DECODING NEURAL MECHANISMS OF SURROUND SUPPRESSION IN FEATURE-BASED ATTENTION

By

Wanghaoming Fang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology – Doctor of Philosophy

2021

ABSTRACT

DECODING NEURAL MECHANISMS OF SURROUND SUPPRESSION IN FEATURE-BASED ATTENTION

By

Wanghaoming Fang

Feature-based attention (FBA) selectively enhances processing of an attended feature at the expense of unattended or task-irrelevant features. Recent studies showed that FBA modulates the perceptual space with both a monotonic profile (i.e., feature-similarity gain) and a non-monotonic profile (i.e., surround suppression). A significant question arises regarding the neural mechanism of the non-monotonic surround suppression effect. Previous studies have suggested that two candidate neuronal mechanisms could underlie these attentional modulations: a shift of neuronal tuning preference toward the attended feature, or a multiplicative gain modulation that scales the overall responses without changing their tuning property. Yet the empirical evidence for these mechanisms is still lacking. In the current work, we explored how these neuronal mechanism manifest at the level of fMRI BOLD measurement using a simulation approach. Specifically, we employed an encoding/decoding approach by first simulating voxel responses from neuronal population assuming either mechanism and then applying a regression-based inverted encoding model (IEM) and a Bayesian method to decode population representations. We found that both methods captured the signature patterns associated with these different neuronal mechanisms. In our second aim, we systematically varied the correlation structure of voxel noise to further compare these different multivariate methods in a stimulus classification task. Our results showed a clear advantage of the Bayesian method over IEM, suggesting that the Bayesian method was superior for deciphering neural representation given the prevalent noise correlation and variable tuning width in the brain. In sum, our current simulation work may

provide a proof of concept for future empirical studies investigating cortical mechanism of FBA using non-invasive methods, as well as guidance for choosing suitable methods in these investigations.

ACKNOWLEDGEMENTS

My life in graduate school has been a truly amazing and extraordinary journey in the past few years. I am also grateful to the wonderful people, who motivated and inspired me throughout the journey to become a better person. I am lucky enough to have such a great family, whose love to me is incomparable. Therefore, I want to first thank my father, Tianci Fang, my mother, Xiaoqing Wang, and my wife, Wenjing Li, who always believe in me and offer incredible support especially amidst the pandemic.

I also have had the pleasure of working with the intelligent minds, especially my doctoral/master guidance committee and collaborators: Dr. Taosheng Liu, Dr. Susan Ravizza, Dr. Jan Brascamp, Dr. Karl Healey, and Dr. Mark Becker. Thank you for your guidance on how to become a rigorous, creative, and collaborative researcher. Moreover, I want to give my special thanks to my advisor, Dr. Taosheng Liu, for sharing your knowledge and perspectives and most importantly, teaching me how to think critically. It has been a great honor to be your student. I am also incredibly grateful to the previous collaborators from New York University, Dr. Marisa Carrasco and Dr. Alex White, who inspired me to become a researcher. Lastly, this work was supported in part by a grant from NSF (2019995).

Thank you all for the great inspirations and being part of my journey. I would not have made it this far without your support and encouragement. Wherever I am in the future, I will always cherish the experiences and wish you all the best.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Our resource-limited visual system is constantly challenged by the information-rich visual environment. To overcome the limitation, visual selective attention filters out task-irrelevant competing distractors and select only a small proportion of task-relevant information for prioritized processing. Such attentional selection can be based on location (i.e., 'spatial attention', Carrasco, 2011) and/or non-spatial features (i.e., 'feature-based attention', Carrasco, 2011; Liu, 2019). The small subset of information that is selected by attention can enjoy benefit such as enhanced behavioral performances and neural responses for the attended location/feature. However, the fate of the unattended location/features remains less clear. If the attentional filter is perfect, one would predict a uniform exclusion of all task-irrelevant information except only the target information (e.g., a step function in a perceptual space). However, empirical studies have shown that it is seldomly the case, especially at the behavioral level (e.g., rarely chance-level performance for the unattended locations/features). Therefore, an important question is how selection of a location or a feature modulates the representation of other locations and features that falls outside of the attentional focus.

Since the beginning of psychological research, how attention modulates our perception has attracted quite some interests of many pioneer psychologists, like Helmholtz, Fechner, and James. Recently studies have proposed different models to describe the profile of visual selective attention. In the spatial domain, early studies characterized the shape of the focus of attention with a popular metaphor – spotlight (Posner, 1980), which well captured the notion that attention selects the most relevant location for enhanced processing at the expense of unattended locations. Later, researchers proposed a gradient structure of the attentional "spotlight" (LaBerge, 1983). It

has been shown that performance monotonically fall off with the distance between the attended and unattended location – a spatial gradient of attentional modulation. Although spatial attention has dominated the studies for decades, location is the not the only stimulus property that we can attend to. The ability to allocate attention to the non-spatial feature(s) of a stimulus is called feature-based attention (FBA). Note, feature is specifically defined in the current work as values within a dimension (e.g., red, or green), although some researchers also use this term to describe a whole feature-dimension regardless of specific values (e.g., dimension-based attention, Found & Muller, 1996; Muller, Heller & Ziegler, 1995).

Feature-based attention can facilitate target selection, even without knowing the exact location. For instance, knowing what color is worn can be helpful when searching for a friend in the Spartan Stadium. This example also illustrates one of the most fundamental properties of feature-based attention – selection of an attended feature spreads globally across the visual field. The global spread of FBA also forms the basis for the highly popular paradigm – visual search, in which participants typically search for a pre-defined target in an array of stimuli. However, in the antecedent case, FBA would not be always helpful especially if your friend wears the same green Spartan T-shirt with the crowd. Therefore, it seems likely that modulation of feature-based attention may also depends on target-distractor similarity.

Some researchers investigate the profile of FBA and suggested that it might also have a monotonic gradient. Based on the visual search paradigm, Duncan & Humphreys (1989) conducted one of the earliest studies on the profile of FBA, in which they found that search efficiency monotonically increased as a target became more different from the distractors, suggesting a monotonic profile. Converging evidence for a monotonic gradient of feature-based attention (FBA) also comes from neural recording studies. For example, early single-unit studies

on FBA have also proposed a monotonic profile in feature space (Fig. 1a), as epitomized by the feature-similarity gain model (Martinez-Trujillo & Treue, 2004; Treue and Martinez-Trujillo, 1999). According to the feature-similarity gain model, the attentional modulation of neuronal activity is a monotonic function of feature similarity. Specifically, attentional enhancement gradually decreases and turns into suppression for features that are progressively more dissimilar to the attended feature. Although the feature-similarity gain model was originally proposed to account for attentional modulation at single-unit level (Martinez-Trujillo & Treue, 2004; Treue & Martinez-Trujillo, 1999), human behavioral and neural imaging studies have also obtained results that generally supports this monotonic profile (Saenz, Buracas & Boynton, 2002, 2003; Liu, Larsson & Carrasco, 2007; Ling, Liu & Carrasco, 2008; Zhang & Luck, 2009; Wang, Miller & Liu, 2015; Ho, et al., 2012; Paltoglou & Neri, 2012).

However, there are several challenges for visual system that are difficult to resolve with only a monotonic selection profile. First, the gradient model predicts a reduction of interference only when the distractor is sufficiently far away (in physical or feature space) from the attended target. But the environment is highly variable that visual information rarely consists of only highly distinct features (e.g., red vs. green). Thus, it is unclear whether and how FBA facilitates selection of a target among similar but different distractors (e.g., finding a beige coffee mug among papers on the desk). Second, while previous studies seem to suggest a linking hypothesis between human behavior and the single-unit findings (i.e., the feature-similarity gain modulation in MT neurons), some methodological concerns, like coarse sampling in feature-space, suggest that the simple gradient model of attention may not be the full story. Third, and most critically, an increasing number of studies recently showed evidence that there exists a non-monotonic profile of attentional modulation in both spatial and feature domain, which cannot be readily

accommodated by the attentional gradient. For example, recent studies of spatial attention that sampled locations more finely have revealed a non-monotonic profile of attention comprised of "surround suppression", such that nearby locations are more suppressed than further locations. This local suppression is thought to allow better distinction between closely located targets and distractors (Hopf et al., 2006; Boehler et al., 2009, 2011; Muller & Kleinschmidt, 2004; Mounts, 2000a, 2000b; Tsotsos, 1995, 2011). In a similar vein, other researchers also reported a non-monotonic attentional modulation when features near the attended feature were probed (Fang, Becker & Liu, 2019; Fang & Liu, 2019; Stormer & Alvarez, 2014; Tombu & Tsotsos, 2008). Consistent with its spatial equivalent, the "surround suppression" effect in FBA enhances signal-to-noise ratio when the target and distractors have similar features (Fig. 1b). Moreover, the two recent studies have further shown that FBA consists of a hybrid profile of both FSG and surround suppression but operates at different similarity scale (Fig. 1c, Fang, Becker & Liu, 2019; Fang & Liu, 2019). Taken together, converging evidence now suggests that there exists a non-monotonic attentional modulation in the vicinity of the attended location/feature.

In the following sections, I first discuss previous evidence for the feature-similarity gain model and the need for a more systematic examination of FBA's profile. Next, I discuss the evidence supporting surround suppression as a canonical mechanism underlying attentional modulation for a number of feature spaces (e.g., color, orientation, motion direction, and spatial frequency) and a more flexible hybrid profile consisting of both surround suppression and feature-similarity gain modulation to adaptively enhance signal-to-noise ratio in isolating a target feature. Lastly, I discuss candidate neural mechanisms of the surround suppression in FBA based on previous findings in both spatial and feature-based attention, arguing that these mechanisms

could be potentially distinguished with recent development of multivariate techniques through computational simulation and modeling.

The Feature-similarity Gain Model

Since the seminal study on attention by Treisman and colleagues (Treisman and Gelade, 1980), human's ability to search for a target based on its defining feature has attracted enormous attention from researchers. While one of the key research topics in early attention studies was the debate on pre-attentive and attentive processing, recent findings have argued against such a rigid dichotomy of this two-stage framework. Even the "pre-attentive" stage that was originally thought to be parallel and capacity-free is subject to FBA's modulation, which was supported by neurophysiological studies that recorded directly from the neurons tuned to different features (Bichot et al., 2005).

Initial Evidence for FSG

Despite the prominence of the classic visual search paradigm, searching for a target also means finding its location which unavoidably involves a shift of spatial attention. This employment of both spatial and feature-based attention does not allow an isolation of a pure feature-based attentional modulation, which has led to claims that the role of FBA is only limited to guiding spatial attention to the target without directly modulating perception (Moore & Egeth, 1998; Shih & Sperling, 1996). Thus, to isolate a pure effect of feature-based attention, a new paradigm was developed utilizing the global spread of FBA. For example, researchers typically focus participants' spatial attention to one location and probe attentional modulation at a different (i.e., spatially ignored) site. Using this method, Martinez-Trujillo and Treue (2004) conducted experiments to directly measure FBA's modulation at the neuronal level. They presented a dot motion stimulus to MT neurons that have receptive fields (RF) in one hemifield,

while the monkey subject attended to the identical stimulus in the opposite hemifield. Because of the global spread of feature-based attention, the neurons in the spatially unattended side were also modulated by attention to motion direction. The authors found that FBA sharpened the population response to an attended motion direction by suppressing neurons preferring the most dissimilar motion direction. But most importantly, the gain factor (i.e., the multiplicative ratio that was applied to change neuronal response by FBA) was found to be a linear function of the similarity between tuning preferences and the attended direction such that attending to a neuron's preferred direction enhanced its response and attending to its non-preferred directions led to suppression of its response. Such findings eventually led to the proposal of the influential *feature-similarity gain model* (Martinez-Trujillo & Treue, 2004; Treue & Martinez-Trujillo, 1999), predicting a monotonic profile of FBA's modulation (Fig. 1a). Later studies further supported the feature-similarity gain model in multiple feature dimensions, based on psychophysical (Saenz, Buracas & Boynton, 2003; Ling, Liu & Carrasco, 2008; Wang, Miller & Liu, 2015; Ho, et al., 2012; Paltoglou & Neri, 2012), neuroimaging (Saenz, Buracas & Boynton, 2002; Liu, Larsson & Carrasco, 2007), electrophysiological (Zhang & Luck, 2009), and single-unit methods (reviewed by Maunsell & Treue, 2006).

Despite the popularity of the feature-similarity gain model, there are a few concerns on the exact interpretation of Martinez-Trujillo and Treue's original findings (Martinez-Trujillo & Treue, 2004). The main point is on their original design, in which the probe's motion direction in the ignored visual hemifield was the same as the direction in hemifield that monkeys attended to. In other words, when the stimulus's motion direction was systematically varied relative to a neuron's tuning preference, the researchers simultaneously changed both the attended feature and the feature in the neuron's RF. This raises two potential issues in using this design to support the

feature-similarity gain model. First, there is a concern regarding the interpretation of "gain". This word originated from earlier spatial attention studies, which described the multiplicative modulation of spatial attention that does not change the feature tuning profile of single neurons (e.g., orientation tuning in McAdam & Maunsell, 1999). However, the covariation of the attended feature and the probe feature in Martinez-Trujillo & Treue's experiments does not allow a full characterization of the neuronal tuning curve. Therefore, their original results may also be explained by other neuronal mechanisms that modifies feature selectivity of neurons (e.g., shift of tuning preference in David, et al., 2008). That being said, a better design would require measuring a full tuning curve each time when the monkey attends to a different motion direction, which is practically difficult to do. Second, Martinez-Trujillo and Treue's original finding describes a sharpening in population response to the attended motion direction. Therefore, the original neuronal evidence for feature-similarity gain model should be limited to only inferring how FBA selects the attended feature, but not about how FBA modulates perceptual representation of the rest of the feature continuum.

Other Evidence for the Feature-similarity Gain Model

Human psychophysical and neural studies provided important complementary evidence for the feature-similarity gain model in the broader context, as these methods rely more on the responses of a population of neurons, or the entire visual system. For neural studies, researchers typically employed a split-display design that is similar to the previous one used by Martinez-Trujillo and Treue – FBA's modulation was measured in the opposite hemifield to where spatial attention was deployed. For example, Saenz and colleagues, measured visual areas' activities when FBA was deployed to two feature dimensions – motion direction and color (Saenz, Buracas & Boynton, 2002). In the motion experiment, participants attended to one hemifield that

contains two overlapping fields of moving dots with opposite motion directions (e.g., upward vs. downward motion) and performed a speed discrimination task. Importantly, another single dot field (e.g., upward motion) was simultaneously presented in the ignored hemifield to provide neural measurement of attentional modulation. The results showed stronger responses across visual areas when the motion direction in the ignored dot field matched the attended direction than when it matched the unattended direction. Similar results were also obtained in color-based attention. Thus, such findings provided early support for the feature-similarity gain model in human visual attention.

Further neural evidence supporting feature-similarity gain model employed a variety of paradigms and tests in other feature dimensions. For example, Liu, Larsson and Carrasco (2007) used adaptation to test orientation-selective modulation in visual cortex at an attended location, where both attended and unattended orientation were superimposed (Liu, Larsson & Carrasco, 2007). The authors found that FBA selectively modulated the adaptation effect both psycho-physically (i.e., measured as behavioral tilt aftereffect) and physiologically (i.e., measured as fMRI response adaptation) for the attended but not the unattended orientation even when both features were spatially superimposed. In a later study, Zhang & Luck obtained similar findings by recording event-related potentials (ERP) to color dots stimuli (Zhang & Luck 2009). Critically, color-based attention resolved competition between two superimposed color dot fields (e.g., red vs. green) by selectively enhancing feed-forward processing of an attended color (e.g. red) over an unattended color (e.g., green) as reflected in the P1 ERP wave. In another study, Serences and Boynton (2007) also tested the feature-selective modulation with superimposed orientation stimuli. They found that the decodability (using multivoxel pattern analysis, MVPA) for an attended orientation was higher than the unattended orientation in the same hemifield,

which is, surprisingly, also true at the mirrored location in the absence of direct sensory stimulation (i.e., blank location). Taken together, neurophysiological studies provided further evidence that the global feature-selective modulation is consistent with the feature-similarity gain model in the human brain (Saenz et al., 2002; Liu, Larsson & Carrasco, 2007; Serences & Boynton, 2007; Zhang & Luck, 2009).

Psychophysical studies that measured the quality of attended feature also provided converging evidence for feature-similarity model. In psychophysics, the perceptual quality of unattended feature was typically evaluated by accuracy (i.e., proportion of correct responses) using a partially valid pre-cue. While neurophysiological studies typically presented a probe stimulus at a spatially ignored location to provide neural measurement of attentional modulation, behavioral studies control spatial attention by presenting all stimuli at the same location (e.g., center of screen), which further reduced the potential role of spatial attention. In an early study, Ling, Liu and Carrasco (2009) investigated how FBA modulates performance in a motion discrimination task when the attended motion direction was embedded in different level of noise (Ling, Liu & Carrasco, 2009). Critically, the author found enhancement from FBA even when the noise of motion was high - a behavioral effect that was consistent with the feature-similarity gain model's prediction in sharpening of population response. However, Ling et al. (2009) only manipulated the motion noise but did not probe performance for unattended directions of motion. Hence their study did not provide direct measurement of the profile of feature-based attention, e.g., how attentional modulation varies as a function of the feature-similarity between attended and unattended features.

Recent studies characterized a more complete functional profile of feature-based attention by systematically varying a target's feature from the attended one. For example, Ho et al. (2012)

measured the perceptual consequences of feature-based attention to motion direction. They employed a partially valid direction pre-cue to manipulate feature-based attention and, critically, measured the FBA's profile by systematically sampling the target's motion direction away from the attended direction in the invalid condition. Although the results showed a non-monotonic profile with the worst performance at 90° instead of the maximum 180°, such a pattern may be explained by axis-tuned motion mechanisms, which would respond equally well to opposite moving directions (Albright, 1984; Conway & Livingstone, 2003; Livingstone & Conway, 2003). Thus, their results were still interpreted as consistent with feature-similarity gain model. However, their finding may be due to a combined effect of spatial attention and feature-based attention as the task was to search for the most coherent motion dot field in an array of four motion stimuli. In another study, Wang, Miller and Liu (2015) also measured FBA's profile in motion direction, with a better control for spatial attention. In their study, participants performed a two-interval-forced-choice (2-IFC) to detect a coherent motion stimulus (i.e., target) against a random motion stimulus (i.e., noise). The stimuli in the 2IFC task were always presented at the screen center such that spatial attention was fixed and remained constant across conditions. The results were similar to Ho et al. (2012)'s findings and, therefore was also consistent with the feature-similarity gain model. The authors also generalized their findings to other feature dimensions, including orientation and color, based on a similar behavioral paradigm. In sum, initial human psychophysical and neural studies provided converging evidence that FBA modulates perception as a monotonic function of feature similarity.

