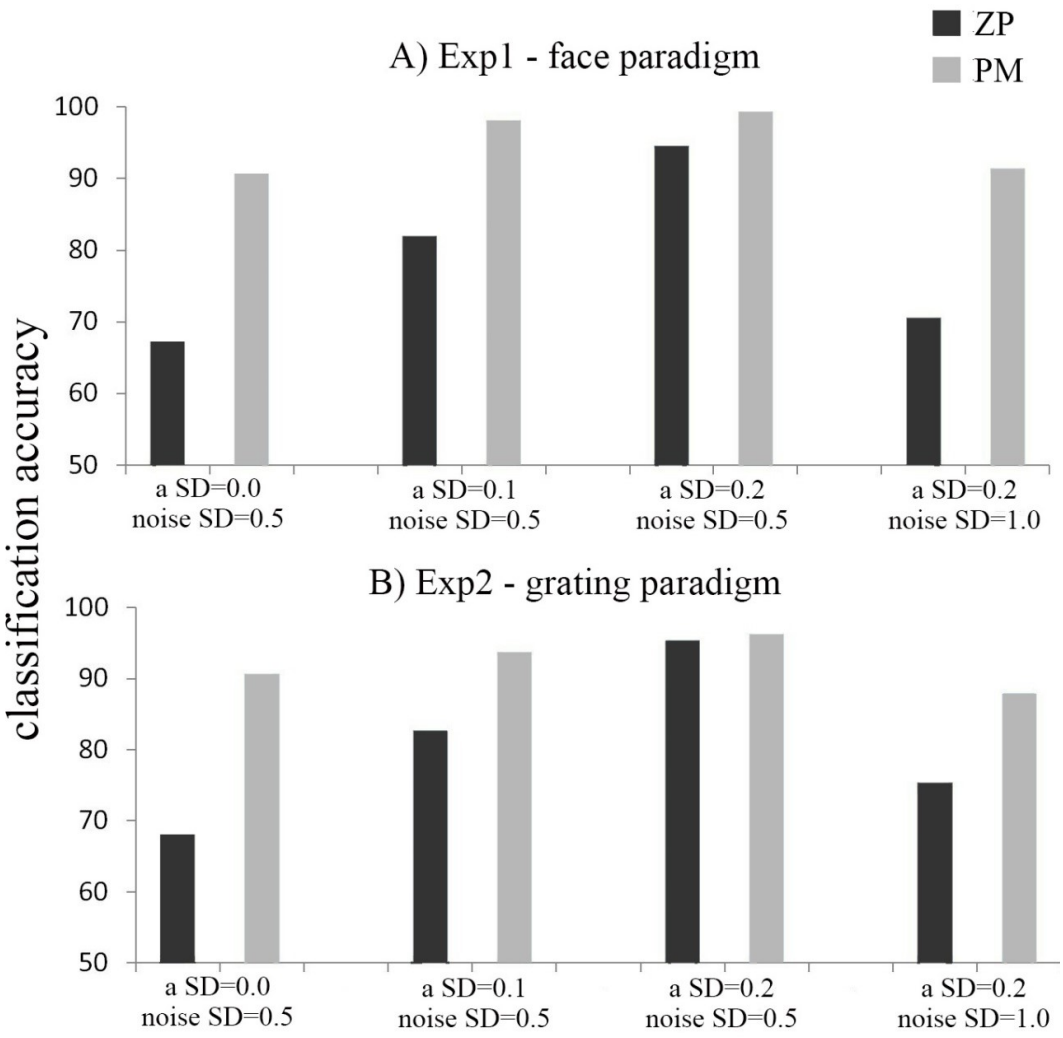


Figure 7.4 compares the local scaling model prediction between the CP of PM and ZP for various levels of *a* and noise variance. The mean *a* was 0.7 for face paradigm and 0.8 for the grating paradigm. 18 subjects were simulated; differences are significant at $p<0.001$.

To try to better understand why increasing the variability of *a* does not enable ZP CP to exceed PM CP, I inspected the pattern across voxels for both empirical datasets after averaging across all trials in either initial and repeat conditions for a typical subject. The results for the gratings are shown in Figure 7.5 (results for faces were very similar). The most obvious result is that voxels with highest initial amplitude are suppressed most, which is in line with the AMA results in Chapters 4-6. The next observation is that no voxels showed an increased response after repetition (contrary to De Gardelle et al., 2012), which suggests that *a* cannot vary so much as to be greater than 1. Most importantly, there is a high similarity (correlation >0.9) between

the patterns for initial and repeated presentations for both the data and the model, even after Z-scoring. The high similarity between data and model patterns suggests that ZP may not be the reason for the discrepancy between data and model in terms of the relative performance of CP based on ZP and CP based on PM. So instead, I explored the second possibility mentioned above, i.e, that PM CP is too high in the model.

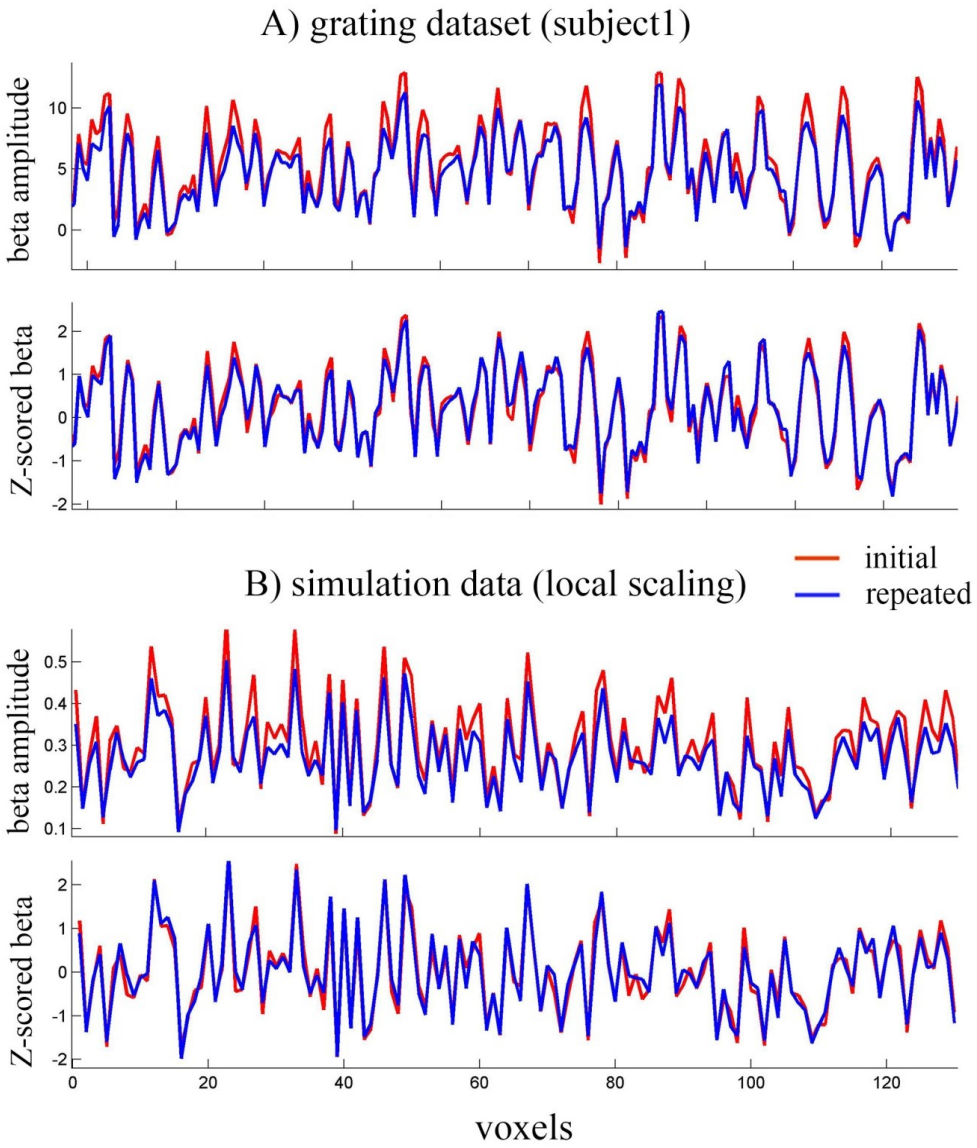


Figure 7.5 shows the voxel pattern (x-axis is voxel number) averaged across the trials for initial (blue) and repeated (red) presentations in one example subject and for simulations of local scaling (using winning parameters values in Figure 6.7).