Potential Issues with the Feature-Similarity Gain Model

Notwithstanding the support for the feature-similarity gain model discussed above, it should be noted that both neural and behavioral findings in human only assessed feature

processing on a coarse scale. A closer examination showed that the previous studies either tested only two orthogonal features (e.g., red vs. green, or upward vs. downward motion direction) or use a coarse sampling in feature space (Saenz, Buracas & Boynton, 2002, 2003; Liu, Larsson & Carrasco, 2007; Serences & Boynton, 2007; Zhang & Luck, 2009; Ho, et al., 2012; Paltoglou & Neri, 2012; Wang, Miller & Liu, 2015). Thus, how feature-based attention modulates the perceptual representation of other similar but different features is unknown.

In addition to the lack of fine sampling in the feature space, another issue posed even more theoretical challenge to the efficiency of the feature-similarity gain model. The key signature of the FSG is the linear modulation dependent on the similarity to the attended feature, which turns enhancement into suppression as the feature become progressively more different from the attended one. While this model can predict a filtering of dissimilar distractor features, it does not seem to be helpful when encountering similar distractor features, which would actually benefit from attentional enhancement because they are similar to the attended feature. Therefore, in recent studies, researchers have turned their attention to investigate the mechanism that underlies attentional modulation in the vicinity of the attended feature.

One line of studies investigated FBA's role in searching a target feature among similar distractors (Navalpakkam and Itti, 2007; Scolari & Serences, 2009; Scolari & Serences, 2010; Scolari et al., 2012). For example, assuming there is a task that requires participants to detect a 55° orientation target among 60° orientation distractors. According to the feature-similarity gain model, one can attend to the 55° orientation such that FBA would enhance the responses of neurons optimally tuned to the target orientation. However, this will also cause the same group of neurons responding more to the distractors (60°) and, therefore, would not increase the overall signal-to-noise ratio (SNR). To resolve this dilemma in FSG model, Navalpakkam and Itti

(2007) proposed an off-channel tuning mechanism such that FBA may be voluntarily deployed to neurons that are sub-optimally tuned to the target feature (e.g., neurons tuned to 50° in the previous example). By shifting attention away to a distant feature, the distractor would be less enhanced as it is more different from the attended "off" channel after shift.

To test this idea, Navalpakkam and Itti instructed participants to search an orientation target (e.g., 55°) among similar and homogenous distractors (e.g., 60°). The authors found that the highest attentional gain was constantly biased and deployed toward the orientation (e.g., 50°) that was further away from the distractors than the target. Later studies also lend support that FBA can be deployed in the off-channel manner to enhance performances in a fine discrimination task (Scolari & Serences, 2009, 2010; Scolari et al., 2012). However, the off-channel gain mechanism may only be facilitative when the target and distractors were linearly separable (D'Zmura, 1991), that is when distractors were sampled from identical side of the attended feature in a feature space. If there are distractors sampled from both sides of the attended feature in a feature space, e.g., 55° orientation embedded in 50° and 60° distractors, the off-channel mechanism may not be helpful, as shifting the attentional gain toward either side in the feature space will result in an enhancement of some distractors. In addition, the off-channel gain requires foreknowledge of both the target and distractors features (Scolari & Serences, 2009, 2010; Scolari et al., 2012). Hence, this mechanism may only facilitate target selection under specific scenarios.

A New Metaphor – the Mexican-hat of FBA

Our visual environment rarely contains homogeneous distractors, or predictable distractor features. In fact, task-irrelevant features may be randomly scattered in a feature space and may also change from time to time (e.g., while driving, the views are constantly changing). Is there a

mechanism of FBA that helps us better adapt to the dynamic and variable visual environment? While feature-similarity gain predicts a suppression of dissimilar features at a coarse scale, some researchers also wonder if there could be any suppressive mechanism in the vicinity of the attended feature to inhibit processing of similar distractors. In the spatial domain, a number of studies have shown that spatial attention elicits a suppressive zone around the attended location to reduce interference from nearby locations (Hopf et al., 2006; Boehler et al., 2009, 2011; Muller & Kleinschmidt, 2004; Mounts, 2000a, 2000b; Tsotsos, 1995, 2011). Importantly, once outside the suppressive zone, behavioral performance or neural activity was not further suppressed – a rebound effect at further locations, which is in line with a non-monotonic "Mexican hat" profile that consists of an excitatory center and suppressive surround.

Recent studies have also extended the investigation of such surround suppression to feature-based attention (Fig. 1b), including the color and orientation domain (Stormer & Alvarez, 2014; Tombu & Tsotsos, 2008). For example, Stormer and Alvarez (2014) addressed whether color-based attention elicits surround suppression to the close neighbors of the attended color. In their study, each hemifield hosted a random motion dot field, in which half the dots were drawn in a target color with the other half drawn in a distractor color. Participants monitored the target dot fields on both sides and reported a brief coherent motion in one of the target dot fields. As one would expect, correct responses were highest when the two dot fields had the same attended color. But what is unexpected was a performance drop when the two attended colors were similar but different, which was then followed by a rebound effect when the two attended colors become more dissimilar. The suppression of similar colors is clearly against what feature-similarity model would predict. Hence, Stormer and Alvarez concluded that the non-monotonic changes of

performances matched the signature pattern of a Mexican-hat profile, therefore suggesting that FBA to colors can elicit a suppressive surround in the color space.

Tombu and Tsotsos (2008) investigated the profile of FBA in the orientation domain. In their study, participants were asked to identify the jaggedness (e.g., straight or jagged) of a grating stimulus that was briefly presented. In addition, researchers also informed participants the most likely orientation of the grating stimulus at the start of a block. Notably, the surround suppression was evident in the results such that the worst performance occurred when the grating's orientation was 45° offset from the attended orientation, which was followed by a rebound at 90° offset from the attended orientation. Such non-monotonic pattern of performance supported the surround suppression effect in attention to orientation.

While the two studies have found initial evidence for a surround suppression effect in FBA, some methodological concerns potentially weakened their conclusions (Stormer & Alvarez, 2014; Tombu & Tsotsos, 2008). First, and critically, there is a lack of baseline condition in these previous studies. A neutral condition is critical to accurately characterize the shape of the attentional profile and rule out alternative interpretations. For example, it is possible that the non-monotonic changes in Stormer & Alvarez's study is caused by perceptual interference when monitoring two colors of different offsets, which could be measured in a neutral condition. To rule out such a confound, the neutral performance should have been subtracted out from the performance under attention condition. In addition, baseline performance is vital in assessing the benefit and cost of attention. Without a proper baseline, it is unclear whether the performance drop reflects a true suppression effect or less enhancement within the surround of the attended feature (Stormer & Alvarez, 2014; Tombu & Tsotsos, 2008).

Second, the task in the previous studies may be sub-optimal for measuring the profile of FBA's modulation. In Stormer and Alvarez's study, participants were required to attend to two colors simultaneously. Recent findings have suggested that there is a limited capability of splitting attention to multiple colors (Liu & Jigo, 2017). Therefore, it is possible that the non-monotonic pattern may be associated with the need to hold and attend to two colors simultaneously. Alternatively, working memory is thought to maintain an attentional template (Desimone & Duncan, 1985; Wolfe, 1994). Therefore, the non-monotonic profile may be due to interference between the templates maintained in working memory instead of a perceptual modulation of FBA to visual input. In addition, Tombu and Tsotsos (2008) employed a task of judging the jaggedness of gratings, which in principle does not require attention to orientation. It is also worth noting that the non-monotonic profile in their study only occurred when the target was jagged but not when the target was straight—a puzzling result that did not have obvious explanations.

Finally, color perception is strongly categorical, which is suggested to play a role in attention. For example, linear non-separability between target and distractors usually lead to inefficient search for a target (Bauer et al., 1998; D'Zmura, 1991). However, recent studies showed that such search can also be much improved when targets and distractors are from different categories than if they come from the same category even when the perceptual similarity between targets and distractors are equated (Daoutis et al., 2006; Hodsoll & Humphreys, 2005). However, the previous study by Stormer and Alvarez used a random selection of colors in a color space. Thus, it is unclear based on their findings how color categories might impact the attentional profile.

Taken together, previous studies provided suggestive but inconclusive evidence that there is a surround suppression effect in feature-based attention. A recent study further tested the attentional profile for color-based attention using a color detection paradigm (Fang, Becker & Liu, 2019). In their study, participants were instructed to detect a coherent color signal against a random noise in a 2-IFC task, in which stimuli were presented at the screen center. Building on the previous studies, the authors made several improvements to better characterize the profile of FBA to colors. First and most importantly, the authors included a neutral condition which provided a proper baseline to establish a genuine suppression effect and better quantify the attentional modulation. Second, participants were instructed to attend a single color, which excluded any potential interference from holding multiple attentional templates (Stormer and Alvarez, 2014). Thirdly, to further reduce task complexity, the signal strength was directly manipulated through color coherence as an analogy to the classic random dot motion kinematogram (Newsome & Pare, 1988). Moreover, they also used a post-cue to reduce response uncertainty, so that performance should reflect FBA's modulation on perception (Pestilli & Carrasco, 2005). As the results showed, the authors found a surround suppression effect that is consistent with previous findings. Interestingly, the authors further revealed that the suppressive surround in color domain also coincided with the color category boundary, which has not been considered in the previous studies (Fig. 1c, left panel). Thus, the surround suppression effect in color-based attention can also be interpreted as a categorical sharpening effect.

The above study naturally raised new questions of whether surround suppression is a specific effect associated with categorical feature like color, or it could also be generalized to other features. In another study, Fang and Liu (2019) conducted a more systematically examination on surround suppression for a series of other important dimensions in early vision

(e.g., orientation, motion direction, & spatial frequency), using a similar 2-IFC task in which participants detect a coherent target feature. They employed a feature cue to direct FBA or an uninformative cue to establish baseline performance. The author found that FBA elicited surround suppression in all three feature dimensions, which suggests that non-monotonic modulation could be a canonical operation of FBA (Fig. 1c). Taken together, these psychophysical studies demonstrate that when the visual attention system faces an unpredictable and dynamic visual environment, it elicits a suppressive surround in feature spaces to enhance the signal-to-noise ratio when the target and distractors have similar features.

An Integrated Model – Flexible Modulation on Different Scales

While such a non-monotonic profile of FBA is in a direct contradiction to the monotonic prediction of the classic feature-similarity gain model, recent studies provided abundant evidence for a surround suppression mechanism that can better isolate a target from similar but different distractors (Fang, Becker & Liu, 2019; Fang & Liu, 2019; Stormer & Alvarez, 2014; Tombu & Tsotsos, 2008). However, if attentional modulation only follows a pure Mexican-hat function, the modulation would continue to rebound to a baseline level for very dissimilar features (Fig.1b). That prediction is inconsistent with previous studies favoring the feature-similarity gain model, which clearly showed a suppression for very dissimilar features. Therefore, the two models appear to be contradictory, and one might wonder which one is correct.

In fact, both models may be correct, but operating on different similarity scale (Fig. 1d). Feature-similarity gain model was mostly supported in studies testing large feature offsets (e.g., red vs. green, upward vs. downward motion), while surround suppression was found in studies using a narrow range near the attended feature. Thus, the final result of FBA's modulation can be regarded as a combination of both feature-similarity gain and surround suppression (Fig. 1d).

This view is further supported by the recent studies (Fig. 1c, Fang, Becker & Liu, 2019; Fang & Liu, 2019). By systematically sampling through feature spaces from a small to large offset, these researchers have consistently revealed a hybrid profile consisting of both surround suppression and feature-similarity gain modulation in dimensions including color, orientation, and motion direction (Fig. 1c). Such a hybrid profile reconciles the findings that feature-similarity gain may be optimal for filtering dissimilar features on a coarse scale, whereas surround suppression can facilitate isolating target from other similar feature on a fine scale (Fig. 1d). Therefore, both feature-similarity gain and surround suppression may be complimentary to each other to better select the desired target information in a complex scene.

Summary

While the feature-similarity gain model remains one of the most prominent models in the attention literature, recent studies have revealed non-monotonic effect that it cannot account for. At a coarse level, the feature-similarity gain predicts a suppression for dissimilar features, which is consistent with behavioral, neuroimaging and single-unit studies. However, on a finer scale, it fails to explain how FBA exclude similar but different distractors to an attended feature. To achieve a more flexible selection of the most task-relevant feature, it is necessary for FBA to efficiently reduce both similar and distinctive distractors in the dynamic environment. A new pattern of attentional modulation, the surround suppression, was discovered such that there is a suppressive zone that enhances the signal-to-noise ratio in the vicinity of the attended feature. Moreover, the classic feature-similarity gain model can be integrated with the surround suppression modulation to enhance the most relevant aspect of the sensory input at the expense of unattended information on both a fine and a coarse similarity scale.

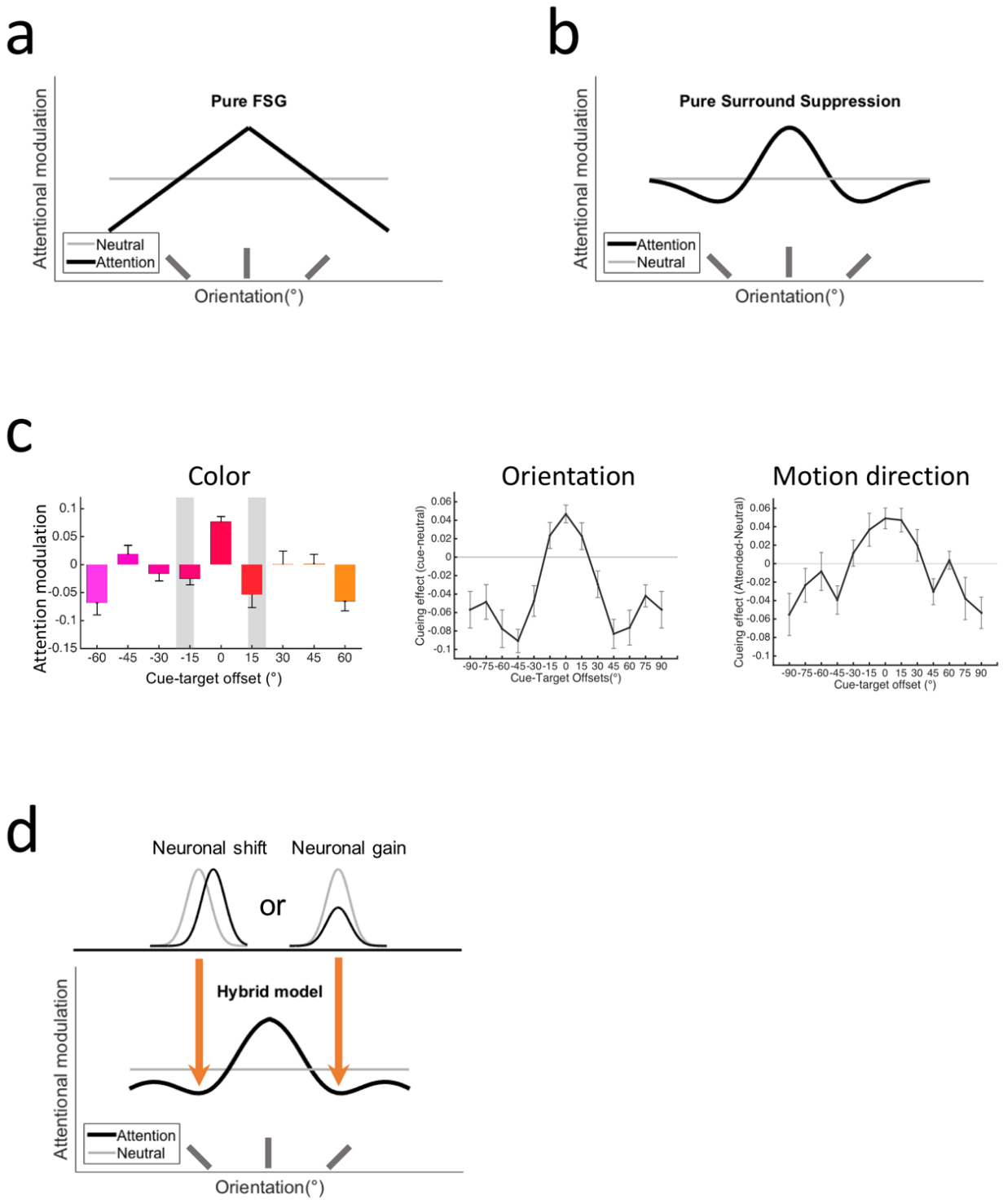


Figure 1. Illustrations for different attentional profiles of FBA. (a) example for a monotonic feature-similarity gain profile in orientation space. (b) example for a pure surround suppression profile in orientation space. (c) empirical behavioral evidence for a hybrid profile in attention to color (Fang, Becker, & Liu, 2019), orientation and motion direction (Fang & Liu, 2019). (d) example for a hybrid profile of FBA (bottom panel) to orientation. Two candidate neural mechanisms underlying surround suppression (top panel), a shift mechanism or a gain mechanism.

CHAPTER 2

NEURAL MECHANISMS OF ATTENTIONAL SURROUND SUPPRESSION

As reviewed in Chapter 1, studies in recent years have shown that FBA can enhance an attended feature at the expense of unattended ones. Yet the neural mechanisms of attentional suppression in feature domain remains unclear, especially in the vicinity of the attended feature. Therefore, the current work focuses on the candidate neural mechanisms underlying the surround suppression that are informed by neurophysiological studies. Importantly, our first aim in the current work is to investigate how recent multivariate methods in computational neuroimaging (e.g., fMRI) may be utilized to distinguish between the candidate neural mechanisms through simulation and computational modeling. In addition, neuronal noise is inherently correlated, which also manifest at the neural population/voxel level. Therefore, the second goal of the current work is to systematically compare the two leading multivariate methods in the presence of correlated neural noise.