Rather than varying *a* across voxels, it is possible to vary *a* across trials. This might reflect variations in attention across trials or stimuli, which affects the amount of adaptation (e.g, more attention given to the initial presentation might increase subsequent adaptation; e.g., Henson & Mouchlianitis, 2007; Moore et al., 2013). This extra variability across trials will increase the overlap in mean responses for initial and repeated presentations, reducing CP based on PM. However, because this trial-variability that is coherent across voxels, it does not harm CP based on patterns (i.e, ZP), as explained in Chapter 2 (since pattern classifiers like SVM depend on the relative activity across the voxels). Therefore, varying *a* across trials will reduce PM but potentially increase ZP by increasing coherency across voxels.

In reality, attentional fluctuations that affect *a* are likely to apply to both initial and repeated presentations (since varying *a* only affects repeated presentations). In other words, such fluctuations are better simulated by adding coherent noise across voxels to all trials (rather than the independent noise assumed so far, e.g, in Figure 7.3). The results are shown in Figure 7.6. Note that the SD of the noise (SD=0.05) was also increased (from SD=0.3 in Figure 7.3), so as to reduce CP based on the PM. However, because this noise is now correlated across voxels, it does not harm ZP, such that ZP is now higher than PM in the model, in order to match the qualitative pattern in data (again note that quantitative fits would require optimizing the variance and covariance of the noise across voxels). To check that this introduction of coherent (rather than independent) voxel noise did not affect any of the results in the previous chapters, I re-ran the grid search analyses for all the 12 models. The results were similar and the local scaling model remained the only model able to fit the data.

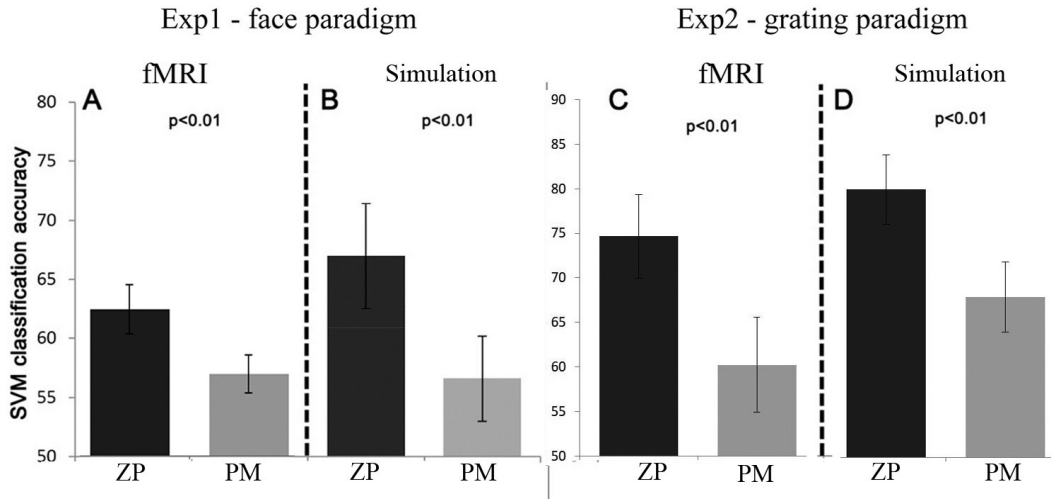


Figure 7.6 compares the CP of ZP and PM in fMRI and simulation for both experiments. Unlike Figure 7.3, Panels B & D now have coherent trial variability SD=0.05.

7.5 Summary

In this chapter, I explored the ability to classify initial versus repeated presentations, regardless of stimulus class. I showed that according to a local scaling model, repetition can produce a systematic change in voxel patterns so as to enable above-chance classification. However, when I compared classification performance based on the mean across voxels (PM) with that based on the mean-corrected and scaled pattern (ZP), the former was higher than the latter in the model, but the latter was higher than the former in both datasets. I showed that this discrepancy cannot be resolved by varying the amount of adaptation across voxels, but it can be resolved by increasing the coherency of noise across voxels (as might be caused by variability across trials in the amount of attention and/or adaptation). In the next, chapter I summarise the results across all chapters.

CHAPTER 8: CONCLUSIONS

8.1 Summary of thesis

This thesis aimed to investigate the neural correlates of stimulus repetition using fMRI. This investigation ranged from mean responses to single trial estimates to voxel patterns and finally to simulated neural tuning-curves. Along the way, I addressed the temporal and the spatial limitations of fMRI, and ended up with several conclusions that affect future fMRI analyses in general, and more specifically how best to study and interpret repetition-related effects in fMRI.

Chapter 1 reviewed the current theories of RS and their potential relations to memory and learning, and emphasised the importance of uncovering the mechanisms underlying RS in humans using a non-invasive neuroimaging tool like fMRI. I also summarised the challenges related to the temporal and spatial limitations of fMRI.

Chapter 2 tackled the temporal limitations in fMRI and examined the optimal fMRI designs for single trial estimation given different sources of variability (both at the level of trials and scans), and concluded that the designs optimal for univariate analyses are not necessarily optimal for multivariate analyses. Instead, different GLM models are better in different situations that depend on the ratio of scan-variability to trial-variability, and the coherency of these types of variance across voxels. In general, the LSA model is better when scan-variability is low or its coherency across voxels is high, while the LSS model is better when trial-variability is low or its coherency across the voxels is high.

Chapter 3 used efficiency measures similar to those in Chapter 2 in order to compare the GLM models on real data from a publically-available face repetition fMRI dataset. In general, LSS appeared to the optimal model for that dataset, which was either due to highly incoherent scan-variabilities (evidenced by the lower SD of LSS estimates across the independent runs) or highly coherent trial-variabilities (evidenced by the higher trial correlations across the voxels in LSS estimates), or a combination of both. The single trial betas from LSS were used in Chapter 4 for a series of univariate and multivariate fMRI analyses in which I formally defined six repetition-related metrics: MAM, CP, WC, BC, AMA, and AMS. These metrics capture various univariate and multivariate aspects of fMRI repetition effects, and when taken together, can be informative about the underlying neuronal adaptation mechanism.

Chapter 5 began to tackle the spatial limitation of fMRI by comparing forward models that use Gaussian tuning-curves to simulate the neural population firing rates, together with parameters that control the amount and distance-sensitivity of neural adaptation. 12 neural adaptation models were proposed, and a grid search of parameter values concluded that only one of them, the local scaling model, could fit all six repetition-related metrics identified in Chapter 4 when using the same parameter settings. Chapter 6 confirmed that the local scaling model was again the only one of the 12 models that could explain the same set of repetition-related metrics in an independent fMRI dataset that used different stimuli, a different protocol and different ROIs (and in which two of the repetition-related effects differed from those in Chapter 5). The simulations in Chapters 5 and 6 illustrate the value of using forward models that map from neurons to voxels, like those considered here, to interpret fMRI data, i.e., map back from voxels to neurons. Despite the simplicity of the models considered, their predictions are not always intuitive, and they therefore help protect against superficial analogies, for example that sharpening of multivoxel fMRI patterns entails the sharpening of neuronal tuning curves.

Finally, Chapter 7 explored the effect of repetition on voxel patterns in order to explain how initial and repeated presentations can be distinguished, independent of the stimulus class. Here, the local scaling model (nor any other of the 12 models)

could not explain how classification performance for initial versus repeated presentations based on the mean-corrected pattern exceeded classification performance on the mean across voxels (i.e. based on univariate RS). Further investigation however suggested that coherent trial-variability (as considered in Chapter 2) could explain this difference in relative classification performance, which was confirmed by simulating the local scaling model together with random trial-variability that was coherent across voxels. Taken together, this thesis's concept of trial-variability, coupled with its formal modelling of neural models of fMRI multi-voxel responses, have significantly advanced our understanding of repetition-related neural mechanisms.

8.2 The Local Scaling Model

The local scaling model is the crowning model in this study, and maps back to the concept of selective suppression illustrated in Figure1.1D of Chapter 1. How does this finding advance our understanding of neural adaptation? We already knew from single-cell recording studies that the repetition-related reduction in neural firing depends on a neuron's preference for the repeated stimulus (Desimone, 1996; Ringach et al., 2002; Kar & Krekelberg, 2016). What I have done is formalize that dependency in terms of three parameters, which implement a nonlinear function in terms of a piecewise linear approximation, extending previous scaling models that ignored this dependency (Weiner et al., 2010; Andresen et al., 2009; Hatfield et al., 2016). But how is this function implemented; i.e., how does the neural adaptation "know" the difference between the neuron's preference and the repeated stimulus?

One possibility is that the amount of adaptation depends on the initial firing rate of a neuron, where this initial firing rate indicates how close the stimulus is to the neuron's preferred stimulus. This activity-dependent scaling, or "fatigue" mechanism (Grill-Spector et al, 2006), may occur because neurons that fire more experience a greater decline in their synaptic resources, and hence are less able to fire subsequently (Abott et al., 1997). This was the model proposed by Andresen et al., (2009), and can be expressed in my formalism as:

$$f_i(j_2) = cG(\theta_{j_2}, \mu_i, \sigma) \qquad c = 1 - aG(\theta_{j_1}, \mu_i, \sigma) \qquad 0 < a \leq 1$$

where j_1 / j_2 refer to first and second presentations of stimulus *j*. In other words, if $G(\theta_{j_2}, \mu_i, \sigma) = G(\theta_{j1}, \mu_i, \sigma) = G$, then $f_i(j_2)$ is non-linearly related to initial firing rate:

$$f_i(j_2) = (1 - aG)G = G - aG^2$$

Thus, for adaptation factor of a=0.8, whereas global scaling would predict that two neurons with initial firing rates of [1.0 0.5] and would fire at rates of [0.8 0.4] after repetition, with this (nonlinear) fatigue, they would fire at rates [0.2 0.3], i.e, the relative rate of firing across neurons would reverse in latter case (because the first neuron has a higher preference for the repeated stimulus). In loose terms, if global scaling means that "neurons that fire more, tire more" (in an additive sense), then activity-dependent scaling states "neurons that fire more, tire much more".

This fatigue model therefore has only two parameters, a and σ, rather than the three used for local scaling. I checked whether the fatigue model could explain the data features in Chapters 4-5 (using a grid search across the same range of values as the previous models). While it could simultaneously explain all six data-features in the face paradigm, it could not fit them in the grating paradigm.[7] In particular, it could not simultaneous produce an increased CP and a decreased AMS for any of the a and σ values examined. Thus the greater flexibility of the three-parameter local scaling model seems necessary to explain all the data-features across both datasets. Therefore, the winning local scaling model in this thesis cannot simply be reduced to activity-dependent adaptation, and it is likely to result from more complex neural/synaptic processes, such as interactions between neurons like those inherent in predictive coding models for example.

8.3 Caveats

There are several limitations and caveats associated with the work described here:

[7] The same was true when I added a piecewise nonlinearity to the activity-dependent adaptation (to mimic the nonlinear distance function used for local scaling), ie with $c = \max(b, 1 - aG)$ where $0 < b < 1$ is the maximum adaptation.

1. Although I detailed the efficiency of GLMs in relation to scan-variability/trial-variability ratio in our simulations, one cannot realistically dissociate between these two variability types in the real fMRI data, because the beta estimates will always include some influence of scan variability (random noise). However, one could infer crudely about the underlying noise structure by identifying the GLM that gives more stable and similar estimates across independent runs using the efficiency measures we discussed in Chapter 3.
2. A real fMRI data have many other sources of noise and variability that I did not model, like head motion, temporal drifts, session-to-session variability, etc. These sources of variability may not be Gaussian.
3. I only performed qualitative fitting, in terms of reproducing significant effects in the data. Future work could do proper quantitative fitting, which would likely require extra scaling parameters. Such quantitative fitting could compare more sophisticated estimates of goodness of fit, which take into account different numbers of free parameters in different models (e.g., Bayesian model evidence, or approximations like the Akaike information criterion or Bayesian information criterion) to formally compare different neural models.
4. The BOLD signal is likely to include components other than just the firing rates of large neurons normally measured in single-cell recording studies, e.g. the BOLD signal may relate more to LFP than MUA, include components from inhibitory interneurons, etc.
5. RS effect may reflect a combination of different adaptation mechanisms, or mechanisms beyond the scaling/sharpening/shifting mechanisms considered here.

8.4 Future studies

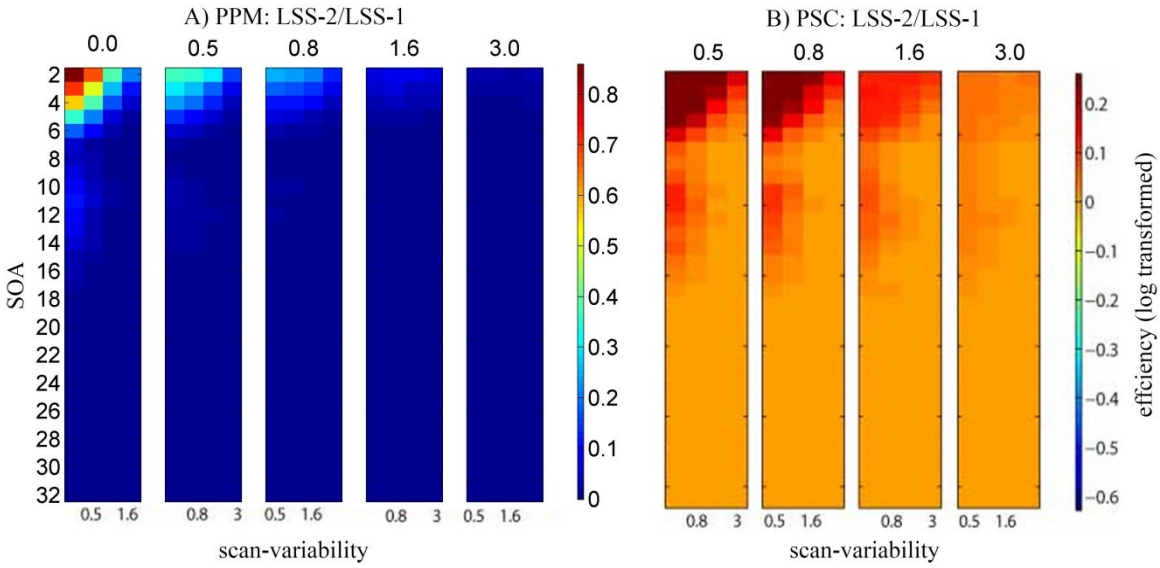
Our simulations demonstrate that repetition-related fMRI metrics vary with the experimental paradigm (e.g. compare Figures 5.3 and 6.5). In this thesis, I could not isolate the reason for these differences between the two paradigms because they differed in several ways. Future studies using more controlled stimulus dimensions (e.g, with continuous sampling, e.g. face morphs) may be able to more directly study the effects of repetition on voxel/neural selectivities, and could use our models to predict the possible outcomes for different paradigms before data are collected. Future studies could also address test some of the predictions by combining single-

cell and fMRI studies (e.g, in nonhuman primates) using the same paradigm and stimuli. Once the selectivity profiles of a number of neurons in an ROI are identified from single-cell recording, it might be possible to constrain the a, b and σ parameters, which could then be used to predict the fMRI repetition effects. Our simple mathematical models could also be used to guide more complex neural network models that relate more directly to neural firing rates and synaptic changes (e.g., Spigler and Wilson 2017). There are many exciting avenues to explore.