Candidate Neuronal Mechanisms of Surround Suppression

While the attentional surround suppression enhances the signal-to-noise at the vicinity of the attended feature, an important question concerns the underlying neural mechanism of the non-monotonic modulation within the suppressive surround. The psychophysical studies above excluded post-perceptual account with a postcue paradigm, which indicated that the surround suppression reduced perceptual sensitivity to the distractors (Fang & Liu, 2019; Fang, Becker, & Liu, 2019). In agreement with the behavioral findings, two electrophysiological studies in humans showed reduced neural responses to features within the surround of the attended feature (Bartsch et al., 2017; Stormer & Alvarez, 2014). Using a frequency-tagging technique, Stormer and Alvarez (2014) found a significantly reduced occipital SSVEP (i.e., steady-state visual

evoked potentials) for colors within the suppressive surround of the attended color. In addition, Bartsch et al., showed with magnetoencephalogram (MEG) that surround suppression emerged in posterior retinotopic visual areas (e.g., VO-1/hV4) within 305 ~ 375 ms after attending to a color. While these results are generally consistent with behavioral effects reviewed above, EEG/MEG measures gross, aggregated signals across large neuronal populations, thus cannot reveal the nature of neuronal level modulations. For example, does surround suppression reduce the overall strength of the stimulus representation, or does it distort the feature space? To further characterize the neural signature of the non-monotonic attentional modulation, we will consider here two prominent neuronal mechanisms underlying FBA (Fig. 1d, top panel): a shift of the tuning preference (i.e., "shift mechanism") or a gain modulation of the tuning curve (i.e., "gain mechanism"), both of which can explain the behavioral surround suppression effect in FBA (Fang et al., 2019; Tsotsos, 2011).

In a crowded scene, both target and distractors are more likely to fall within the same receptive field (RF) and compete for representation. In the spatial domain, one way that spatial attention biases neuronal responses toward the attended stimulus may be through changing the spatial profile of its RF – shifting toward and shrinking around the attended location (Moran & Duncan, 1985). Previous studies have indeed found that spatial attention shifted RF toward an attended location in multiple visual areas, including macaque medial temporal area (Anton-Exrleben, Stephan & Treue, 2009; Womelsdorf et al. 2006, 2008), and V4 (Connor et al. 1997). In human visual cortex, a recent fMRI study showed that spatial attention attracted population receptive field (pRFs) toward an attended location (Klein, Harvey & Dumoulin, 2014). Similarly, recent evidence also suggested that FBA can elicit neuronal tuning shift toward an attended feature (e.g., Fig. 1d top panel, David et al., 2008; Ibos & Freedman, 2014).

Would similar neuronal shift mechanism underlie surround suppression in feature-based attention? Interestingly, recent studies have provided initial insights into such possibility using computational modeling to explore the potential connection between neuronal tuning shift and behavioral surround suppression in feature domain (Fang et al., 2019). Building on the single-unit findings on neuronal tuning shift, Fang, Becker, & Liu (2019) have implemented a computational model with population neural coding and Bayesian read-out rule (Pouget, Dayan & Zemel, 2000, 2003; Ma, Beck, Latham, & Pouget, 2006). Under known physiological constraints, the simple model in their study consisted of a bank of neurons spanning a feature space (e.g., color). The authors simulated their behavioral experiment (i.e., 2IFC) under both attention and neutral condition to measure the profile of feature-based attention's modulation. Interestingly, the neuronal tuning shift successfully led to surround suppression in behavior, which suggested a hitherto unknown relationship between the previous physiological findings and the Mexican-hat profile of behavior. At an intuitive level, the tuning shift that occurred within the vicinity of an attended feature created a vacuum, which weakened representation of features in the suppressive surround. Therefore, it is possible that FBA can elicit a suppressive zone by shifting nearby neurons' tuning preference toward the attended feature.

While changes in the neuronal tuning profile might underlie the attentional surround suppression, it is not the only possible account. Visual attention can also cause a response gain change, when attention is directed to a location (McAdam & Maunsell, 1999), or feature (Treue & Martinez-Trujillo, 1999). At the neuronal level, the response gain modulation can be implemented as a multiplicative factor applied to the tuning curve without changing its tuning preference or width (Fig. 1d top panel). At the behavioral level, such gain modulation enhances perceptual sensitivity of an attended feature, which is analogous to an upscaling of the local

contrast of the attended feature (Herrmann et al., 2012). For example, Herrmann and colleagues found that perceptual sensitivity of attended orientations was higher than the unattended condition across all contrast levels (e.g. from 5% to 85%), which was consistent with a multiplicative response gain modulation at the neuronal level. In the current work, we hypothesized that a similar gain mechanism might also underlie the surround suppression such that the multiplicative gain modulation is a non-monotonic function of the similarity between tuning preference and attended feature on a fine similarity scale.

In the absence of direct physiological data, computational models can provide useful insights on this non-monotonic gain mechanism. For example, the selective tuning model may explain the surround suppression effect in space-based and, potentially, in FBA (Tsotsos, 1995, 2011). The selective tuning (ST) model is a multi-layered computational model that is initially proposed to account for visual processing in the spatial domain (e.g., crowding, spatial resolution). The model has a similar hierarchical structure (e.g., larger RF size in higher level) as the human visual system. In ST model, attentional surround suppression can be elicited through a top-down winner-take-all mechanism, which initiates feedback modulation to inhibit units less tuned to the attended location in earlier layers. This top-down influence can produce spatial surround suppression in early units and is able to account for findings in the spatial domain (Hopf et al., 2006; Boehler et al., 2009, 2011; Muller & Kleinschmidt, 2004; Mounts, 2000a; 2000b). In feature domain, the selective tuning also assumed that the feedback modulation on neuronal tuning curves could be a gain modulation that downscales neural response within the suppressive surround in feature space (Tombu & Tsotsos, 2008; Tsotsos, 2011; Bartsch et al., 2017).

Neural Decoding at the Population Level

In short, surround suppression in FBA could arise from two candidate neuronal mechanisms – a tuning shift mechanism or a gain mechanism (Fig. 1d), both of which can explain the non-monotonic modulation of FBA (Fang et al., 2019; Tsotsos, 2011). A significant question is how to distinguish between these candidate neuronal mechanisms using non-invasive neural measures from the human brain. As a proof of concept, we believe that it is necessary to establish a link between the neuronal mechanisms and their manifestation in aggregated neural measures from human cortex (e.g., at voxel level using fMRI).

Although single-unit studies provide invaluable knowledge of attentional mechanism, it is also unlikely that a few single neurons determine the behavioral response in any task. Information conveyed by neuronal populations likely bear more intimate relationship to representation of stimulus and ultimately behavior (Pouget, Dayan & Zemel, 2000, 2003; Ma, Beck, Latham, & Pouget, 2006). This population-based view has gained increasing recognition in recent years in system and cognitive neuroscience (Churchland et al., 2012; Mante et al., 2013; Sprague, Saproo, & Serences, 2015; Fusi et al., 2016). A challenge is that currently, we do not know how or whether the two neuronal mechanisms could be distinguished at the fMRI voxel level. Therefore, our first goal is to fill this gap by decoding and differentiating manifestations of the neuronal mechanisms at the voxel level through simulation and computational modeling. Because of the limitation in spatial resolution, classic univariate analysis in fMRI imaging only captures the overall responses across neuronal populations, therefore obscuring the underlying multivariate pattern information. Early studies showed that a linear pattern classifier (i.e., multi-voxel pattern analysis, MVPA) can identify the presence of certain stimulus information by training and testing a linear classifier on the spatial pattern of

voxel responses within a region of interest (Kamitani & Tong, 2005). One might wonder whether it is possible to go beyond the voxel level and extract sub-voxel information to distinguish between different neuronal mechanisms. A recent multivariate technique in computational neuroimaging may circumvent such limitation and decode information beyond the resolution of single voxel, which is therefore suitable for current study.

To establish a direct link between the neuronal mechanisms and their modulation on population responses profile, we employed an encoding/decoding model approach in computational neuroimaging (Naselaris et al., 2011; Brouwer & Heeger, 2009). Such an approach has been used by a variety of studies from low-level perceptual phenomenon (e.g., cross-orientation suppression, Brouwer & Heeger, 2011) to higher-level cognition (e.g., working memory, Ester et al., 2013, 2015). A voxel-based encoding model provides a functional description between stimulus input and voxel responses (Naselaris et al., 2011). It starts by encoding different stimulus (e.g., orientation) using hypothetical receptive fields or channels that are informed by physiological evidence. At the voxel level, the response of a single voxel can be modeled as an aggregation of different neuronal population or hypothetical channels. Therefore, it is possible to build a direct mapping through linearly weighted combination to link the encoder's stimulus-evoked responses and voxel responses across neuronal populations. One can fit the encoding model to empirically observed voxel responses (e.g., training data) and analytically estimate the linear weights using linear regression method (i.e., least square estimation). For example, Brouwer and Heeger initially employed a channel-encoding model to examine the neural representation of a continuous color space in visual areas, in which the hypothetical channels resembled the known selectivity of color tuning curves (Brouwer & Heeger, 2009, 2013). While mean voxel responses in visual areas did not reliably differ for

different stimulus colors, Brouwer and Heeger (2009) were able to accurately reconstruct the representations for colors in different visual areas using the forward channel-encoding model, with a similar accuracy as more conventional pattern classification decoding method (Kamitani & Tong, 2005).

After estimating the best-fit encoding model, inversion of the encoding model (i.e., inverted encoding model, IEM) permits one to reconstruct individual channel's responses from a new set of voxel responses measured under different task conditions or cognitive states (e.g., attention). Importantly, the reconstructed channel responses through inversion produces tuned response profiles like population response profile, which may provide important insight into the mechanisms of a variety of cognitive task, including feature-based attention (Scolari et al., 2012; Saproo & Serences, 2014; Ester et al., 2016). For instance, Ester and colleagues (2016) used the inverted encoding model to investigate whether the frontoparietal regions contain continuous or categorical representation of attentional control signal for FBA to orientation. When participants attended to the orientation of gratings, the researchers found that reconstructed channel responses using voxels from frontoparietal regions showed a similar profile to those reconstructed from visual areas, suggesting a continuous representation of sensory information in attentional control regions. Moreover, the peak location of the reconstructed channel response profile may also reveal perceptual distortion in sensory regions caused by higher-level cognitive processes. In another study, Ester et al. (2020) investigated the neural basis of categorical learning by training participants to categorize orientations into two arbitrary groups. In visual areas, the profile of reconstructed channel responses around category boundaries showed a shift toward the center of the category after learning, suggesting a perceptual distortion in orientation space through

learning. Taken together, the IEM method provide a promising way to explore mechanisms at the neuronal population level beyond the limitation of single voxel.

However, caution is suggested when using the reconstructed channel response profiles to infer the underlying neural mechanism. For instance, Liu et al. (2018) tested a well-known property of contrast-invariant orientation tuning in primary visual cortex using the IEM method. Surprisingly, they found an increase in the width of reconstructed channel responses when stimuli's contrast was reduced, inconsistent with findings from single-unit recording studies (Sclar & Freeman, 1982). Their computational model further showed that such changes in the reconstructed responses do not necessarily indicate corresponding changes in the tuning width of neuron, but instead, can be explained by reduced signal-to-noise ratio as contrast is reduced. This latter result raises a reverse-inference issue. At its core, this reflects a lack of examination on the relationship between single-unit activities and the population level responses (e.g., BOLD signal), which further necessitates our current work in bridging the gap across different levels of measurements. By simulating and decoding different neural mechanisms at the population level, the current work may serve as a reference point for investigating the neural mechanisms of surround suppression in future empirical studies.

For decoding purpose, we also considered a Bayesian method, which further transforms the reconstructed channel response function into a posterior probability distribution of the stimulus, given an observed voxel pattern (van Bergen et al., 2015). It was further pointed out that the reconstructed channel response function is contingent on the initial assumption about the channel's specific shape, which is not surprising given that the IEM is essentially a linear regression model (Liu et al., 2018; Gardner & Liu, 2019). The Bayesian approach follows the same structure as the IEM analysis but further models the structure (i.e., covariance matrix) of

correlated voxel noise. Using Bayes' rules and a flat prior, posterior probability can be computed for a stimulus given the observed voxel responses under the assumption of a normal distribution of errors. More importantly, it has been shown that the reconstructed probability distribution is invariant to the model's assumptions of the channels as the Bayesian method reconstructs information about stimulus rather than parameters of the channel (Gardner & Liu, 2019). Thus, it would be useful to assess whether the Bayesian method can also differentiate the candidate neural mechanisms of attentional modulation (i.e., surround suppression) at the population level. In addition, noise correlation is prevalent across neuronal population, which also manifests at the voxel level. Yet how the voxel-wise correlated noise affects different multivariate methods remain uncharted both in the current research domain, as well as in the general literature of fMRI decoding. As a comparison to the standard IEM approach, our second aim is to extend the Bayesian method to evaluate the possible neural mechanisms of surround suppression and to compare both methods in the presence of correlated voxel noise.

In summary, in the current study, we explored the candidate neural mechanisms of surround suppression in FBA using model simulations. We first generated synthetic voxel responses using a neural population model, which implemented different mechanisms of surround suppression at the neuronal level. To differentiate different neural mechanisms at the level of aggregate population measures (i.e., fMRI), we decoded the population codes using both the inverted encoding model method and a Bayesian method. We expect that sub-voxel signature patterns may be identified for different neural mechanism, which can provide a comprehensive description of attentional mechanisms across different levels of measurement. We also hypothesized that the Bayesian method could be more suitable for decoding purpose in the presence of noise correlation. To test this hypothesis, we systematically manipulated the

structure of noise correlation among voxels to compare the two multivariate methods (i.e., standard IEM and Bayesian method). We expect that the current work should provide theoretical and practical guidelines for future empirical studies investigating cortical mechanisms of FBA using non-invasive methods in the human brain.

CHAPTER 3

NEURAL MECHANISMS OF SURROUND SUPPRESSION: SHIFT VS. GAIN

The goal of current simulation is to distinguish the candidate neural mechanisms of the non-monotonic surround suppression at the fMRI voxel level. We modeled the attentional modulation as a hybrid profile, which consisted of the surround suppression on a fine scale and feature-similarity gain modulation on a coarse scale (Fang, Becker, & Liu, 2019; Fang & Liu, 2019). Neurons within the suppressive surround can either shift tuning preferences (i.e., *shift mechanism*) or only changes the overall response amplitude (i.e., *gain mechanism*). To evaluate the population codes of these neuronal mechanisms, we conducted model simulations that consist of three steps. We first described a generative model that simulated the voxel responses under different neuronal mechanisms of FBA (i.e., shift or gain mechanism), and then specified how a channel-encoding model was fitted to the voxel response from a neutral training data set. In the final step, we employed two parallel methods to decode the population codes of the simulated voxels responses. Specifically, we inverted the best-fit channel model as a measure of the population responses and estimated the posterior probability of stimulus at different offsets from the attended one. To test the generalizability of our findings, we repeated the simulations under different combinations of neuronal tuning width parameters and voxel noise parameters.

Method

Population Encoding Model

Neutral condition. Each run of the simulation consisted of three steps. In the first step, we built a population model to generate synthetic neural responses (e.g., voxel responses in fMRI) under the neutral conditions. As shown in Figure 2a, the model contains a bank of identical,

uniformly distributed, orientation-tuned neurons spanning from 0° to 180° in the orientation space. Each neuron's tuning curve is assumed to be a von Mise function, which has the form of

$$f_t(s) = \mathrm{e}^{k\cos(s-\mu_t)} \cdot a \ + b \quad (1)$$

where $f_t(s)$ is the *t-th* neuron's response to an orientation stimulus *s*. κ determines the bandwidth of neuronal tuning curves, which is the same for all neurons but can be varied across different simulations. μ_t is the neuron's preferred orientation, which is evenly distributed from 0° to 179° in 1° increment. *a* determines the amplitude of the neuron's response, and *b* represents the baseline activity. We set the baseline activity for each neuron to be 0 and normalized the area under the tuning curve to be 1 such that the average response across the whole neuronal population remains equal across different tuning width *k*. All neurons (180 in total) are assumed to be independent. For tuning width, we further transformed the κ to full width at half-maximum (in degrees) to facilitate the interpretation of results.

Candidate Neuronal Mechanism of Surround Suppression

Attentional modulation. For the attentional condition, the monotonic feature-similarity gain modulation (FSG, Fig. 2e) is specified as:

$$FSG_t = \beta - \alpha \cdot |offset_t| \quad (2)$$

$$offset_t \ = \mu_t - \mu_{att} \quad (3)$$

where FSG_t is the feature-similarity gain modulation for *t-th* neuron. Both α, β are parameters (slope and intercept respectively) controlling the overall shape of the linear feature-similarity gain (Fig. 2e). μ_t is a neuron's tuning preference, and μ_{att} is the attended orientation (i.e., 90°). By definition, feature-similarity gain only depends on similarity between neuronal tuning preferences and the attended orientation. Therefore, it is expected that the decoded population

response profiles across different offsets will show a monotonic profile under a pure FSG modulation (Fang, Becker, & Liu, 2019). In a preliminary simulation (Fig. 21), we also verified this prediction as a basic check of our model implementation: a feature-similarity gain modulation alone is unable to explain the non-monotonic surround suppression.

To model the non-monotonic FBA, we simulated a hybrid profile of modulation on the neuronal population (Fig. 1a), which consisted of a surround suppression on a fine scale and a feature-similarity gain modulation on a coarse scale (Fang, Becker, & Liu, 2019; Fang & Liu, 2019). Within the suppressive surround, FBA could either elicit a gain change of the neurons' overall response (i.e., gain mechanism, Fig. 2d), or shift their tuning preferences (i.e., shift mechanism, Fig. 2c) toward the attended feature. In addition to simulating the non-monotonic surround suppression on a fine scale (e.g., range of the suppressive surround: ±45° offsets), we also implemented a monotonic feature-similarity gain modulation on a coarse scale (e.g., ±90° offset). Therefore, the overall profile of attentional modulation shows a hybrid shape (Fang, Becker, & Liu, 2019; Fang & Liu, 2019). In the next part, I will describe the implementation of the different neuronal mechanisms underlying surround suppression.