APPENDICES

Appendix 1

The ratio of PPM and PSC for LSS-2 to LSS-1 models is shown in supplimentary Figure 1. As expected, the simulations showed that distinguishing non-target trials by condition (LSS-2) is always better, particularly for short SOAs and low ratios of trial-variabilty to scan-variability. Therefore, in all subsequent analyses, I used the more standard form of LSS i.e LSS-N (Mumford et al., 2012) where N refers to the number of the conditions in the experiment.

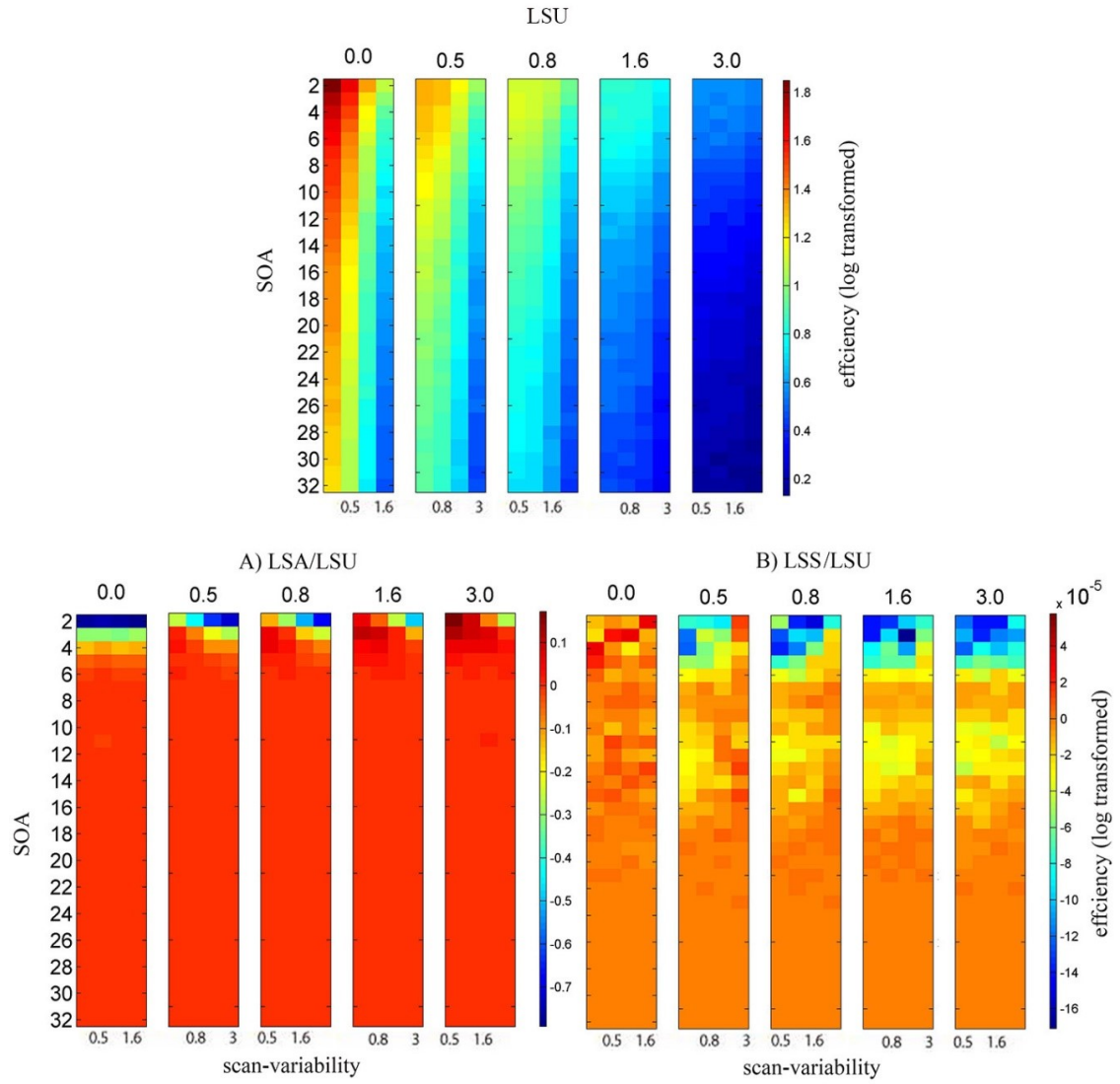


Supp. Figure 1. Efficiency comparison between LSS-N (here N = 2) and LSS-1 A) Ratio of PPM, B) Ratio of PSC LSS-2. See Figure 2.2 legend for more details.

Appendix 2

While LSU was not the focus of this thesis (because it does not provide estimates for individual trials), for completeness I compared the ratio of PPM for LSA to LSU and LSS to LSU models (supplimentary Figure 2) (the results are noisier than in the

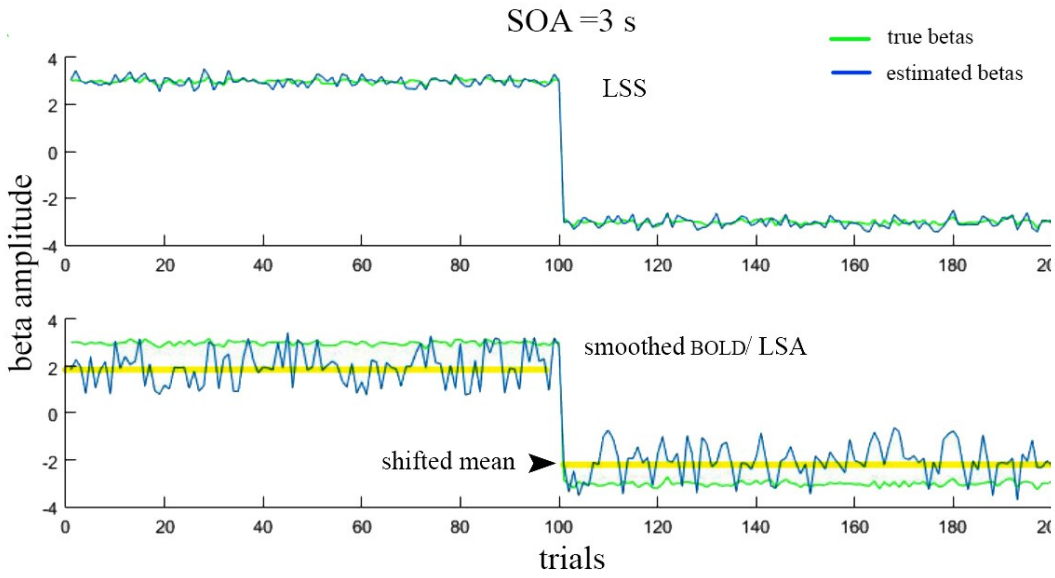
figures in Chapter 2 because I only used 1000 iterations). Neverthelsses, we can see that LSA/LSU is qualitatively similar to LSA/LSS (Figure 2.2) indicating that LSS behaves similarly to the standard LSU when estimitating the average beta. Indeed direct comparsion between LSS and LSU produced only tiny differences (indicated by the small scale on colourbar). Nonetheless, there is some evidence that LSU is superior at short SOAs, particularly when trial-variability is high, consistent with the empirical results in Figure 3.4 of Chapter 3.



Supp. Figure 2 shows PPM for A) LSU, ratio of PPM for B) LSA relative to LSU, and C) LSS (LSS-2) relative to LSU. See Figure 2.2 legend for more details.

Appendix 3

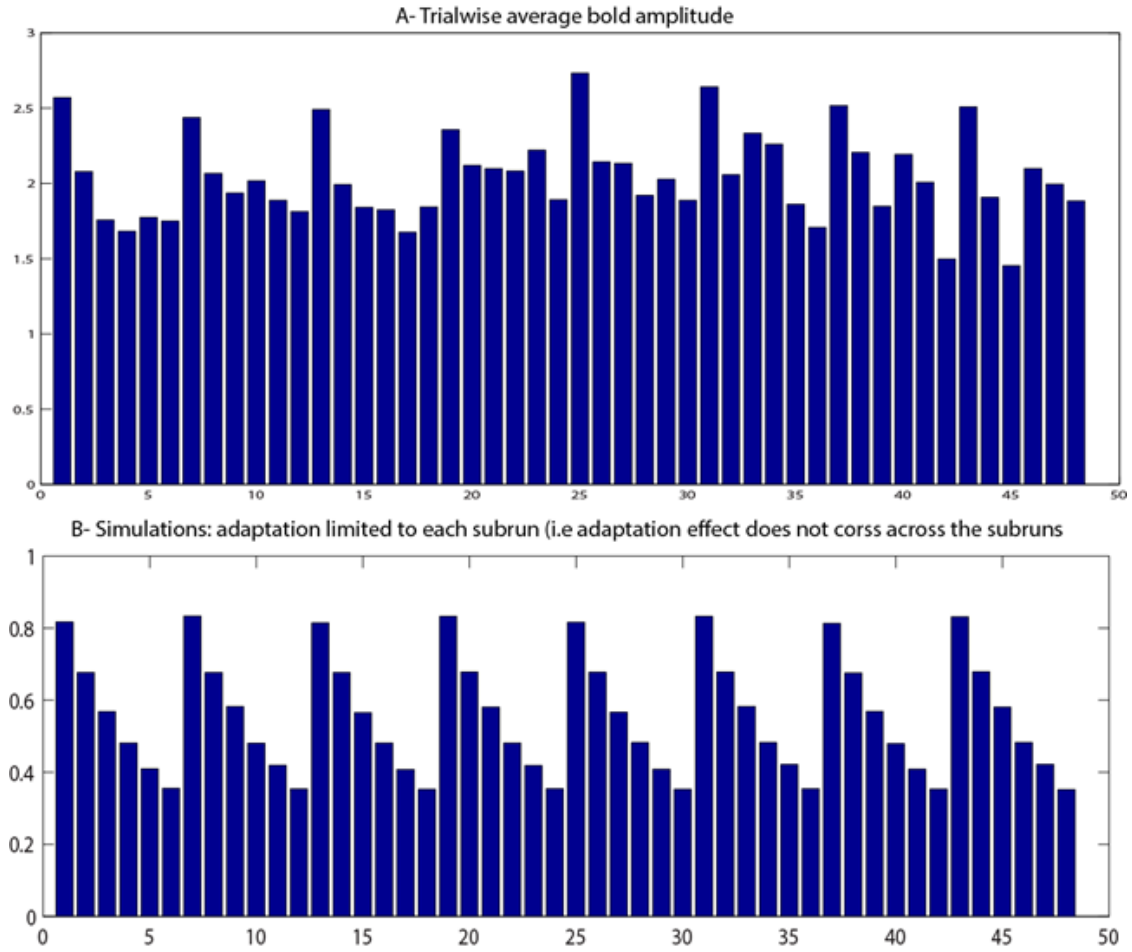
As explained in Chapter 2, LSS method becomes more beneficial in case of higher scan variability ratio to trial variability and this was attributed to its smoothing effect which neutralises the excessive Gaussian scan noise. I have tested to see if LSA combined with an explicitly smoothed BOLD signal using a low pass filter with a window of 5 scans has the same benefits as LSS. The answer is revealed in the supplementary figure below which compares LSS to BOLD smoothing + LSA. It can be noted that these two methods are not equivalent. More importantly, the later shifts the mean beta toward the baseline.



Sup. Figure3: Simulations comparing the standard LSS beta estimates to that of LSA on a smoothed BOLD in a randomised design that has two trial types. The trials have been re-ordered in the picture where trials from 1 to 100 have a true beta magnitude of 3 while trials from 101-200 have a true beta magnitude of -3. Simulated scan variability =0.3 and trial variability =0.1

Appendix 4

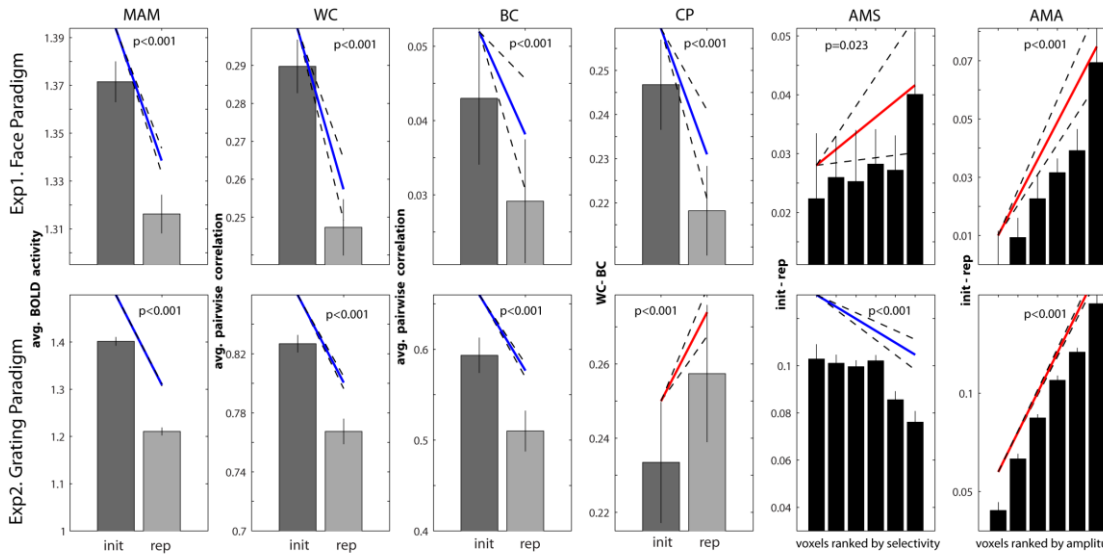
For the grating dataset, I did not reset the adaptation factor, *c*, to 1 before each initial presentation; rather, I set the factor to 1 only at the start of each subrun (see section 6.4). This is because the data suggest that the lengthy subrun breaks allowed the repetition effects to fade away, as shown in supplementary Figure 4 below:



Sup. Figure 4: Showing A) adaptation effects were limited to the subruns and the BOLD activity at the start of each subrun were the same as the first trials in each subrun B) showing the effect of re-setting the adaptation factor between the independent runs in our simulation to match the empirical data results.

Appendix 5:

Although our simulations matched the findings from both datasets qualitatively, there were some quantitative differences. Most of these differences can be trivially attributed to the unknown (and uninteresting) scaling between neural activity and fMRI BOLD signal. However, one more obvious disparity was that there was a positive correlation between stimulus classes (BC) in both datasets, yet the simulations produced a negative correlation. The positive BC in the data could owe to several factors that were not modelled in the simulations, such as correlated variance across trials or temporal drift (extrinsic scanner factors). Alternatively there could be intrinsic factors such as neural populations within a voxel that are not selective, responding equally to both stimulus classes (i.e flat tuning curves). Such diversity in the neural tuning curves has been reported in single-cell literature (Shapley et al., 2003; Schmolesky et al., 2000). These additional contributions or assumptions are not of theoretical interest for this thesis. Nonetheless, as a sanity check, I confirmed that a positive BC could be achieved by adding a proportion of neural populations with flat tuning curves that respond and adapt equally to both stimulus types (Supplementary Figure 4). Importantly, this addition did not change the overall conclusions, i.e. local scaling was still the wining model in both datasets even after adding this extra type of neurons to produce more positively correlated neural activity.