In the first scenario, we implemented the gain mechanism as the neuronal mechanism underlying surround suppression, which only affects the overall responsivity of neuronal tuning curves without changing their preferred orientation (Fig. 2d). For a neuron, the gain modulation is simulated by a scaling parameter, which is multiplied with the neuronal tuning function (similar to parameter *a* in Eq. 1). Across the entire orientation space and for different neuronal group, the multiplicative gain modulation of FBA is implemented as a piecewise function (Fig. 2d):

$$G_t = \begin{cases} A_1 \cdot e^{-\frac{(\mu_t-\mu_{att})^2}{2w_1^2}} - A_2 \cdot e^{-\frac{(\mu_t-\mu_{att})^2}{2w_2^2}} + L, & \text{if } |\mu_t - \mu_{att}| < 1.25SS_{range} \\ \beta - \alpha \cdot |\mu_t - \mu_{att}|, & \text{otherwise} \end{cases} \quad (4)$$

where $\{A_1, w_1, A_2, w_2, L, \alpha, \beta\}$ are the parameters controlling the overall shape of the piecewise function. SS_{range} represents the offset (i.e., 45°), where the maximum surround suppression occurred (Fig. 2d). μ_t and μ_{att} are neuronal tuning preference and attended orientation. For neurons that are within a range of $1.25SS_{range}$ from the attended feature, the piece-wise function simulates a non-monotonic surround suppression modulation using a difference of Gaussian function. Once outside the suppressive surround (i.e., |offsets| >= $1.25SS_{range}$), there is a further suppression up to ±90° offset (i.e. feature-similarity gain modulation, Eq. 2). As illustrated in Fig. 2d, the overall shape of the FBA therefore has a hybrid profile on both sides of the attended feature, which is similar to empirical findings (Fang, Becker, & Liu, 2019; Fang & Liu, 2019).

In the second scenario, we also assumed hybrid profile of FBA modulation across the entire orientation space. The critical difference is a shift in neuronal tuning within the suppressive surround (Fig. 2c) toward the attended orientation. Our previous simulation showed that a tuning shift in the vicinity of attended feature can elicit a suppressive surround (Fang, Becker, & Liu, 2019; Fang & Liu, 2019). In conjunction with the monotonic feature-similarity gain modulation (e.g., a monotonic function), the stimulation can further explain the suppression effects found at different scales. Following the previous works, the hybrid profile of FBA in the second scenario was implemented as a combination of range-limited (i.e., up to the suppressive surround then gradually stop) neuronal tuning shift and a feature-similarity gain modulation.

For feature-similarity gain modulation, we used the same linear function as in Eq. 2. We then implemented an attention-induced shift in tuning preference toward the attended feature

(i.e., matched filter). This shift is assumed to be proportional to the distance between tuning preference and the attended feature in our previous model containing uniformly tuned units (Fang et al., 2019), which was specified by a piece-wise linear function.

$$Shift_t = \begin{cases} 0.5 \cdot \ offset_t, & if\ |offset_t| \le SS_{range} \\ 2 \cdot sgn\ (offset_t) \cdot (1.25w - |offset_t|), & if\ SS_{range} < |offset_t| \le 1.25SS_{range} \\ 0, & if\ |offset_t| > 1.25SS_{range} \end{cases} \quad (5)$$

where *sgn* is the sign function, and SS_{range} = 45°, in which maximum surround suppression occurs. This results in a larger shift as neurons move further away from the attended feature followed by a reduced shift outside the suppressive surround (i.e., rebound). Once $Shift_t$ declines to 0, the tuning shift also stops. Under this scenario, neuronal responses were calculated in the same fashion as in Eq. 1, except that neuron's preferred orientation (μ_t), was replaced by (μ_t - $Shift_k$), representing a shift in tuning preference.

Step 1: Simulating Voxel Responses Under Neutral and Attentional Condition

We then simulated response of each voxel (N = 100 in total), which contains neurons tuned to all possible orientations. Each voxel was simulated as linear combination of neuronal responses, which is defined as:

$$v_i(s) = \sum_{t=1}^{180} W_{it}^{neuron} f_t(\mathrm{s}) \quad (6)$$

where $v_i(s)$ is the i-th voxel's tuning curve. W_{it}^{neuron} is the linear weight of t-th neuron in this voxel. The linear weight (W_{it}^{neuron}) contains 180 numbers drawn from a uniform distribution between [0, 1], which describe the relative contribution of different neuronal populations to a voxel's response. After the weighted sum, voxel response is scaled (average response across all voxels: ~ 1) such that it is close to the common range as blood-oxygen-level-dependent (BOLD)

responses (percent of signal change). To generate the final voxel response, we further added correlated noise *e* to the voxel responses (Step.1 in Fig. 2e).

$$\mathbf{B} = \mathbf{v}(\mathrm{s}) + \boldsymbol{e} \tag{7}$$

The noise term *e* is randomly sampled from a multivariate Normal distribution with a mean of 0 and voxel-by-voxel covariance matrix of Σ:

$$\boldsymbol{e} \sim N(0, \boldsymbol{\Sigma}) \tag{8}$$

The covariance matrix Σ has the form as:

$$\Sigma_{ij} = \tau_i \tau_j R_{ij} \tag{9}$$

where the pairwise covariance, Σ_{ij}, between the i-th and j-th voxel was computed as the product of voxel standard deviation τ_i, τ_j, and their pairwise correlation, R_{ij}.

The standard deviations (τ) of voxel responses are proportional to the average voxel responses before adding the noise term:

$$\tau_{\mathrm{i}}(\mathrm{s}) = \frac{\lambda}{m} \sum_{i=1}^{m} v_i(s) \tag{10}$$

where *m* is the total number of voxels (100 in total). λ is the proportion between voxel standard deviation (τ) and average response of all voxels to a certain orientation stimulus. The voxel-by-voxel correlation, *R,* is constructed by a combination of a voxel-tuning-dependent correlation R^{tuning} and arbitrary correlation R^{arb} that is independent of voxel's tuning property, which can be caused by thermal and physiological variabilities in fMRI signal. It has been shown that such a correlation structure can well explain the voxel-wise noise correlation in empirical fMRI data (van Bergen et al., 2015; van Bergen & Jehee, 2018). Therefore, we employed a similar structure to generate correlated noise in the current model. The tuning-dependent correlation coefficient R^{tuning} is given by (cf. van Bergen & Jehee, 2018):

$$R_{ij}^{tuning} = r \cdot (1 - I_{ij}) \cdot corr(v_i(s), v_j(s)) + I_{ij} \quad (11)$$

where the *r* is a scaling parameter that controls the strength of correlation between voxels. *I* is an identity matrix. For the correlated voxel noise that is independent of voxel tuning property, we refer to it as arbitrary noise, R^{arb}. To create the arbitrary noise, we shuffled the R^{tuning} such that columns and rows of R^{tuning} were reordered in the same randomized order. This is to ensure that R^{arb} is still a symmetric matrix after being shuffled. Critically, R^{arb} has the same overall correlation but now noise correlation does not depend on the tuning property. In other words, the R^{arb} installs noise that is randomly correlated. Having defined both R^{arb} and R^{tuning}, the final correlation matrix *R* is generated as a combination of R^{tuning} and R^{arb}, which is described as:

$$R_{ij} = (1 - I_{ij})[p \cdot R_{ij}^{tuning} + (1 - p) \cdot R_{ij}^{arb}] + I_{ij} \quad (12)$$

where *p* is a parameter from [0, 1] that specifies the relative contribution of R^{tuning} and R^{arb}. We analyzed the covariance matrix from empirical data (cf. Liu, et al., 2018) with Eq. 15 (see below) to estimate the ratio of *tuning-dependent* and the *tuning-independent noise,* which yielded a ratio of 2.5:1. Therefore, we fixed the *p* parameter to be 2.5/(2.5 +1) = 0.71 in the main simulation. In the benchmark test section (see Chapter 4), we also explored the effect of varying the ratio between R^{tuning} and R^{arb} on decoding of population codes using different multivariate methods.

Simulated Experiment. We simulated voxel responses under a neutral (Fig. 2a) and an attentional condition (Fig. 2c – 2e). Eight orientations were sampled evenly through the whole orientation space (i.e., 0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°). In the neutral condition, each orientation was presented for 32 trials, yielding a total of 256 trials in total. No attentional modulation (i.e., gain or shift) was applied to the neutral data. In the attentional condition, we fixed the attended orientation at 90° (i.e., full space ranges from 0 ° to 179° at 1° increment), resulting eight offset conditions (i.e., offset: -90°, -67.5°, -45°, 22.5°, 0°, 22.5°, 45°, 67.5° for the

corresponding orientations above) that yielded 256 trials (32 trials/offset). As noted above, we simulated a hybrid profile for FBA following recent empirical findings, which revealed a surround suppression on a fine scale (e.g., ±45° offset) and a feature-similarity gain modulation on a coarse scale (Fang, Becker, & Liu, 2019; Fang & Liu, 2019). For both neutral and attentional condition, we modeled voxels' responses using the same linear weight (W^{neuron} in Eq. 6) and covariance matrix (Σ in Eq. 8) described above. In other words, we assumed that FBA did not alter the covariance structure of the noise between the neutral and attentional condition. We simulated another set of data under neutral condition to use as a validation data set, which was used to compare the accuracy in a benchmark test between the different decoding methods.

Step 2: Fitting a Channel-Encoding Model to Voxel Patterns

Fit channel-encoding model. In the second stage, we employed a channel encoding model as proposed by Brouwer and Heeger (2009), to characterize the orientation tuning function. These channels serve as model basis functions that span a model-based information space as an analogy to the activity space spanned by neuronal population, where each axis is a neuronal population. Similar to how the voxels can be treated as a linear combination of neurons, the channel encoding model assumes a voxel's response can be expressed as a linearly weighted combination of a set of channels, which are hypothetical tuning curves evenly distributed in the orientation space. Similarly, the linear weights specified the contribution of each hypothetical channel to a voxel's response. We fitted the channel encoding model to training data from the previous step (e.g., neutral condition) to estimate the linear weights for each channel.

The channel-encoding model consisted of 8 evenly distributed channels (i.e., model basis functions) covering the full orientation space (0° to 179°). Each channel is a half-wave rectified sinusoidal raised to power of 7 (Fig. 2b), which yields an equivalent bandwidth of 25° at half

maximum. As neuronal tuning widths can be variable in visual cortex, we also varied the channel basis function with two additional sets of bandwidths (45° or 65°). Therefore, we employed three subtypes (e.g., channel bandwidth: 25°, 45°, 65°) of channel-encoding models in our simulation.

The hypothetical channels' responses across trials can be expressed as a matrix of $\underline{n}$ by h matrix, C_{train}, where $n = 256$ is the number of trials in a training data set, and $h = 8$ is the number of channels (i.e., a total of 8 channels). The training data is a set of voxel responses simulated under the neutral condition, which is a n by m (i.e., m = 100, number of voxels) matrix B_{train}. W is the linear weight matrix of h by m, where each column describes the channel's contribution to a voxel's response. Therefore, the relation between voxel responses and the channel basis function is given by:

$$\boldsymbol{B_{train} = C_{train}\, W} \tag{13}$$

Given both B_{train} and C_{train}, the least-square estimation of W is defined as:

$$\widehat{\boldsymbol{W}} = \left(\boldsymbol{C_{train}^{T} C_{train}}\right)^{-1} \boldsymbol{C_{train}^{T}\, B_{train}} \tag{14}$$

To further estimate the structure of variability within the voxel responses, we fitted a noise model to the residual term after removing the best fitting voxel response, $\mathbf{C_{train}}\widehat{\mathbf{W}}$. The noise model assumes that the covariance of the voxel noise consists of both a voxel tuning independent component and a voxel tuning dependent component, which is defined as follows (see van Bergen et al., 2015 for a detailed derivation):

$$\boldsymbol{\Omega} = \rho\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + (1-\rho)\,\mathbf{I} \circ \boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + \sigma^{2}\widehat{\boldsymbol{W}}^{T}\widehat{\boldsymbol{W}} \tag{15}$$

where $\boldsymbol{\Omega}$ represents the covariance matrix of voxel noise, ρ scales voxel noise irrespectively of their tuning similarity (i.e., equivalent to tuning-independent noise R^{arb} in Eq. 12), $\boldsymbol{\tau}$ is the standard deviation of voxel response, σ is the standard deviation of model's channel (i.e., model

basis function), and $\widehat{\mathbf{W}}$ is the estimated linear weight matrix from Eq. 14 (see above), **I** is the identity matrix.

Assuming Gaussian distribution of the voxels' residual term, we used the maximum likelihood estimation to fit the noise model by finding the parameters $\hat{\mathbf{q}}$ that maximize the joint probability of the given voxel responses.

$$\hat{\mathbf{q}} = \text{argmax}\left(\sum_{1}^{n} \ln(p(\boldsymbol{B}|s, \boldsymbol{\theta}))\right) \tag{16}$$

where $\hat{\mathbf{q}} = \{\hat{\rho}, \hat{\tau}, \hat{\sigma}, \widehat{W}\}$, n is number of trials, and $p(\boldsymbol{B_n}|s, \boldsymbol{\theta})$ is the conditional probability of voxel response given a stimulus s in a single trial.

The conditional probability $p(\boldsymbol{B}|s, \boldsymbol{\theta})$ is defined as:

$$p(\boldsymbol{B}|s, \boldsymbol{\theta}) \propto \exp\left(\left(\boldsymbol{B_{train}} - \boldsymbol{C_{train}}\widehat{\boldsymbol{W}}\right)\boldsymbol{\Omega^{-1}}\left(\boldsymbol{B_{train}} - \boldsymbol{C_{train}}\widehat{\boldsymbol{W}}\right)^{\mathrm{T}}\right) \tag{17}$$

Step 3: Decoding Population Codes

Decoding neural responses. After estimating the best fit encoding model using training data set, inversion of the encoding model can be used to decode information of stimuli given a test data set of voxel responses. Test data sets (*n* by *m* matrix) were generated under both the neutral (i.e., validation data set) and attentional condition (i.e., for decoding the modulation of surround suppression at the population level). Decoding was performed using 3 different sets of channel-base function (i.e., width: 25°, 45°, 65°).

Inversion of Eq. 13 on testing data B_{test} can reconstruct channels' responses, $\hat{\mathbf{C}}_{\mathbf{test}}$, to a test stimulus. The reconstructed channel response is denoted as channel response function (CRF), which is considered as an approximation of neural population response of a certain stimulus (Scolari et al., 2012; Garcia et al., 2013; Ester et al., 2016; Sprague, Boynton & Serences, 2019, also see Gardner & Liu, 2019).

$$\widehat{\boldsymbol{C}}_{\boldsymbol{test}} = \boldsymbol{B}_{\boldsymbol{test}}\widehat{\boldsymbol{W}}^{\boldsymbol{T}}\left(\widehat{\boldsymbol{W}}\,\widehat{\boldsymbol{W}}^{\boldsymbol{T}}\right)^{-\mathbf{1}} \quad (18)$$

We also used the estimated covariance matrix, **Ω,** to generate posterior probability of a stimulus given the test data, using the same method derived by van Bergen and colleagues (van Bergen et al., 2015). After applying Bayes' rule with a flat prior, the posterior probability of a stimulus given a voxel response is defined as (see van Bergen et al., 2015 for a detailed derivation):

$$p(s|\boldsymbol{B}, \hat{\mathbf{q}}) = \frac{p(\boldsymbol{B}|s, \hat{\mathbf{q}})}{\int p(\boldsymbol{B}|s, \hat{\mathbf{q}})p(s)ds} \quad (19)$$

where the conditional probability $p(\boldsymbol{B}|s, \hat{\mathbf{q}})$ is computed using covariance matrix **Ω**, the normalization term $\int p(\boldsymbol{B}|s, \hat{\mathbf{q}})p(s)ds$ is computed numerically by summing all possible values of $p(\boldsymbol{B}|s, \hat{\mathbf{q}})$ spanning the whole orientation space (0° to 179°, at 1° increment).

Identifying Signature Patterns at Neural Population Level

To further evaluate the difference between the surround suppression's underlying mechanism at the population level, we also manipulated two independent variables: neuronal tuning width (κ in Eq. 1, transformed into degrees of full width at half maximum) and voxel standard deviation (τ). We set nine different neuronal tuning width (i.e., κ, equivalent full bandwidth at half maximum: 25°, 30°, 35°, 40°, 45°, 50°, 55°, 60°, 65°) and eight voxel standard deviation (λ: 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, & 35% of average voxel response before the noise term was added). We performed ten independent simulations for each combination of a neuronal tuning width (κ) and voxel standard deviation (τ).

For each run of the simulation, we performed decoding using channel-encoding model with three different sets of channel basis function (i.e., channel width: 25°, 45°, or 65°). We fitted the circular Gaussian template (Eq. 1, four free parameters) to the reconstructed channel

response function and estimated posterior probability distribution for each individual cue-target offset (i.e., -90°, -67.5°, -45°, 22.5°, 0°, 22.5°, 45°, 67.5°). We then compared how different attentional mechanisms affected the fitted parameters (i.e., mean, width, amplitude, and baseline). We analyzed each of the ten runs separately and then averaged results across all runs.

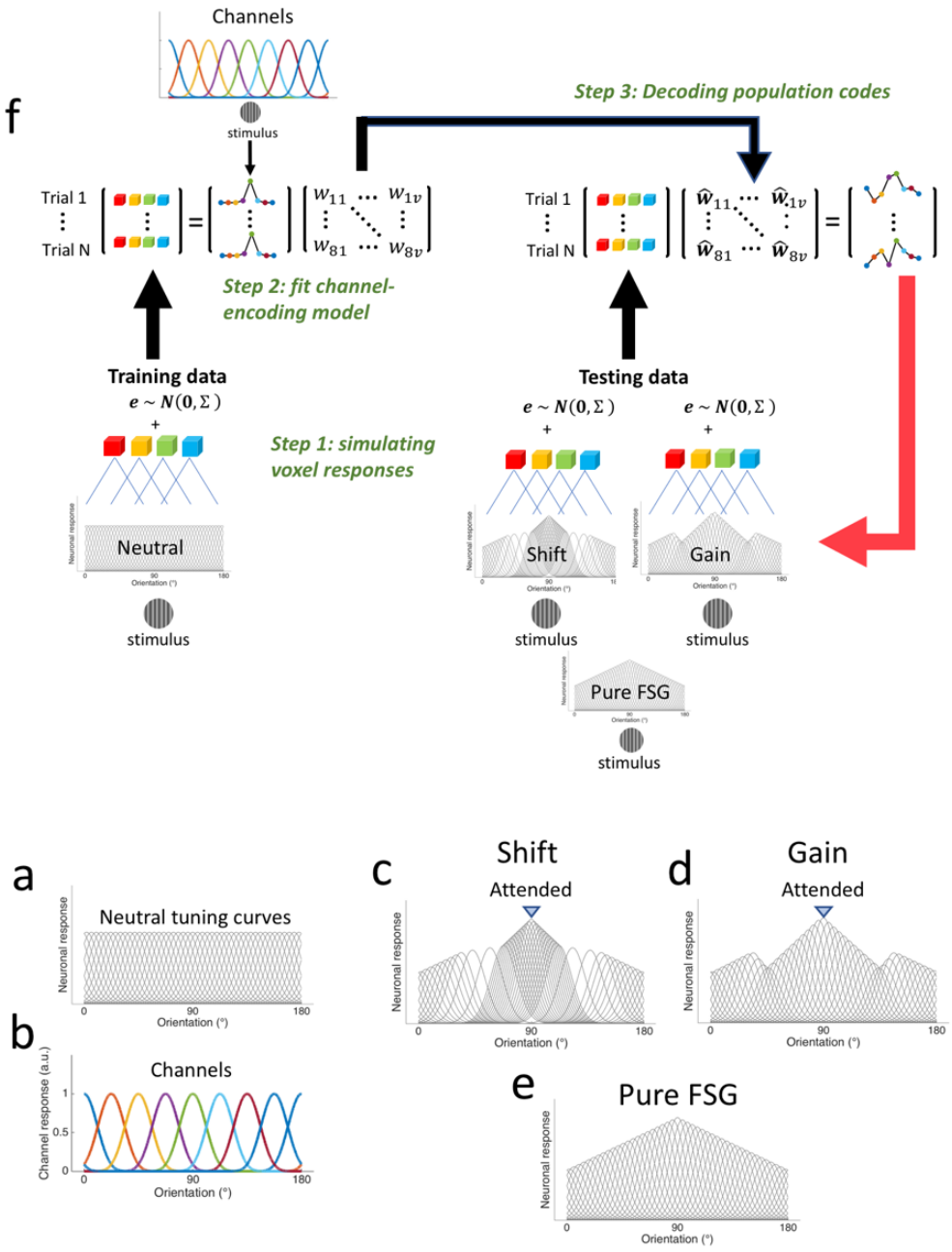


Figure 2. Model architecture and simulation schematic. (a) neuronal tuning curves under a neutral condition. (b) Idealized orientation-tuned channels (i.e., 8 in total) in the channel-encoding model. (c) neuronal tuning curves under a hybrid modulation of both feature-similarity gain and tuning shift. (d) neuronal tuning curves under a hybrid gain modulation of both feature-similarity gain and surround suppression gain. (e) neuronal tuning curve under a pure feature-similarity gain (FSG) modulation. (f) simulation consisted of 3 critical steps. Step 1: simulating voxel responses with voxel-wise correlated noise under both neutral (i.e., training data set) and attentional condition (i.e., testing data set). Step 2: fitting channel-encoding model to estimated channel weights. Step 3: Decoding population codes by inverting the best-fit channel-encoding model. The reconstructed population profiles establish a direct link to neuronal mechanisms (e.g., red arrow).

Results

In the current simulation work, we employed an encoding/decoding model approach to assess the candidate neuronal mechanisms of surround suppression at the population level using two multivariate methods. Under a neutral condition, our model assumed a bank of neurons that were evenly distributed across the entire orientation space. The neuronal responses to an orientation stimulus (e.g., 0° orientation) were linearly combined using random weights to generate voxel response. We also implemented correlated voxel noise that was sampled from a multivariate normal distribution with a covariance matrix that described a mixture of both tuning-dependent and tuning-independent correlation. We trained a channel-encoding model under the neutral condition using three sets of channel basis function. We then inverted the best-fit model to reconstruct the channel response function and posterior stimulus probability distributions under different attentional conditions. To evaluate the candidate neuronal mechanisms for surround suppression, we contrasted their manifestation at the population level under different parameter combinations of neuronal tuning width and voxel variability. To better explain the findings, results shown below were obtained from a specific combination of neuronal tuning width (κ in Eq. 1, equivalent to 40° in orientation space), and voxel variance (λ = 0.15, Eq. 6). Full simulation results are shown in Figures 9 to 21.

Signature Patterns at Neural Population level

For orientations at different offset (e.g., 0°, ±22.5°, ±45°, ±67.5°, 90°) relative to the attended orientation, we first reconstructed their individual channel response functions (CRF) to evaluate the attentional modulations on population responses (Fig. 3a – 3c & Fig. 4a – 4c). We also employed a Bayesian method to decode the probabilistic stimulus representation at each offset (Fig. 3d – 3f & Fig. 4d – 4f) after analyzing the correlated noise structure (i.e., the

covariance matrix) among voxels (van Bergen et al., 2015). The estimated probability distribution showed a continuous distribution in the orientation space, with peak location representing the most likely stimulus, and the width representing the stimulus uncertainty. As shown in Fig. 3 & Fig. 4, orientation stimuli at different offsets were decoded using both methods under the attentional (i.e., solid line) and the neutral condition (i.e., dashed line). The decoding analysis was repeated using three different set of channel basis function (i.e., 25° - Fig. 3a, 3c, 4a, & 4c, 45° - Fig. 3b, 3e, 4b, & 4e, 65° - Fig. 3d, 3f, 4d, & 4f)

Tuning shift mechanism. In the first scenario, we assumed that the attentional surround suppression was caused by a shift of neuronal tuning preference toward the attended feature (i.e., shifting mechanism). Interestingly, we first observed that such attentional attraction elicited an inflation of width at the attended orientation (i.e., Fig. 3 all panels, 0° offset pink solid curve) in both the reconstructed CRF (Fig. 3a-3c) and the posterior probability distribution (Fig. 3d-3f), when comparing with the neutral condition (i.e., pink dashed line). For the neurons that were originally tuned to the nearby features from the attended feature, FBA shifted their tuning preference to become more responsive to the attended feature than in the neutral condition. This shift essentially led to an over-abundance of neurons tuned to the attended feature. Therefore, we observed such an inflation in the neuronal population profile for the attended feature. As the tuning shift was imperfect and did not completely overlap with the attended feature (David et al., 2008), a gradient shape was seen in the inflated neuronal population profile for 0° offset.

As the presented orientation deviated from the attended orientation (e.g., Fig. 3 all panels, cyan curves at ±22.5°, blue curves at ±45°), we found a repulsion effect (i.e., shift away from the attended feature) in both the reconstructed CRF (Fig. 3a-3c) and posterior probabilities (Fig. 3b-3f) as compared to the neutral condition. This is because the “labeled-line” architecture of the

model, i.e., the orientation labels of all the neurons remained the same in the attention and neutral condition. Thus, a nearby feature could activate neurons tuned to further-away features (e.g., a 25° stimulus activating neurons tuned to 35°), causing the decoder to classify the stimulus as repulsed from the attended feature. The repulsion effect only appeared in the intermediate offsets (e.g., ±22.5° and ±45°), and as the stimulus deviated further away from the attended orientation, such repulsion effect disappeared. This was caused by a gradual stop in the tuning shift for neurons at larger offsets, as implemented in the model.

Therefore, neuronal tuning shift mechanism is manifested as a repulsion effect in the population response around the suppressive surround, which creates an attentional distortion for similar but different features from the attended one. Such findings are further consistent with our previous simulation work, which suggested that the surround suppression might enhance feature resolution through repulsion (Fang et al., 2019).

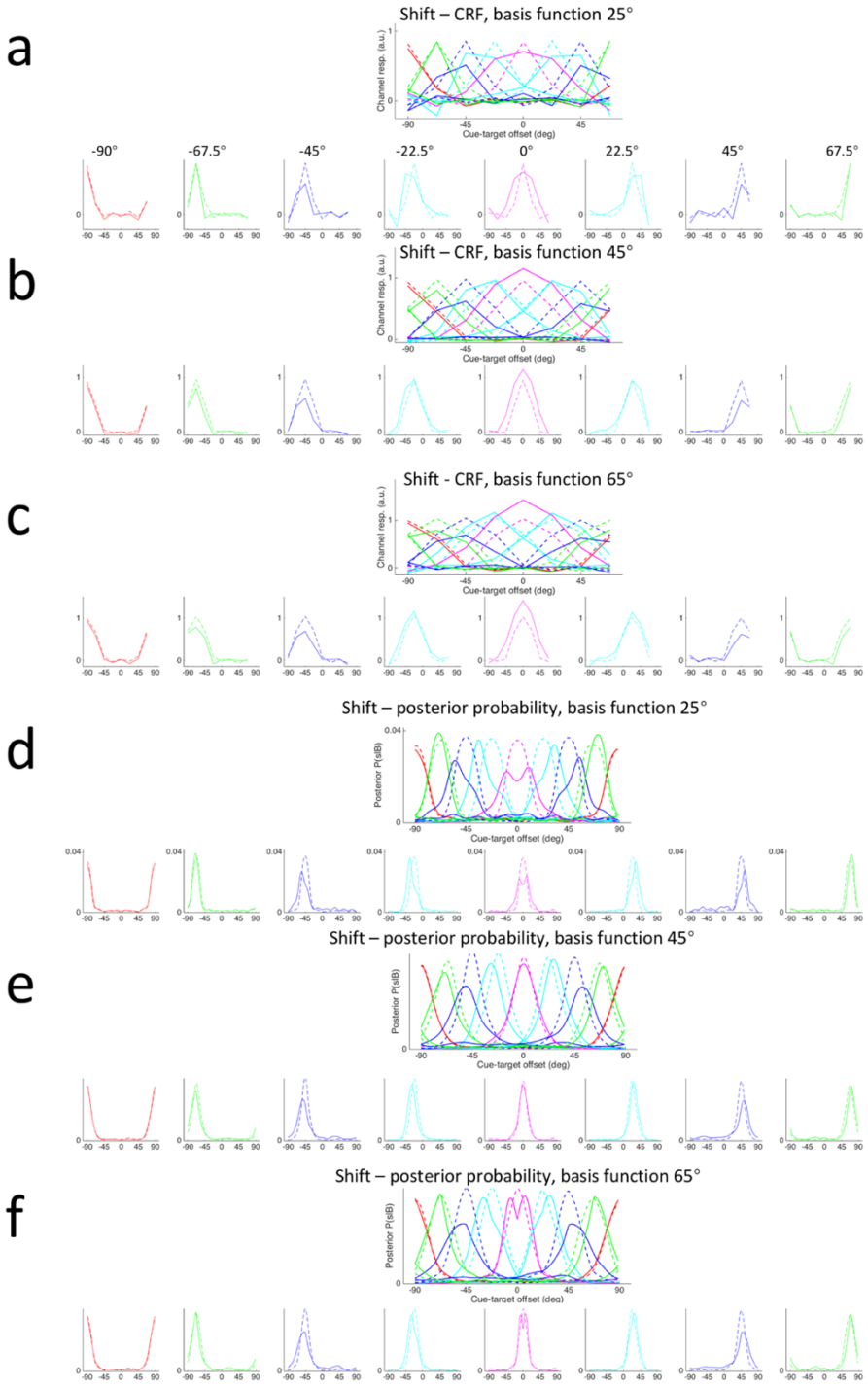


Figure 3. Shifting mechanism - population level profile for individual cue-target offset. (**a**) Reconstructed channel response function (CRF) under attentional (solid curves), and neutral condition (dashed curves). Channel Basis function: 25°. Top panel: reconstructed CRF for each individual cue-target offset plotted in colors (e.g., magenta: 0°, cyan: ±22.5°, blue: ±45°, green: ±67.5°, red: 90°). Bottom panels: reconstructed CRF plotted separately for each offset. (**b**) & (**c**), same as (**a**) except that channel basis function's width was 45° in (**b**) and 65° in (**c**). (**d**) Estimated posterior probability distribution (attentional: solid curves, neutral: dashed curves). Channel basis function: 25°. (**e**) & (**f**), same as (**d**) except that channel basis function's width was 45° in (**e**) and 65° in (**f**).

Gain mechanism. In the second scenario, we assumed that attentional surround suppression only affected the amplitude of neuronal tuning curve without changing the tuning preference of neurons within the surround. For the gain mechanism, we observed a qualitative different pattern with no significant repulsion effect as found with the shifting paradigm. Instead, the most obvious pattern is located at ±45° offset manifested as a reduction of the reconstructed CRF (Fig. 4a – 4c, blue solid at ±45° offset) and a downscale of posterior probability distributions at the intermediate offsets (Fig. 4d – 4f, blue solid at ±45° offset), which was a consequence of the suppression on neuronal gain that we implemented when generating the data.

Importantly, the changes in the overall responsivity of CRF showed a non-monotonic pattern, such that there was a significant reduction of the recovered CRF at intermediate offsets (e.g., Fig. 4a – 4c, blue solid curves) followed by a rebound at larger offsets. As shown in Fig. 4d – 4f, the posterior probability distributions paralleled the CRFs' patterns and showed a similar non-monotonic change. The lowered probability distribution within the suppressive surround suggested an increase in the uncertainty of the stimulus. This is consistent with observations from neurophysiological studies that gain modulation is equivalent to changing the local contrast of stimuli (Treue & Martinez-Trujillo, 1999; Reynolds et al., 2000; McAdam & Maunsell, 2000; Martinez-Trujillo & Treue, 2002).

In short, the gain mechanism can elicit a surround suppression modulation, but without a distortion of feature space. This is a qualitative different population pattern as compared to those under the tuning shift mechanism.

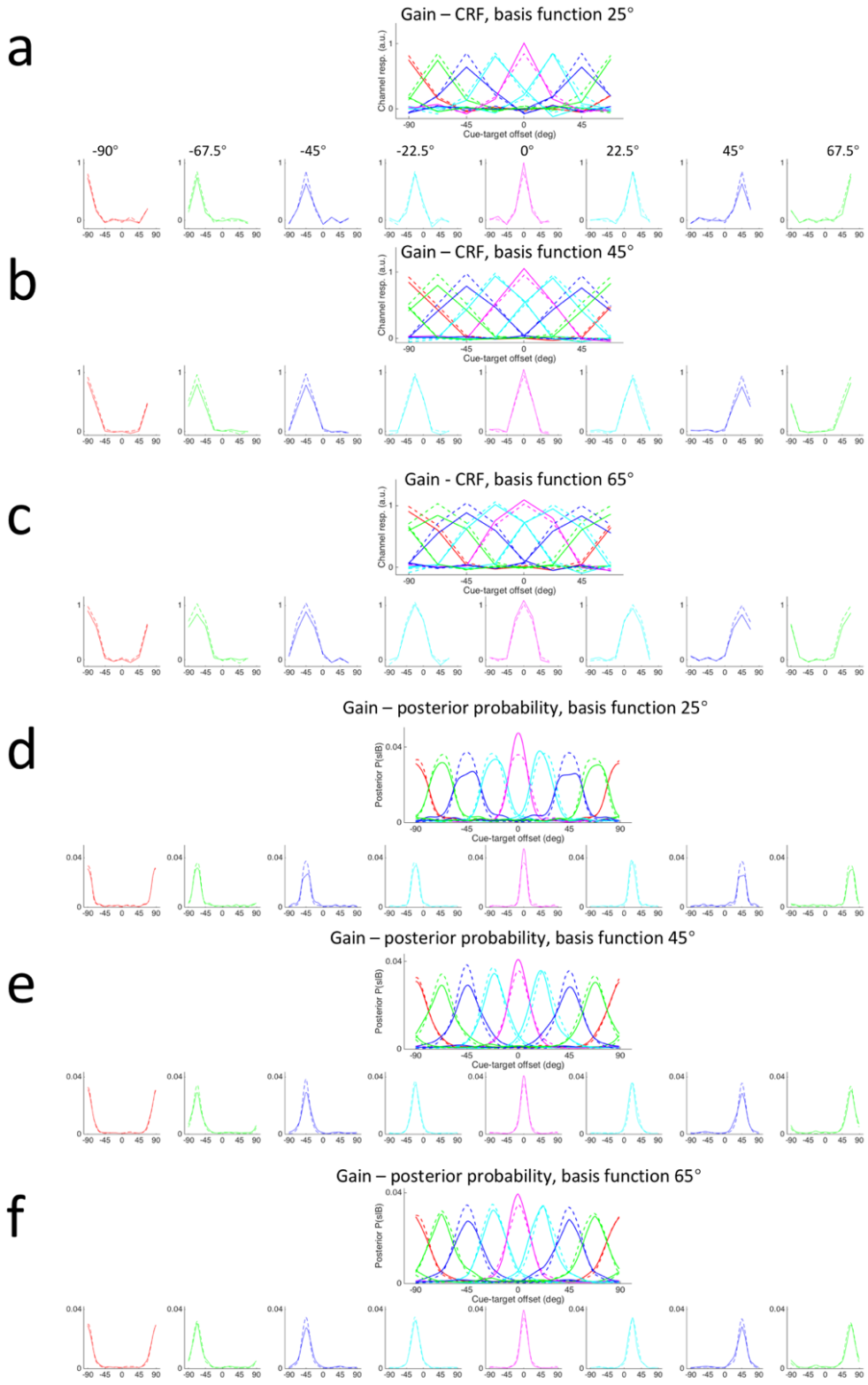


Figure 4. Gain mechanism – population level profile for individual cue-target offset. Figure convention is same as in Figure 3. (**a**), (**b**), & (**c**), reconstructed channel response function using standard IEM method. Channel basis function: 25° for (**a**), 45° for (**b**), and 65° for (**c**). (**d**), (**e**) & (**f**), posterior probability distributions using channel basis function of 25°, 45° & 65° respectively.

Further Comparison Between the Shift and the Gain Mechanisms

To further visualize and quantify how different mechanisms of surround suppression modulate the information contained within voxel responses, we fitted the circular Gaussian template (Eq. 1) separately to the CRF and posterior probability distribution that were averaged across trials for each individual offset condition. This allowed us to examine how attention impacted the four parameters of the fitted Gaussian: baseline, amplitude, mean, and width. This analysis was based on averaging the fitted parameters across the 10 simulation runs for each combination of neuronal tuning width and voxel variance, for each of the channel-basis function (e.g., channel width: 25°, 45°, 65°). We found that three parameters (amplitude, mean and width) were significantly modulated as a function of offsets except the baseline parameter which did not show any systematic profile in accordance with the non-monotonic attentional modulation. We will not further consider the baseline parameter, and next we discuss the other three parameters.

Amplitude. The estimated amplitude parameter for both CRF (Fig. 5e & 5f, black) and posterior probability distribution (Fig. 5e & 5f, red) followed a non-monotonic Mexican-hat pattern across most conditions. Interestingly, these patterns were similar under both the shift (Fig. 5e) and the gain mechanism (Fig. 5f). This result serves as a validation for the claim that both shift and gain mechanism can underlie the surround suppression modulation, as both mechanisms reduced the strength (i.e., amplitude) of neural representations for distractors within the suppressive surround. This also means that the amplitude parameter cannot serve as a robust indicator to differentiate between the two neural mechanisms due to their similar monotonic pattern. We now turn to the results for the other two parameters: mean and width, which provided qualitatively different patterns between shift and gain mechanisms.