Sup Figure 5. Local scaling prediction for all the 6 criteria after adding correlated neural activity to make BC positive. Around 10% of correlated neural activities (neurons with flat tuning curves) were added to the voxels for the face paradigm, and around 50% added for the grating paradigm. (init = initial, rep = repeated).

Appendix 6

Supplementary Table 1: Wining parameter combinations in Experiment 1 & Experiment 2

grating dataset			**face dataset**		
t-stat 99% confidence			**t-stat 99% confidence**		
a	***b***	***σ***	***a***	***b***	***σ***
0.7	0.1	0.4	0.6	0.1	0.1
0.8	0.1	0.4	0.7	0.1	0.1
0.7	0.2	0.4	0.7	0.1	0.2
0.8	0.2	0.4	0.8	0.1	0.2

0.8	0.4	0.4	0.9	0.1	0.2
0.8	0.6	0.4	0.6	0.2	0.2
0.8	0.1	0.6	0.7	0.2	0.2
0.9	0.1	0.6	0.8	0.2	0.2
0.8	0.2	0.6	0.9	0.2	0.2
0.9	0.2	0.6	0.6	0.4	0.2
0.8	0.4	0.6	0.7	0.4	0.2
0.9	0.4	0.6	0.8	0.4	0.2
0.8	0.6	0.6	0.9	0.4	0.2
0.9	0.6	0.6	0.6	0.6	0.2
0.9	0.8	0.6	0.7	0.6	0.2
0.9	0.1	0.8	0.8	0.6	0.2
0.9	0.2	0.8	0.9	0.6	0.2
0.9	0.4	0.8	0.6	0.8	0.2
0.9	0.6	0.8	0.7	0.8	0.2
0.9	0.8	0.8	0.8	0.8	0.2
0.9	0.1	1	0.9	0.8	0.2
0.9	0.2	1	0.7	1	0.2
0.9	0.4	1	0.7	1.3	0.2
0.9	0.6	1	0.9	1.3	0.4
0.9	0.8	1	0.9	1.7	0.4

REFERENCES

Abbenhuis, M. A., Raaijmakers, W. G. M., Raaijmakers, J. G. W., & Van Woerden, G. J. M. (1990). Episodic memory in dementia of the Alzheimer type and in normal ageing: Similar impairment in automatic processing. *The Quarterly Journal of Experimental Psychology*, *42*(3), 569-583.

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, *275*(5297), 221-224.

Abdulrahman, H., & Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, *125*, 756-766.

Aguirre, G. K., Zarahn, E., & D'esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage*, *8*(4), 360-369.

Alnk, A., Walther, A., Krugliak, A., van den Bosch, J. J., & Kriegeskorte, N. (2015). Mind the drift-improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. *bioRxiv*, 032391.

Alink, A., Krugliak, A., Walther, A., & Kriegeskorte, N. (2013). fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Frontiers in psychology*, *4*.

Alink, A., Walther, A., Krugliak, A., & Kriegeskorte, N. (2017). Local opposite orientation preferences in V1: fMRI sensitivity to fine-grained pattern information. *Scientific Reports*, *7*(1), 7128.

Andresen, D. R., Vinberg, J., & Grill-Spector, K. (2009). The representation of object viewpoint in human visual cortex. *Neuroimage*, *45*(2), 522-536.

Arendt, T., Holzer, M., Stöbe, A., GÄrtner, U., LÜTH, H. J., Brückner, M. K., & Ueberham, U. (2000). Activated mitogenic signaling induces a process of dedifferentiation in Alzheimer's disease that eventually results in cell death. *Annals of the New York Academy of Sciences*, *920*(1), 249-255.

Avidan, G., Hasson, U., Hendler, T., Zohary, E., & Malach, R. (2002). Analysis of the neuronal selectivity underlying low fMRI signals. *Current Biology*, *12*(12), 964-972.

Bacciu, D., & Starita, A. (2008). Competitive repetition suppression (CoRe) clustering: A biologically inspired learning model with application to robust clustering. *IEEE transactions on neural networks*, *19*(11), 1922-1941.

Bachatene, L., Bharmauria, V., Cattan, S., Rouat, J., & Molotchnikoff, S. (2015). Reprogramming of orientation columns in visual cortex: a domino effect. *Scientific reports*, *5*.

Baum, S. H., & Beauchamp, M. S. (2014). Greater BOLD variability in older compared with younger adults during audiovisual speech perception. *PloS one*, *9*(10), e111121.

Baylis, G. C., & Rolls, E. T. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Experimental brain research*, *65*(3), 614-622.

Berry, C. J., Shanks, D. R., Speeken brink, M., & Henson, R. N. (2012). Models of recognition, repetition priming, and fluency: exploring a new framework. *Psychological review*, *119*(1), 40.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS biology*, *14*(11), e1002577.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, *12*(8), 3139-3161.

Brignell, C. J., Browne, W. J., Dryden, I. L., & Francis, S. T. (2015). Mixed Effect Modelling of Single Trial Variability in Ultra-High Field fMRI. *arXiv preprint arXiv:1501.05763*.

Brown, M. W., Wilson, F. A. W., & Riches, I. P. (1987). Neuronal evidence that inferomedial temporal cortex is more important than hippocampus in certain processes underlying recognition memory. *Brain research*, *409*(1), 158-162.

Buckner, R. L., Goodman, J., Burock, M., Rotte, M., Koutstaal, W., Schacter, D. & Dale, A. M. (1998). Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI. *Neuron*, *20*(2), 285-296.

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223.

Carp, J., Gmeindl, L., & Reuter-Lorenz, P. A. (2010). Age differences in the neural representation of working memory revealed by multi-voxel pattern analysis. *Frontiers in Human Neuroscience*, *4*.

Carr, V. A., Rissman, J., & Wagner, A. D. (2010). Imaging the human medial temporal lobe with high-resolution fMRI. *Neuron*, *65*(3), 298-308.

Chawla, D., Rees, G., & Friston, K. J. (1999). The physiological basis of attentional modulation in extrastriate visual areas. *Nature neuroscience*, *2*(7), 671.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human brain mapping*, *8*(2-3), 109-114.

Davis, H. P., Cohen, A., Gandy, M., Colombo, P., VanDusseldorp, G., Simolke, N., & Romano, J. (1990). Lexical priming deficits as a function of age. *Behavioral Neuroscience*, *104*(2), 288.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *Neuroimage*, *97*, 271-283.

de Beeck, H. P. O. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage*, *49*(3), 1943-1948.

De Gardelle, V., Waszczuk, M., Egner, T., & Summerfield, C. (2012). Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. *Cerebral Cortex*, *23*(9), 2235-2244.

De Gardelle, V., Stokes, M., Johnen, V. M., Wyart, V., & Summerfield, C. (2013). Overlapping multivoxel patterns for two levels of visual expectation. *Frontiers in human neuroscience*, *7*.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences*, *93*(24), 13494-13499.

Dragoi, V., Rivadulla, C., & Sur, M. (2001). Foci of orientation plasticity in visual cortex. *Nature*, *411*(6833), 80-86.

Dragoi, V., Sharma, J., & Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, *28*(1), 287-298.

Engell, A. D., & McCarthy, G. (2014). Repetition suppression of face-selective evoked and induced EEG recorded from human cortex. *Human brain mapping*, *35*(8), 4155-4162.

Ewbank, M. P., & Henson, R. N. (2012). Explaining away repetition effects via predictive coding. *Cognitive neuroscience*, *3*(3-4), 239-240.