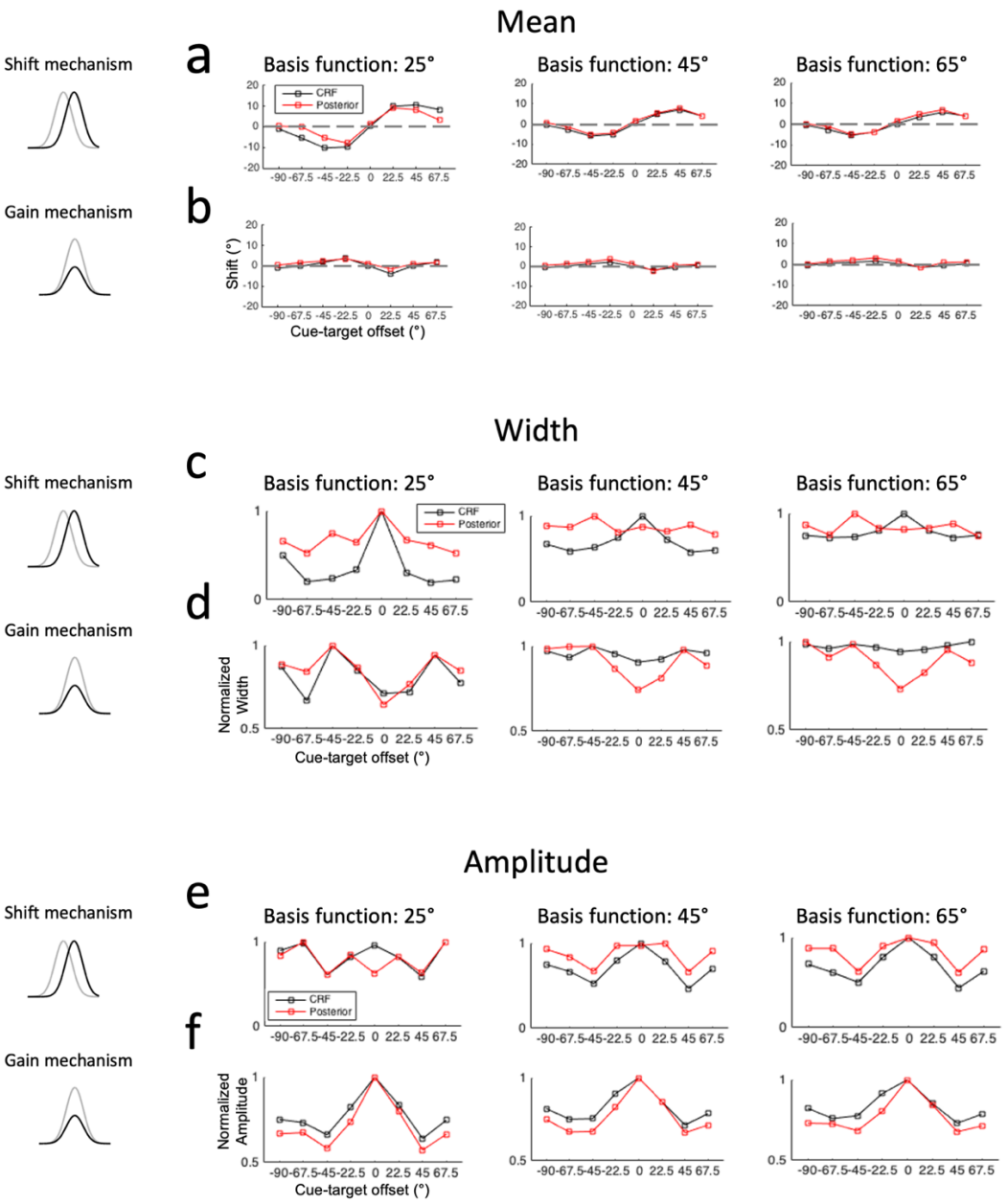


Figure 5. Further comparison between the shift and gain mechanism. (a) Shift mechanism – Shift in estimated means for each offset. Shift in estimated mean was plotted as a function of offset. The amount of shift was computed by subtracting the estimated orientation (i.e., mean of fitted von Mises function) from the actual stimulus orientation after fitting the CRF (black) and posterior probability distribution (red). Different panels represent results obtained using different channel basis function (i.e., width: 25°, 45° and 65° from left to right). (b) Shift in estimated mean for gain mechanism. Note, the Gain mechanism elicited a qualitatively different pattern from the shift mechanism as shown in (a). (c) Shift mechanism - Estimated width was normalized relative to the maximum value after fitting von Mises function at each offset to CRF (black) and posterior probability distribution (red). (d) Normalized widths for gain mechanism after fitting von Mises function to CRF (black) and posterior probability distribution (red). Left: channel basis function's width is 25°, middle: 45°, right: 65°. (e) Shift mechanism - Estimated amplitude was normalized relative to the maximum value after fitting von Mises function at each offset to CRF (black) and posterior probability distribution (red). (f) Normalized amplitude for gain mechanism.

Mean. For each offset condition, we computed the amount of shift by subtracting the estimated orientation (i.e., mean of the fitted circular Gaussian function) from the actual stimulus orientation for both CRF (Fig. 5a & 5b, black) and posterior probability distribution (Fig. 5a & 5b, red). Results were obtained using different channel basis functions (i.e., plotted in columns, channel width increased from left to right). Negative values indicate shifting toward smaller orientation value than original orientation and positive values indicate shifting toward larger orientation value than the original orientation. As shown in Fig. 5a, shifting mechanism resulted in a significant amount of repulsion in the estimated stimulus value such that smaller orientation than the attended one was shifted to even smaller value, and larger orientation shifted to larger value (cf. Fig. 3, ±45°). The results that were obtained using different channel basis functions showed similar patterns (different columns). The amount of repulsive distortion was also similar between CRF (Fig. 5a, black) and posterior probability distribution (Fig. 5a, red). Interestingly, the magnitude of shift was reduced as the basis function's width increased from 25° (Fig. 5a, left panel) to 65° (Fig. 5a, right panel).

For the gain mechanism (Fig. 4b), we observed a qualitatively different pattern in the estimated orientation. Compare to the significant deviation caused by the shifting mechanism, the majority of the estimated orientations under the gain modulation still overlapped with the actual, except those at ±22.5° offset (Fig. 5b), which showed a weak trend of attraction of the attended feature (cf., cyan curves in Fig. 4). The changes in mean also were less pronounced as the channel basis function became wider (Fig. 5b left to right panels). In short, we observed a qualitative different pattern in how different neuronal mechanism can modulate the orientation decoded from the population activities.

Width. We also found a difference in the width for reconstructed CRF and posterior probability distribution at different offsets. To better visualize and compare the patterns of width change, results shown in Fig. 5c & Fig. 5d were normalized relative to the maximum value (original unit in degrees). For the CRF (plotted in black), the shift mechanism led to an increase in estimated width of the reconstructed channel responses functions at the attended orientation (0° offset in Fig. 5c, also shown in Fig. 3), which was caused by attracting neurons in the neighboring zone toward the attended orientation. However, the gain mechanism showed the opposite pattern (Fig. 5d, black line), in which width was smaller at the attended orientation (i.e., 0° offset). Another observation is that the data pattern in CRF width was sensitive to the changes in the basis function, such that the difference between shift (Fig. 5c black) and gain mechanism (Fig. 5d, black) became less obvious as channel basis function became wider (from left panel to right panel in Fig. 5c & Fig. 5d).

The posterior probability function also showed qualitatively different patterns between shift mechanism (Fig. 5c, red) and gain mechanism (Fig. 5d, red). For tuning shift mechanism (Fig. 5c, red), the widths of reconstructed posterior probabilities were larger for the attended orientation (0° offset) and decreases as the offset became larger. However, for the gain mechanism (Fig. 5d, red), we found the opposite pattern. The width of the posterior probability displayed an inversed Mexican hat shape with smallest width at the attended orientation (0° offset). However, it's worth noting that the qualitative different pattern in width was only robustly observed under low noise (i.e., voxel variance), and can quickly become indistinguishable as noise increased (e.g., Fig. 12). Lastly, the distinction between shift and gain mechanism remained robust when the basis function changed (red in Fig. 5c & 5d, from left to

right panels), while the difference was much reduced with the CRF (black in Fig. 5c & 5d, left to right).

Summary

To summarize, we simulated two candidate neural mechanisms of surround suppression of FBA in a forward encoding model: a shift mechanism, or a gain mechanism. We then decoded their manifestations at population level using two multivariate methods: the standard IEM, and a Bayesian method. We found both multivariate methods showed that different neuronal mechanisms were associated with unique patterns at the population level in the vicinity of the suppressive surround. Importantly, the tuning shift mechanism elicited a distortion in the feature space, which manifested as a repulsion effect that shifted the nearby feature away from the attended one. The observed pattern was different for the gain mechanism, which manifested as reduced channel responses in the suppressive surround without a significant repulsion. This important distinction was robust across most combinations of neural tuning width and voxel noise. Within the suppressive surround, orientations (i.e., mean of the circular Gaussian template, Eq. 1) estimated from reconstructed CRF and posterior probability distribution both supported the repulsion effect that was elicited only by the shift mechanism, but not by the gain mechanism. We also found the estimated width parameter may also differentiate the two neuronal mechanisms when the voxel noise was low. Therefore, the simulation results suggested that it is possible to distinguish the neuronal mechanisms at the population level using both the standard IEM method and the Bayesian method.

CHAPTER 4

COMPARISON BETWEEN THE MULTIVARIATE METHODS

In the previous simulation, we showed that both methods decoded the signature patterns associated with different neural mechanisms for surround suppression in FBA. However, the linear regression nature of the standard IEM suggests that it is constrained by the initial assumption of the channel basis functions. In fact, the results described in the previous section also hinted that the IEM method may be more dependent on the channel basis function. In addition, a second advantage of considering the Bayesian method is that it was initially proposed to model the correlated noise structure among voxels. However, there is no systematic investigation on how noise correlation affects the multivariate methods, especially the standard inverted encoding model. Therefore, the second aim of the current work is to conduct a systematic comparison between the two multivariate methods in the presence of correlated voxel noises. In particular, we tested the impact of two key facets of these models: assumptions of the basis function and the amount and nature of the correlated noise on performance of the Bayesian method and IEM. For this purpose, we analyzed the simulated data using three sets of channel basis function (width: 25°, 45°, and 65°) as in previous chapter, and systematically varied the basis function and the structure of noise correlation to evaluate the two multivariate methods.

Method

Comparison Based on Stimulus Classification

Benchmark test of decoding. Stimulus classification is one of the most common tasks in in neuroimaging data analysis, which reveals stimulus information that is hidden underneath the seemingly "random" patterns of voxel responses. To further compare the standard inverted encoding method and the Bayesian method, we employed a classification task to compare them

under a neutral condition. The schematic of simulation was the same as in the previous chapter but for simplicity and without losing generality, here we only consider a neutral condition without attentional modulation. After fitting the encoding model to the training data set, we reconstructed the channel response function and posterior probability distribution from a validation data set, which had the same stimulus condition (i.e., 8 orientations, 32 trials each) as the training data set. We classified the stimulus label on each trial into one of eight possible stimulus categories (i.e., 8 possible stimuli: 0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°). We then computed classification accuracy across all trials. The final accuracy level is averaged across ten runs of simulation under each combination of parameters.

Classification with reconstructed channel response function (CRF). For the CRF method, we first generated predicted CRFs for eight possible orientations (i.e., 8 possible stimuli: 0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°). The eight predicted CRFs was then correlated individually with the reconstructed CRF on each trial. We computed the eight correlation coefficients on each trial and chose the maximum as the classified stimulus label for that trial (Brouwer & Heeger, 2009). Classification is correct if the classified stimulus label is the same as the true label.

Classification with posterior probability distribution. Classification using posterior probability distribution is more straightforward than the CRF, as it is a smooth distribution of probabilities for all stimuli (i.e., 0° to 179° at 1° increment) across the orientation space. Therefore, we computed the probabilities (within a ±5° range) at the eight possible orientation (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) and chose the largest one. Note, we also tried range smaller than ±5° and yielded highly similar results. Therefore, the exact range would not affect our interpretation. We repeated this for each trial and computed the averaged classification accuracy across 10 runs of simulation.

Manipulation on Correlated Voxel Noise

Importantly, the Bayesian method further modeled the difference source of the covariance among voxel response, while the standard inverted encoding method is based on the least-square fitting without assuming any correlated noise structure. We reasoned that the major difference between the two methods may reside in detecting the changes in the correlation structure and how they may be affected by the shape of channel basis function. Therefore, it is worthwhile to further examine how the correlation structure (cf. Eq. 11 & 12) and different channel basis function affects the decoding performance for both the standard inverted encoding method and the Bayesian method. Therefore, we manipulated the magnitude of voxel correlation (controlled by parameter r in Eq. 11) and different sources of the voxel correlation (i.e., p in Eq. 12, proportion of R^{tuning}), under different neuronal tuning width (i.e., 9 values from 25° to 65°).

Correlation magnitude. In the first scenario, we systematically varied the maximum correlation strength from 0.1 to 1 at a step-size of 0.1 (i.e., 10 levels in total), while fixing the ratio between tuning-dependent correlation R^{tuning} and tuning-independent correlation R^{arb} (2.5:1). Meanwhile, we also varied the neural tuning widths (i.e., from 25° to 65° at a step-size of 5°) as in the previous chapter.

Correlation ratio (R^{tuning}). In the second scenario, we fixed the maximum correlation magnitude (at 0.4) but varied the proportion of tuning-dependent correlation (R^{tuning} in Eq. 11 & 12) from 0 to 1 using a step-size of 0.1. The neural tuning widths were also sampled from 25° to 65° with a step-size of 5°. When either the correlation magnitude (i.e., r, Eq. 11) or the proportion of R^{tuning} (i.e., p, Eq. 12) was manipulated, we fixed the voxel standard deviation (i.e., $\lambda = 0.15$, Eq. 10). In other words, the diagonal term of the covariance matrix is constant, while we systematically manipulated the off-diagonal covariance among voxels.

Varying both correlation magnitude and correlation ratio together. In the last scenario, we varied both the magnitude and proportion of R^{tuning} together, which could also have created a much larger parameter space than other scenarios. As the purpose was to explore the interaction between correlation magnitude and ratio, we further constrained both the neuronal tuning width (bandwidth: 40°), and voxel standard deviation (i.e., $\lambda = 0.15$, Eq. 10).

Note, we only varied these noise parameters for the benchmark test but kept them fixed when we simulated and compared between different neuronal mechanisms in the previous chapter for simplicity.

Channel basis function. In each of the three scenarios above, we performed the Benchmark decoding test with three sets of channel basis function (width: 25°, 45°, and 65°) to explore how different basis functions affects decoding performance of different methods.

Results

Benchmark Test – Stimulus Classification

As both the Bayesian method and IEM can distinguish the neuronal mechanisms of FBA to a similar extend (see previous chapter), we further explored their difference in decoding ability with a benchmark test in stimulus classification. We systematically varied the covariance (i.e., off-diagonal terms) of the covariance matrix of voxel responses, while fixing the diagonal terms (i.e., τ, voxel variance) and varied the width of channel basis function. We first measured the classification accuracy in the benchmark test as a function of the changes in the correlation structure. We then computed the difference in classification accuracy between the Bayesian method and the standard inverted encoding method (i.e., Bayesian minus IEM). The benchmark tests were repeated with three sets of channel basis function (width: 25°, 45°, & 65°).

Correlation magnitude. We first varied the maximum correlation strength from 0.1 to 1 while fixing the ratio between R^{tuning} and R^{arb} (2.5:1). Fig. 6b shows the classification accuracy as a function of overall correlation strength between voxels (controlled by parameter r in Eq. 11) under different combination of neuronal tuning width (i.e., 25° to 65°) and voxel variance (λ = 0.15, Eq. 10). Both the Bayesian method (Fig. 6b, solid line) and standard inverted encoding method (Fig. 6b, right panel) showed a drop in the decoding accuracy as the overall correlation magnitude increased across different values of neuronal tuning width. The classification accuracy of the Bayesian method remained higher than the standard IEM method (Fig. 6a), which was mainly attributed to conditions with wide neural tuning curve (e.g., 45° to 65°, light green to yellow). However, when the neural tuning widths were close to the channel width, the difference in decoding accuracies became close to 0 (Fig. 6a, dark green). This indicated that IEM's decoding accuracy only reached similar level to the Bayesian method when neural tuning width and channel width were similar.

This pattern was also observed when the channel width was varied (e.g., 45° and 65°). As shown in Fig. 6c (channel basis function: 45°) and Fig. 6e (channel basis function: 65°), Bayesian method outperformed the IEM method in most conditions. Yet the most advantageous conditions occurred when neural tuning is narrowest (i.e., 25°), which is opposite to what was seen in Fig. 6a. This further suggested that IEM method's decoding performance was sensitive to the match between neural tuning width and channel width.

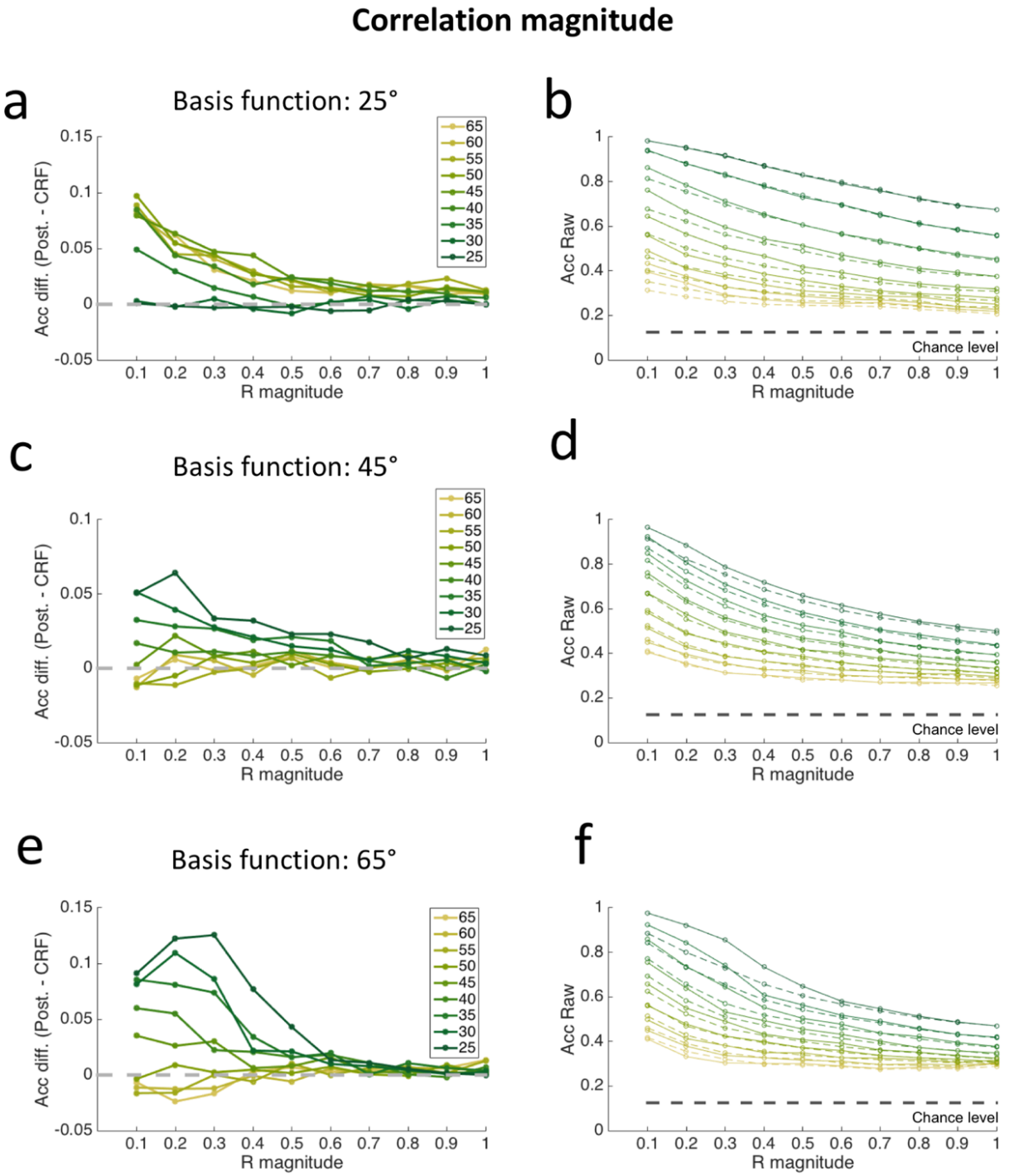


Figure 6. Benchmark test - correlation magnitude. (a), (c), & (e) Difference in classification accuracy (Bayesian method – CRF method) was plotted as a function of maximum correlation strength (R magnitude). Colors represent different neuronal tuning widths. Width of channel basis function: (a) – 25°, (c) – 45°, (e) – 65°. (b), (d), (f) raw classification accuracy using Bayesian method (solid lines) and CRF method (dashed lines). Channel basis function: (b) – 25°, (d) – 45°, (f) – 65°. The dashed lines in (b), (d), & (f) represent chance level (0.125) performance in the 8-way classification task.

Correlation ratio (R^{tuning}). The proportion of R^{tuning} within the overall correlation was manipulated while fixing the maximum correlation strength constant (i.e., maximum correlation: 0.4). As shown in the right column of Fig. 7, both classification methods became less accurate as the voxel noise changes from arbitrary to tuning-dependent correlation. However, the Bayesian method had an overall higher classification accuracy than the IEM method for the majority of data points. The advantage was also seen all three sets of channel basis function (Fig. 7a – 25°, Fig. 7c – 45°, Fig. 7e – 65°). Furthermore, we observed a similar pattern between neural width and channel width for IEM as described in the previous section. When the channel width was narrow (Fig. 7a – 25°), IEM performed worse than Bayesian method under broad neural tuning width (Fig. 7a, light green to yellow). As the channel width increased (Fig. 7c – 45°, and Fig. 7e – 65°), IEM performed worse than Bayesian method mostly for narrow neural tuning width (Fig. 7c & 7e, dark greens). In other words, the IEM method was comparable to the Bayesian method only when the neural tuning width and channel width matched.

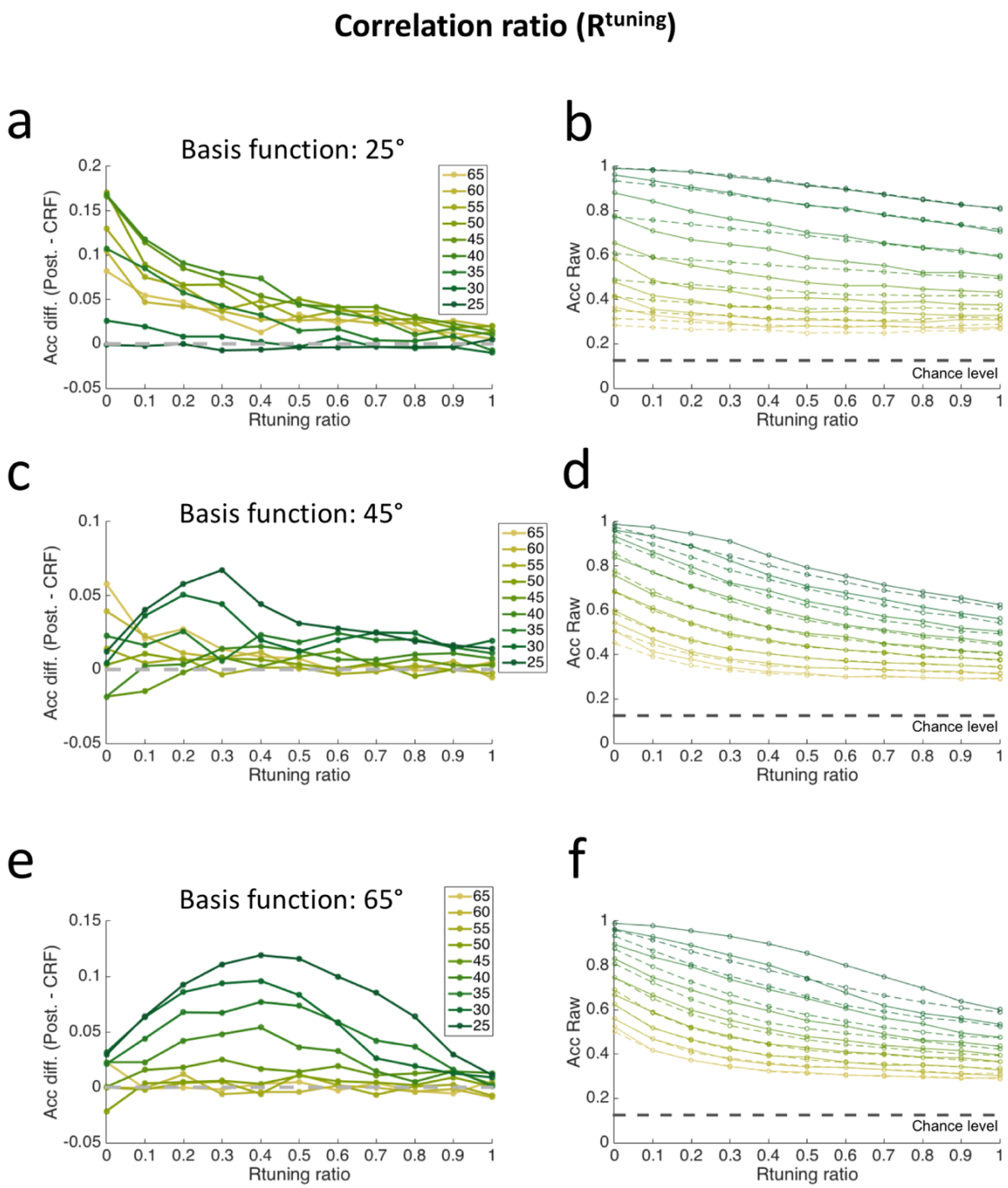


Figure 7. Benchmark test - correlation ratio. (a), (c), & (e) Difference in classification accuracy (Bayesian method – CRF method) was plotted as a function of ratio between tuning dependent and tuning-independent noise (1 meaning completely tuning-dependent noise). Colors represent different neuronal tuning widths. Width of channel basis function: (a) – 25°, (c) – 45°, (e) – 65°. (b), (d), (f) raw classification accuracy using Bayesian method (solid lines) and CRF method (dashed lines). Channel basis function: (b) – 25°, (d) – 45°, (f) – 65°. The dashed lines in (b), (d), & (f) represent chance level (0.125) performance in the 8-way classification task.

Varying both correlation magnitude and correlation ratio together. In the last scenario, we further explored the parameter space by varying both the magnitude and ratio of correlation component together, when neuronal tuning width (i.e., 40°) and voxel variance (λ = 0.15 in Eq. 10) were fixed. Consistent with findings above, we found that a larger correlation magnitude tends to cause a reduction in the overall classification accuracy for both methods (i.e., different colors in the right column of Fig. 8, solid – Bayesian method, dashed – standard IEM). Meanwhile, as the correlation became more tuning-dependent, we found a reduction in classification accuracy (i.e., along the x axis of Fig. 8b, 8d, & 8f). Similar to the findings above, the Bayesian method again outperformed the standard inverted encoding method for the majority of the data points in the parameter space when the neural tuning width was different from the channel width (Fig. 8a: neural – 40°, channel – 25°, Fig. 8e: neural – 40°, channel – 65°). IEM performed to a similar level as the Bayesian method when the neural tuning width was similar the channel width (Fig. 8c: neural – 40°, channel – 45°).

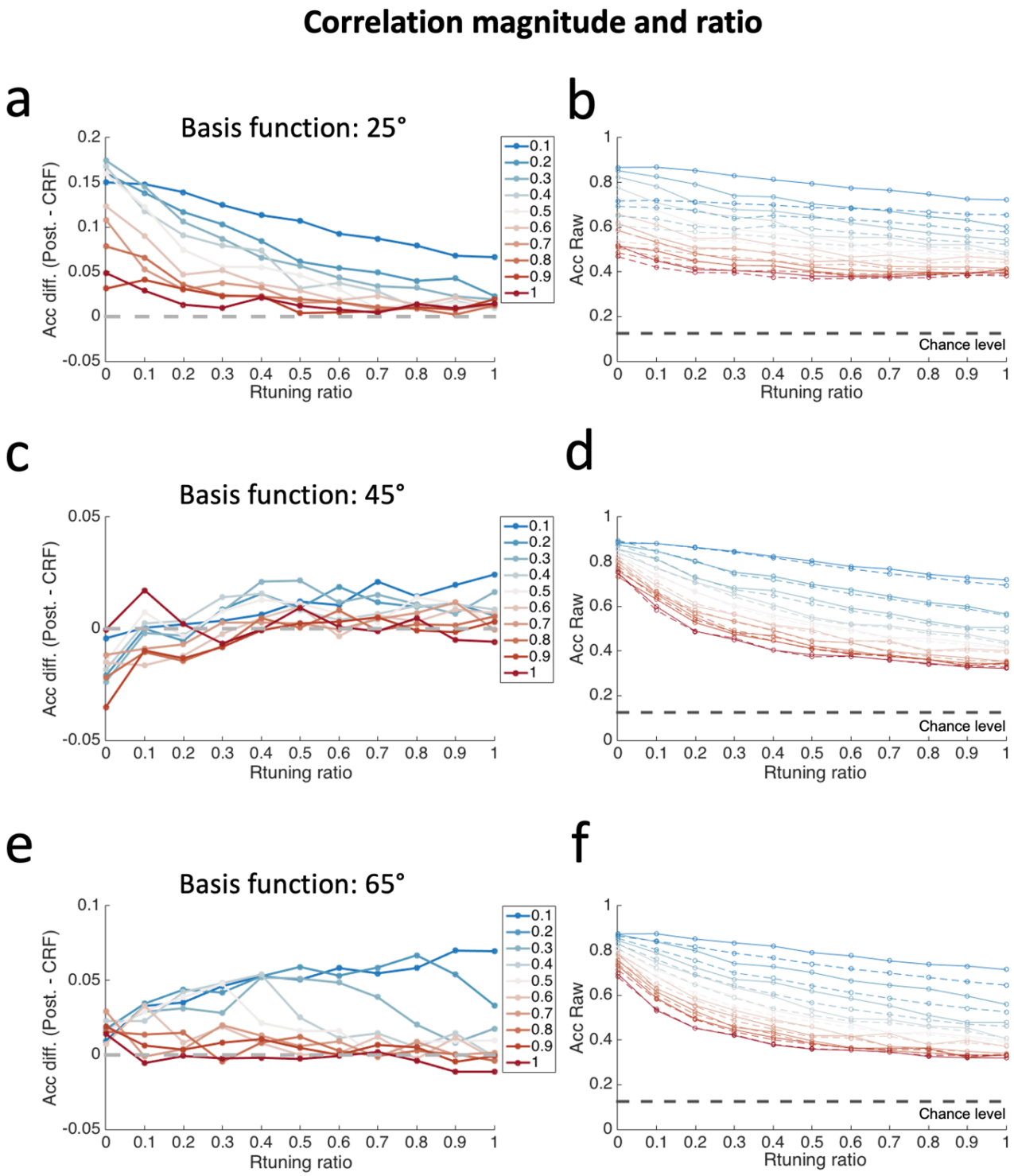


Figure 8. Benchmark test - correlation magnitude and ratio. (a), (c), & (e) Difference in classification accuracy (Bayesian method – CRF method) was plotted as a function of ratio between tuning dependent and tuning-independent noise (1 meaning completely tuning-dependent noise). Colors represent different maximum correlation strength (R max). Neural tuning width was fixed at 40°. Width of channel basis function: (a) – 25°, (c) – 45°, (e) – 65°. (b), (d), (f) raw classification accuracy using Bayesian method (solid lines) and CRF method (dashed lines). Channel basis function: (b) – 25°, (d) – 45°, (f) – 65°. The dashed lines in (b), (d), & (f) represent chance level (0.125) performance in the 8-way classification task.

Summary

In this simulation, we systematically compared the performances of both multivariate methods (the Bayesian method and IEM) in a classification benchmark test by varying the structure of noise correlation (correlation magnitude, ratio between tuning-dependent and tuning-independent correlation, or both) and the basis function (25°, 45°, 65°). While both methods' performance dropped as the noise become more correlated, or more tuning-dependent, the results evidently showed that the Bayesian method performed better than IEM across vast majority of all data points when the parameter space of noise correlation was examined. Through varying the basis function (e.g., 25°, 45°, 65°), the results further showed that IEM's performance became much worse when there was a mismatch between the channel basis function and neural tuning width. Such findings further demonstrate that regression-based IEM is constrained by initial assumption on its channel basis function.

CHAPTER 5

GENERAL DISCUSSION

We employed an encoding/decoding model to explore the neuronal mechanism of the attentional surround suppression in the feature domain. We fist constructed a generative model to simulate neural responses with two alternative neuronal mechanisms (tuning shift or gain) underlying the attention surround suppression. We then decoded such attentional modulation on simulated fMRI voxel responses using the standard inverted encoding method and a Bayesian method. Our results revealed that each neuronal mechanism produced its own signature pattern at the population level. This result can serve as a prior prediction for further empirical studies to adjudicate between different neural mechanisms of feature-based attention. Furthermore, while both methods can differentiate the neural mechanisms, we found that the Bayesian method is more robust than the standard inverted encoding method in the presence of correlated noise.

Distinguishing Neural Mechanisms of Surround Suppression in FBA

Both single-unit electrophysiological method and neuroimaging method are critical in investigating the neural mechanisms of FBA. Therefore, it is important to bridge attentional mechanisms measured across different levels. At single-unit level, FBA can elicit either a neuronal tuning shift or a multiplicative change of neuronal responsivity, which may equally explain the non-monotonic surround suppression effect in behavior (Fang et al., 2019; Tsotsos, 2011). However, such neuronal-level mechanism is likely hidden at the fMRI voxel level using traditional univariate analysis or a pure decoding approach (e.g., multivariate pattern analysis, MPVA), which are powerful in detecting differences in activation patterns across condition while being agnostic to how the differences are created. The strength of the current approach is that we explicitly coded the underlying neuronal mechanism of surround suppression into voxel

responses using a generative model (Fig. 2c & 2d). From the voxels with known neuronal modulation, we then decoded the population response profile to establish a direct link between the neuronal mechanisms and their population measures (Fig. 2f). Importantly, each neuronal mechanism was shown to have its own signature pattern in population response (Fig. 3 - 4). By examining the population response profile within the suppressive surround, our simulation may shed further light on how different neuronal mechanisms can explain the non-monotonic effect of FBA.

Tuning Shift Mechanism

One way that FBA can elicit surround suppression is by shifting neurons within the suppressive surround toward the attended feature. Our simulation showed that this inward shift of neuronal tuning was transformed into an outward shift at the population level (Fig. 3), which can repulse similar but task-irrelevant features further away from the attended feature. In the spatial domain, the RF shift was suggested to increase the perceived distance between the attentional focus and nearby location – termed the attentional repulsion effect (Suzuki, & Cavanaugh, 1997). Other researcher further suggested a linking hypothesis between such distortion in physical space and enhanced spatial resolution by attention (Anton-Erxleben and Carrasco, 2013).

The surround suppression effect of FBA may also employ a similar repulsion effect to enhance feature resolution. For example, nearby tuning curves (e.g., ±45° offset) that are centered in the neighborhood of the attentional focus can be attracted such that they can also be activated by orientations near the attended one. However, the labels of the affected tuning curves still represent the original orientation, which creates a repulsion in the perceived orientation away from the attentional focus. This could be equivalent as physically moving the nearby

distractors away from an attended orientation, which reduces interference. In addition, as attention attracts tuning curves toward attended orientation, such shift would also cause some part of the nearby orientation space to be under-represented. This is because the attentional shift also modified the neighboring neuron's preferred orientation toward the feature in focus, resulting in a suboptimal response to the original orientations that they code for. Such weakened neural responses can be used to suppress distractors in the vicinity of the attended orientation.