Fahy, F. L., Riches, I. P., & Brown, M. W. (1993). Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Experimental Brain Research*, *96*(3), 457-472.

Fair, D. A., Schlaggar, B. L., Cohen, A. L., Miezin, F. M., Dosenbach, N. U., Wenger, K. K., & Petersen, S. E. (2007). A method for using blocked and event-

related fMRI data to study “resting state” functional connectivity. *Neuroimage*, *35*(1), 396-405.

Fang, F., Murray, S. O., & He, S. (2006). Duration-dependent fMRI adaptation and distributed viewer-centered face representation in human visual cortex. *Cerebral Cortex*, *17*(6), 1402-1411.

Fox, M. D., Snyder, A. Z., Zacks, J. M., & Raichle, M. E. (2006). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature neuroscience*, *9*(1), 23-25.

Fransson, P. (2005). Spontaneous low-frequency BOLD signal fluctuations: An fMRI investigation of the resting-state default mode of brain function hypothesis. *Human brain mapping*, *26*(1), 15-29.

Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. *Neuroimage*, *16*(2), 484-512.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1211-1221.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A. N. D. R. E. W., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: characterizing differential responses. *Neuroimage*, *7*(1), 30-40.

Friston, K. J., Zarahn, E. O. R. N. A., Josephs, O., Henson, R. N. A., & Dale, A. M. (1999). Stochastic designs in event-related fMRI. *Neuroimage*, *10*(5), 607-619.

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2010). Blood oxygen level-dependent signal variability is more than just noise. *Journal of Neuroscience*, *30*(14), 4914-4921.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., Stephan, K. E., Baldeweg, T., & Friston, K. J. (2009). Repetition suppression and plasticity in the human brain. *Neuroimage*, *48*(1), 269-279.

George, N., Dolan, R. J., Fink, G. R., Baylis, G. C., Russell, C., & Driver, J. (1999). Contrast polarity and face recognition in the human fusiform gyrus. *Nature neuroscience*, *2*(6), 574-580.

Goense, J. B., & Logothetis, N. K. (2008). Neurophysiology of the BOLD fMRI signal in awake monkeys. *Current Biology*, *18*(9), 631-640.

Grady, C. L., & Garrett, D. D. (2014). Understanding variability in the BOLD signal and why it matters for aging. *Brain imaging and behavior*, *8*(2), 274-283.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta psychologica*, *107*(1), 293-321.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, *10*(1), 14-23.

Grotheer, M., & Kovács, G. (2015). The relationship between stimulus repetitions and fulfilled expectations. *Neuropsychologia*, *67*, 175-182.

Grotheer, M., & Kovács, G. (2016). Can predictive coding explain repetition suppression?. *Cortex*, *80*, 113-124.

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, *21*(4), 1639-1651.

Hartman, M., & Hasher, L. (1991). Aging and suppression: Memory for previously relevant information. *Psychology and aging*, *6*(4), 587.

Hatfield, M., McCloskey, M., & Park, S. (2016). Neural representation of object orientation: A dissociation between MVPA and Repetition Suppression. *Neuroimage*, *139*, 136-148.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-2430.

Heeger, D. J., Huk, A. C., Geisler, W. S., & Albrecht, D. G. (2000). Spikes versus BOLD: what does neuroimaging tell us about neuronal activity?. *Nature neuroscience*, *3*(7), 631-633.

Henriksson, L., Nurminen, L., Hyvärinen, A., & Vanni, S. (2008). Spatial frequency tuning in human retinotopic visual areas. *Journal of Vision*, *8*(10), 5-5.

Henson, R. N. (2016). Repetition suppression to faces in the fusiform face area: A personal and dynamic journey. *cortex*, *80*, 174-184.

Henson, R. N. (2012). Repetition accelerates neural dynamics: In defense of facilitation models. *Cognitive neuroscience*, *3*(3-4), 240-241.

Henson, R. N. A., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, *41*(3), 263-270.

Henson, R. N. A., Shallice, T., Gorno-Tempini, M. L., & Dolan, R. J. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex*, *12*(2), 178-186.

Henson, R. N., & Mouchlianitis, E. (2007). Effect of spatial attention on stimulus-specific haemodynamic repetition effects. *Neuroimage*, *35*(3), 1317-1329.

Henson, R., Shallice, T., & Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science*, *287*(5456), 1269-1272.

Henson, R. N. (2012). Repetition accelerates neural dynamics: In defense of facilitation models. *Cognitive neuroscience*, *3*(3-4), 240-241.

Henson, R. N. A. (2003). Neuroimaging studies of priming. *Progress in neurobiology*, *70*(1), 53-81.

Horner, A. J., & Henson, R. N. (2008). Priming, response learning and repetition suppression. *Neuropsychologia*, *46*(7), 1979-1991.

James, T. W., Humphrey, G. K., Gati, J. S., Menon, R. S., & Goodale, M. A. (1999). Repetition priming and the time course of object recognition: an fMRI study. *Neuroreport*, *10*(5), 1019-1023.

Jeyabalaratnam, J., Bharmauria, V., Bachatene, L., Cattan, S., Angers, A., & Molotchnikoff, S. (2013). Adaptation shifts preferred orientation of tuning curve in the mouse visual cortex. *PloS one*, *8*(5), e64294.

Josephs, O., & Henson, R. N. (1999). Event-related functional magnetic resonance imaging: modelling, inference and optimization. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *354*(1387), 1215-1228.

Kamitani, Y., & Sawahata, Y. (2010). Spatial smoothing hurts localization but not

information: pitfalls for brain mappers. *Neuroimage*, *49*(3), 1949-1952.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, *8*(5), 679-685.

Kar, K., & Krekelberg, B. (2016). Testing the assumptions underlying fMRI adaptation using intracortical recordings in area MT. *Cortex*, *80*, 21-34.

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265-270.

Kovács, G., & Vogels, R. (2014). When does repetition suppression depend on repetition probability?. *Frontiers in human neuroscience*, *8*.

Krekelberg, B., Boynton, G. M., & van Wezel, R. J. (2006). Adaptation: from single cells to BOLD signals. *Trends in neurosciences*, *29*(5), 250-256.

Kriegeskorte, N., Cusack, R., & Bandettini, P. (2010). How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter?. *Neuroimage*, *49*(3), 1965-1976.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National academy of Sciences of the United States of America*, *103*(10), 3863-3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National academy of Sciences of the United States of America*, *103*(10), 3863-3868.

Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of neurophysiology*, *69*(6), 1918-1929.

Linke, A. C., Vicente-Grabovetsky, A., & Cusack, R. (2011). Stimulus-specific suppression preserves information in auditory short-term memory. *Proceedings of the National Academy of Sciences*, *108*(31), 12961-12966.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869-878.

Lövdén, M., Li, S. C., Shing, Y. L., & Lindenberger, U. (2007). Within-person trial-to-trial variability precedes and predicts cognitive decline in old and very old age: Longitudinal data from the Berlin Aging Study. *Neuropsychologia*, *45*(12), 2827-2838.

Lueschow, A., Miller, E. K., & Desimone, R. (1994). Inferior temporal mechanisms for invariant object recognition. *Cerebral Cortex*, *4*(5), 523-531.

Lu, Y., Jiang, T., & Zang, Y. (2005). Single-trial variable model for event-related fMRI data analysis. *IEEE transactions on medical imaging*, *24*(2), 236-245.

Martin, A., Lalonde, F. M., Wiggs, C. L., Weisberg, J., Ungerleider, L. G., & Haxby, J. V. (1995). Repeated presentation of objects reduces activity in ventral occipitotemporal cortex: A fMRI study of repetition priming. In *Society for Neuroscience Abstracts* (Vol. 21, p. 1497).

McDonald, C. R., Thesen, T., Carlson, C., Blumberg, M., Girard, H. M., Trongnetrpunya, A., ... & Cash, S. S. (2010). Multimodal imaging of repetition priming: using fMRI, MEG, and intracranial EEG to reveal spatiotemporal profiles of word processing. *Neuroimage*, *53*(2), 707-717.

McMahon, D. B., & Olson, C. R. (2007). Repetition suppression in monkey inferotemporal cortex: relation to behavioral priming. *Journal of neurophysiology*, *97*(5), 3532-3543.
Miller, E. K., & Desimone, R. (1993). Scopolamine affects short-term memory but not inferior temporal neurons. *Neuroreport*, *4*(1), 81.

Miller, E. K., Gochin, P. M., & Gross, C. G. (1991). Habituation-like decrease in the responses of neurons in inferior temporal cortex of the macaque. *Visual neuroscience*, *7*(4), 357-362.

Mitchell, D. B., & Schmitt, F. A. (2006). Short-and long-term implicit memory in aging and Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, *13*(3-4), 611-635.

Moore, K. S., Yi, D. J., & Chun, M. (2013). The effect of attention on repetition suppression and multivoxel pattern similarity. *Journal of cognitive neuroscience*, *25*(8), 1305-1314.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636-2643.

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage*, *103*, 130-138.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social cognitive and affective neuroscience*, *4*(1), 101-109.

Myung, J. I., Tang, Y., & Pitt, M. A. (2009). Evaluation and comparison of computational models. *Methods in enzymology*, *454*, 287-304.

Naccache, L., & Dehaene, S. (2001). The priming method: imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral cortex*, *11*(10), 966-974.

Neumann, J., Lohmann, G., Zysset, S., & von Cramon, D. Y. (2003). Within-subject variability of BOLD response dynamics. *Neuroimage*, *19*(3), 784-796.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, *10*(9), 424-430.

Pourtois, G., Schwartz, S., Spiridon, M., Martuzzi, R., & Vuilleumier, P. (2008). Object representations for multiple visual categories overlap in lateral occipital and medial fusiform cortex. *Cerebral Cortex*, *19*(8), 1806-1819.

Rao, R. P. (1999). An optimal estimation approach to visual perception and learning. *Vision research*, *39*(11), 1963-1989.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79-87.

Rees, G., Friston, K., & Koch, C. (2000). A direct quantitative relationship between the functional properties of human and macaque V5. *Nature neuroscience*, *3*(7), 716-723.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual review of psychology*, *39*(1), 475-543.

Ringach, D. L., Bredfeldt, C. E., Shapley, R. M., & Hawken, M. J. (2002). Suppression of neural responses to nonoptimal stimuli correlates with tuning selectivity in macaque V1. *Journal of Neurophysiology*, *87*(2), 1018-1027.

Rissman, J., Gazzaley, A., & D'esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*(2), 752-763.

Rissman, J., Greely, H. T., & Wagner, A. D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences*, *107*(21), 9849-9854.

Rossion, B. (2008). Constraining the cortical face network by neuroimaging studies of acquired prosopagnosia. *Neuroimage*, *40*(2), 423-426.

Sapountzis, P., Schluppeck, D., Bowtell, R., & Peirce, J. W. (2010). A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. *Neuroimage*, *49*(2), 1632-1640.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B*, *372*(1711), 20160049.

Schmolesky, M. T., Wang, Y., Pu, M., & Leventhal, A. G. (2000). Degradation of stimulus selectivity of visual cortical cells in senescent rhesus monkeys. *Nature neuroscience*, *3*(4), 384-390.

Shapley, R., Hawken, M., & Ringach, D. L. (2003). Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. *Neuron*, *38*(5), 689-699.

Spigler, G., & Wilson, S. P. (2017). Familiarization: A theory of repetition suppression predicts interference between overlapping cortical representations. *PLoS One*, *12*(6), e0179306.

Stern, C. E., Corkin, S., González, R. G., Guimaraes, A. R., Baker, J. R., Jennings, P. J., ... & Rosen, B. R. (1996). The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, *93*(16), 8660-8665.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745-756.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature neuroscience*, *11*(9), 1004-1006.

Summerfield, C., Wyart, V., Johnen, V. M., & De Gardelle, V. (2011). Human scalp electroencephalography reveals that repetition suppression varies with expectation. *Frontiers in Human Neuroscience*, *5*.

Swindale, N. V. (1998). Orientation tuning curves: empirical description and estimation of parameters. *Biological cybernetics*, *78*(1), 45-56.

Talairach, J., & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging.

Todorovic, A., & de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *Journal of Neuroscience*, *32*(39), 13389-13395.

Tsvetanov, K. A., Henson, R. N., Tyler, L. K., Davis, S. W., Shafto, M. A., Taylor, J. R., ... & Rowe, J. B. (2015). The effect of ageing on fMRI: correction for the confounding effects of vascular reactivity evaluated by joint fMRI and MEG in 335 adults. *Human brain mapping*, *36*(6), 2248-2269.

Turner, B. (2010). A comparison of methods for the use of pattern classification on rapid event-related fMRI data. In *Poster session presented at the Annual Meeting of the Society for Neuroscience, San Diego, CA*.

Utzerath, C., John-Saaltink, E., Buitelaar, J., & Lange, F. P. (2017). Repetition suppression to objects is modulated by stimulus-specific expectations. *Scientific Reports*, *7*(1), 8781.

Van Hooser, S. D., Heimel, J. A. F., Chung, S., Nelson, S. B., & Toth, L. J. (2005). Orientation selectivity without orientation maps in visual cortex of a highly visual mammal. *Journal of Neuroscience*, *25*(1), 19-28.

Vidal, J. R., Perrone-Bertolotti, M., Levy, J., De Palma, L., Minotti, L., Kahane, P., ... & Lachaux, J. P. (2014). Neural repetition suppression in ventral occipito-

temporal cortex occurs during conscious and unconscious processing of frequent stimuli. *Neuroimage*, *95*, 129-135.

Visser, R. M., de Haan, M. I., Beemsterboer, T., Haver, P., Kindt, M., & Scholte, H. S. (2016). Quantifying learning-dependent changes in the brain: Single-trial multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology*, *53*(8), 1117-1127.

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. *Journal of Neuroscience*, *31*(33), 12021-12028.

Vogels, R., Sáry, G., & Orban, G. A. (1995). How task-related are the responses of inferior temporal neurons?. *Visual neuroscience*, *12*(2), 207-214.

Wakeman, D. G., & Henson, R. N. (2015). A multi-subject, multi-modal human neuroimaging dataset. *Scientific data*, *2*, sdata20151.

Ward, E. J., Chun, M. M., & Kuhl, B. A. (2013). Repetition suppression and multi-voxel pattern similarity differentially track implicit and explicit visual memory. *Journal of Neuroscience*, *33*(37), 14749-14757.

Weiner, K. S., Sayres, R., Vinberg, J., & Grill-Spector, K. (2010). fMRI-adaptation and category selectivity in human ventral temporal cortex: regional differences across time scales. *Journal of neurophysiology*, *103*(6), 3349-3365.

Whittington, J. C., & Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current opinion in neurobiology*, *8*(2), 227-233.

Wilson, R. S., Segawa, E., Hizel, L. P., Boyle, P. A., & Bennett, D. A. (2012). Terminal dedifferentiation of cognitive abilities. *Neurology*, *78*(15), 1116-1122.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, *330*(6000), 97-101.

Zarahn, E., Aguirre, G. K., & D'Esposito, M. (1997). Empirical analyses of BOLD fMRI statistics. *Neuroimage*, *5*(3), 179-197.

Zeithamova, D., de Araujo Sanchez, M. A., & Adke, A. (2017). Trial timing and pattern-information analyses of fMRI data. *NeuroImage*, *153*, 221-231.

Zhang, H., Tian, J., Liu, J., Li, J., & Lee, K. (2009). Intrinsically organized network for face perception during the resting state. *Neuroscience letters*, *454*(1), 1-5.

End of thesis

Hunar Abdulrahman

2018