Gain Mechanism

Gain modulation is another way that FBA may elicit the surround suppression effect. Both feature-based attention and spatial attention can modulate perceptual sensitivity (e.g., measured as d') to luminance contrast through either a response gain or a contrast gain (spatial attention: Herrmann et al., 2010; FBA: Herrmann, Heeger & Carrasco, 2012), which has been well captured in the highly influential normalization model of attention (Reynolds & Heegers, 2009). At the neuronal level, the gain modulation can be implemented as a multiplicative factor applied to neuronal tuning curve, which modulates the overall amplitude without changing their tuning preference (spatial attention: Reynolds et al., 2000; McAdam & Maunsell, 1999; feature-based attention: Treue & Martinez-Trujillo, 1999; McAdam & Maunsell, 2000). We simulated a non-monotonic gain modulation across the neuronal population with a difference of Gaussian profile, which consisted of an excitatory component and a suppressive component. A recent single-unit study indicated that the suppressive Gaussian component may be explained by a tuned normalization pool that is modulated by FBA (Yoo, Martinez-Trujillo, et al., 2021). Consistent with the multiplicative modulation on neurons, our simulation also showed a reduction of the overall response at population level within the suppressive surround (Fig. 4).

Interestingly, the multiplicative gain modulation that attention exerts at the neuronal level produces a similar effect as changing the effective contrast of an attended stimulus (Reynolds et al., 2000; Martinez-Trujillo & Treue, 2002). Recently, researchers evaluated how attention changes the perceived intensity due to attention (Carrasco, Ling & Read, 2004; Liu, Fuller, & Carrasco, 2006; Liu, Abrams & Carrasco, 2007). The findings were consistent with an altered appearance of the attended stimulus, as if it had a higher contrast (Carrasco, Ling & Read, 2004, Liu, Abrams & Carrasco, 2009) or more coherent motion (Liu, Fuller, & Carrasco, 2006), which was also accompanied by an enhanced processing in early visual areas (e.g., neural correlates for altered contrast, Liu, Pestilli, & Carrasco, 2005, Dugué et al., 2020). It is possible that surround suppression may also prevent distractors' interference by reducing their effective contrast and perceptual salience. Therefore, an interesting direction for future studies could be to investigate how FBA modulates the appearance of stimulus feature within the suppressive surround of an attended feature.

Despite the absence of direct neural evidence for a non-monotonic gain mechanism, computational models suggested that FBA may directly exert gain modulation in sensory visual areas to suppress similar but different features in the surround of an attended feature. For example, this possibility is supported by the selective tuning model (see Candidate Neuronal Mechanisms of Surround Suppression Section in Chapter 2), which assumes a top-down feedback modulation that progresses backward along the visual hierarchy and directly inhibits units less tuned to the attended one in earlier layers (Tsotsos, 1995, 2011). Alternative to the direct suppressive gain modulation predicted by the selective tuning model, it is also possible that the top-down feedback modulation may first modulate excitatory cells, which indirectly implement surround suppression through lateral inhibition.

Furthermore, the tuning shift mechanism and the gain mechanism may not be mutually exclusive in feature-based attention. A recent study using a two-layer feedforward model showed that a multiplicative gain modulation in low-level regions can lead to tuning shift in higher regions through linear integration (Ibos & Freedman, 2014). Therefore, it is possible that the tuning shift mechanism may build up progressively in downstream regions of the gain modulation. In principle, our modeling framework could be extended to simulate a multi-layer network, which will allow us to further explore in future studies whether the two mechanisms modulate different stages of the visual processing hierarchy.

Source of Surround Suppression in FBA

While the shift or gain mechanism could underlie the surround suppression's modulation in sensory regions, they likely rely on top-down feedback modulation originated from attentional control areas. Bartsch and colleagues recently explored this hypothesis by measuring the temporal dynamic of surround suppression in feature-based attention using MEG (Bartsch et al., 2017). To manipulate feature-based attention, they used a 2-alternative-forced-choice task, in which participants reported the location of a red target against a green distractor on one side of screen. To measure the profile of FBA, a probe stimulus was presented in the opposite hemifield of the target stimuli, whose color was systematically sampled away from the attended red color. The authors found that the FBA first exhibited a coarse selection profile of the attended red color (205 ~ 275 ms) in anterior ventral extrastriate areas (areas anterior to VO and lateral to PHC). Following this initial coarse selection (after ~100 ms), there was a suppression of colors near the attended red color, which suggest the emergence of surround suppression. In addition, source localization analysis revealed that this refinement of attentional profile occurred in more

posterior retinotopic visual areas (VO-1/hV4). Taken together, these findings support a role of top-down modulation in FBA's surround suppression.

At a larger scale, recent work further suggests that the frontoparietal network (FPN) is ultimately responsible for the top-down control during feature-based attention (for reviews, see Scolari, Seidl-Rathkopf, & Kastner, 2015, Liu et al., 2019). For example, a recent study showed that FPN population activity is correlated with behavioral performance in a feature-based attention task and disrupting this network (e.g., with transcranial magnetic stimulation, TMS) impaired behavioral performance, hence suggesting the causal role of FPN in determining feature-based attentional modulation (Jigo, Gong, & Liu, 2018). In sum, current evidence suggests that top-down feedback is necessary in eliciting the suppressive surround in both space-based (Boehler et al., 2009, 2011) and feature-based attention (Bartsch et al., 2017). Given the critical role of frontoparietal network in attentional control, it is possible that the FPN also operates as the source region that generates surround suppression. For example, it may either change the gain of visual cortical neurons or shift their tuning. Therefore, future studies may further investigate the relationship between FPN and the non-monotonic surround suppression modulation of FBA.

A Priori Modeling Framework for Future Empirical Studies

So far, we have discussed that the current approach can be suitable for unifying neural mechanisms of surround suppression across different levels of measurements. To further examine whether our simulated results represent a robust effect between different neuronal mechanisms, we simulated under a wide range of parameter combinations including tuning width and noise level. Importantly, to provide an equal footing for comparing the different neuronal mechanisms, we used the same neutral dataset to train different models but test the model on

attentional data generated by different neuronal mechanisms. Therefore, the only difference between attentional conditions is the underlying neuronal mechanisms. Such an analytic scheme provides an unbiased way to distinguish different mechanisms without introducing spurious results due to overlap between training and testing data (Sprague, Boynton, & Serences, 2019). Under the current modeling framework, we established a linking hypothesis between candidate attentional modulations at the neuronal level and their manifestation at the fMRI voxel level.

Bridging neural mechanisms measured at different levels is a nontrivial endeavor. For example, it is worth noting that the repulsion effect at the population level is in the opposite direction of the neuronal shift, which was toward the attended feature. This observation suggests a disconnect between single neurons' behavior and their collective behavior at the population level, which is not unique to the repulsion effect. For instance, spatial attention has been found to modulate the neuronal responses to contrast. While single-unit studies typically found a contrast gain or a response gain modulation depending on the relative size of attentional field and stimulus (Reynolds & Heeger, 2009), neuroimaging studies more often report an additive improvement (e.g., vertical shift) in the contrast response function (Buracas and Boynton, 2007; Li et al., 2008; Murray, 2008; Pestilli et al., 2011). Simulation further showed that such finding can be well explained by the normalization model assuming different balance between the attentional field and stimulus size encountered by a neuronal population, which resembles a combination of contrast gain and response gain modulation across the entire population (Hara et al., 2014). Therefore, a forward simulation combined with appropriate decoding method is necessary to relate neuronal level and population level phenomena.

In addition, it is recently suggested that directly inferring the underlying neural mechanisms using the inverted encoding model may be inappropriate as the signal-to-noise ratio

of the fitted model can also change the property of the reconstructed channels (Liu et al., 2018). When there is a limited number of candidate mechanisms, a forward simulation approach like ours may provide a grounded solution to this problem by building a direct link between the neuronal mechanism and its population pattern. When the simulated population response showed qualitative difference as the current results demonstrate, one may use the simulation as a priori prediction for guiding empirical work and interpreting the findings. Admittedly, there may not be a simple solution to unequivocally assay the neural mechanism across different level of measurements. However, for early visual processing that are well studied and can be reasonably constrained with physiological knowledge, the encoding/decoding approach may help researchers to better understand the model behavior after inverting the encoding model and to avoid misinterpreting the changes in reconstructed channels by conducting simulation under different parameter combinations. Taken together, the current computational modeling and simulation may also serve as a general framework and reference point for interpreting empirical findings in future studies on the neural mechanism of FBA.

Comparison Between the Multivariate Methods

The encoding model approach has been widely employed to generate a functional description of a brain area and make quantitative prediction of voxel response. However, inverting the encoding model only leads to reconstruction of initial model assumption, stimulating debates in whether reconstructed channel response can represent the population response (Gardner & Liu, 2019, also see Sprague et al., 2019). The problem at its core is that the inverting approach only reconstructed the intermediate step (Gardner & Liu, 2019), which is different from previous applications that further perform stimulus identification (Kay et al., 2008) or reconstruction (Brouwer & Heeger, 2009). It is suggested that the Bayesian technique

developed by van Bergen and colleagues (van Bergen et al., 2015) could be more suitable to explore underlying neural mechanisms as it aims to recover information on stimulus instead of initial assumption of the encoding model (Liu et la., 2019; Gardner & Liu, 2019).

Therefore, we compared the standard inverted encoding model with the Bayesian decoding technique that further transforms the reconstructed channel response function into the probability distribution of the stimulus given the voxel response patterns. As reviewed in Chapter 2, IEM method has been fruitfully used to characterize neural mechanisms of higher order cognitive function at sub-voxel level, including perceptual distortion effect (e.g., shift of reconstructed population response profile) caused by categorical learning (Ester et al., 2016). The Bayesian method produces a full probability distribution of all the possible stimuli given a particular neural response, which in principle contains stimulus information rather than merely producing a point estimation of the most likely stimulus. Indeed, our simulation demonstrated that the Bayesian technique was also suitable for exploring the underlying mechanisms of complex cognitive functions, such as differentiating the shift vs. gain mechanisms in FBA. Moreover, for a benchmark stimulus classification task, the Bayesian method provides more reliable results when the channel basis function was varied (Fig. 6 - 8). In fact, the standard IEM method only performed to a similar level as the Bayesian method when there was a match between the channel basis function and neural tuning curve (i.e., similar width). Given that the neural tuning width is variable in the human brain, it may impose a greater challenge for the standard IEM than the Bayesian method.

Another important property of the Bayesian method is the explicit modeling of the noise structure among voxels, which are inherently correlated. Voxel-wise correlation, likely a result of neuronal correlations (Averbeck, Latham, & Pouget, 2006; Cohen & Kohn, 2011; Kohn,

Coen-Cagli, Kanitscheider, & Pouget, 2016), may have a detrimental effect on the accuracy of neural representation of stimulus, while only a handful of studies have recently begun to characterize its impact on the population activity at the fMRI BOLD level (van Bergen et al., 2015; van Bergen & Jehee, 2018). In the benchmark classification task, we systematically varied the covariance among voxels and found that the Bayesian method outperforms the standard inverted encoding method. Such advantage of the Bayesian method is likely due to the fact that it attempts to capture the correlation structure of the noise, while the standard inverted encoding method does not explicitly model the structure of the noise. Indeed, it assumes all the voxels are independent. Our results thus imply that when noise correlation is not extremely high or the correlation structure changes among experimental conditions, the Bayesian method produces superior results than the inverted encoding method.

Conclusions

While the feature-similarity gain model remains one of the most influential models of attention, recent studies have revealed non-monotonic effect in behavior that it cannot account for. At a coarse level, the feature-similarity gain predicts a suppression for dissimilar features, which is consistent with behavioral, neural imaging and single-unit studies. However, on a finer scale, it fails to explain how FBA exclude similar but different distractors to an attended feature. In recent years, an increasing number of studies showed that FBA can elicit a non-monotonic surround suppression, which enhances the signal-to-noise ratio in the vicinity of an attended feature. In fact, both the surround suppression and the feature-similarity gain modulation may be at work but on different similarity scale to enhance the most relevant aspect of the sensory input at the expense of unattended information.

The first aim of the current work was to investigate candidate neural mechanisms underlying the non-monotonic profile of FBA through simulation and computational modeling. The attentional template stored in working memory may exert a top-down modulation in eliciting the suppressive zone. One possibility is that the top-down feedback signal can shift the tuning preference of sensory neurons toward the attended feature to further enhance target representation (i.e., matched filter). Alternatively, top-down feedback may operate via a multiplicative gain mechanism without changing other properties of neuronal tuning. Interestingly, previous studies suggested different linking hypotheses between the candidate neural mechanisms and perceptual differences at the appearance level. Therefore, one interesting direction for future studies is to further test how the surround suppression may affect the perceptual appearance of stimulus in behavioral studies (e.g., contrast change or distortion in feature space).

Our simulation demonstrated that the candidate neural mechanisms can be distinguished at the fMRI voxel level using non-invasive neuroimaging method. This is made possible with the most recent developments in neural decoding technique in computational neuroimaging – a inverted encoding model technique that reconstructs population-level response profile and a Bayesian technique that further transform the population-level response into probability distribution in the feature space. Both methods decoded signature patterns associated with different candidate neural mechanisms. Therefore, it is possible to use the findings in the current simulation as a priori predictions for future studies and further examine the candidate mechanisms of surround suppression in the human brain.

Furthermore, our simulation work may provide a modeling framework for empirical studies using non-invasive methods like fMRI. The encoding/decoding approach in the current

simulation work can help bridge the gap in neural mechanisms across different levels of measurements, which provides a solution to the reverse-inference issue often found in model-based analysis. Specifically, one can implement the neural mechanism in a forward simulation, and then decode it at a different level (e.g., voxel level). In this way, our current work should contribute to the general effort in better understanding the underlying neural mechanisms of cognitive functions.

Lastly, the current work further revealed advantages of the Bayesian technique over the IEM method in the presence of correlated voxel noise. First, the Bayesian method captures the correlated noise structure, while the IEM does not. This is important given that neural noises are intrinsically correlated, which can greatly impact neural representations. Second, the Bayesian method aims to reconstruct stimulus information (i.e., probability distribution), while the IEM aims to recover model assumption. As our results suggested, the Bayesian method provides more accurate estimation of stimulus and is less influenced by initial model assumptions (e.g., channel basis function). Therefore, these new findings may provide further guidance for future empirical studies when considering different decoding methods.

Table 1. List of variables in the model simulation

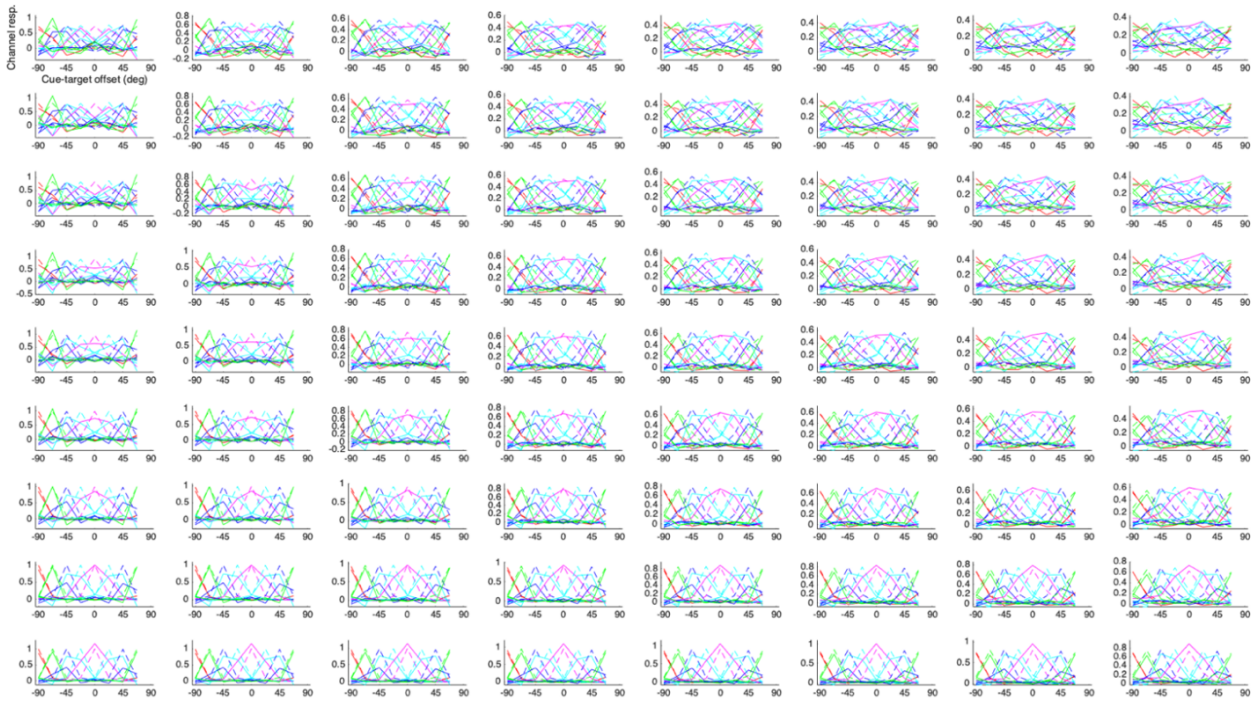
Parameter	**Description**
von Mises function (neuronal tuning curve)	
s	Stimulus orientation in degrees
$f_t(s)$	Orientation tuning curve of t-th neuronal population
μ_t	Tuning preference of t-th neuronal population
κ	Concentration parameter controlling neuronal tuning width
a, b	Amplitude and baseline for circular Gaussian (von Mises) function
Attentional modulation at the neuronal level	
μ_{att}	Attended orientation stimulus (90°)
G_t	Gain modulation for t-th neuronal population
A_1, w_1, A_2, w_2, L	Parameters controlling the overall shape of Difference of Gaussian function, which simulates the non-monotonic gain modulation

Table 1 (Cont'd)

α, β	Parameters controlling the linear feature-similarity gain modulation
SS_{range}	Surround suppression range (±45°)
fMRI Voxel response	
$v_i(s)$	i-th voxel's tuning curve
W^{neuron}	Linear weights combining neuronal response into voxel response
e	Simulated voxel noise draw from a multivariate distribution
Voxel-wise noise correlation	
Σ	Covariance matrix of simulated voxel noise
τ	Standard deviation of voxel response
λ	Proportion of voxel standard deviation relative to mean voxel response
R^{tuning}, R^{arb}	Tuning-dependent and tuning-independent noise correlation among voxels
p	Proportion of R^{tuning} in the overall correlation among voxels
r	Parameter scaling the maximum correlation strength
Decoding (IEM & Bayesian method)	
B_{train}, B_{test}	Training data set of voxel responses and testing data set of voxel responses
$\widehat{W}$	Estimated channel weights using training data set using standard IEM
C_{train}	Predicted channel response for training data set
Ω	Estimated covariance matrix from training data set using Bayesian method
$\hat{C}_{test}$	Reconstructed channel response function using test data set using standard IEM

APPENDIX

a



b

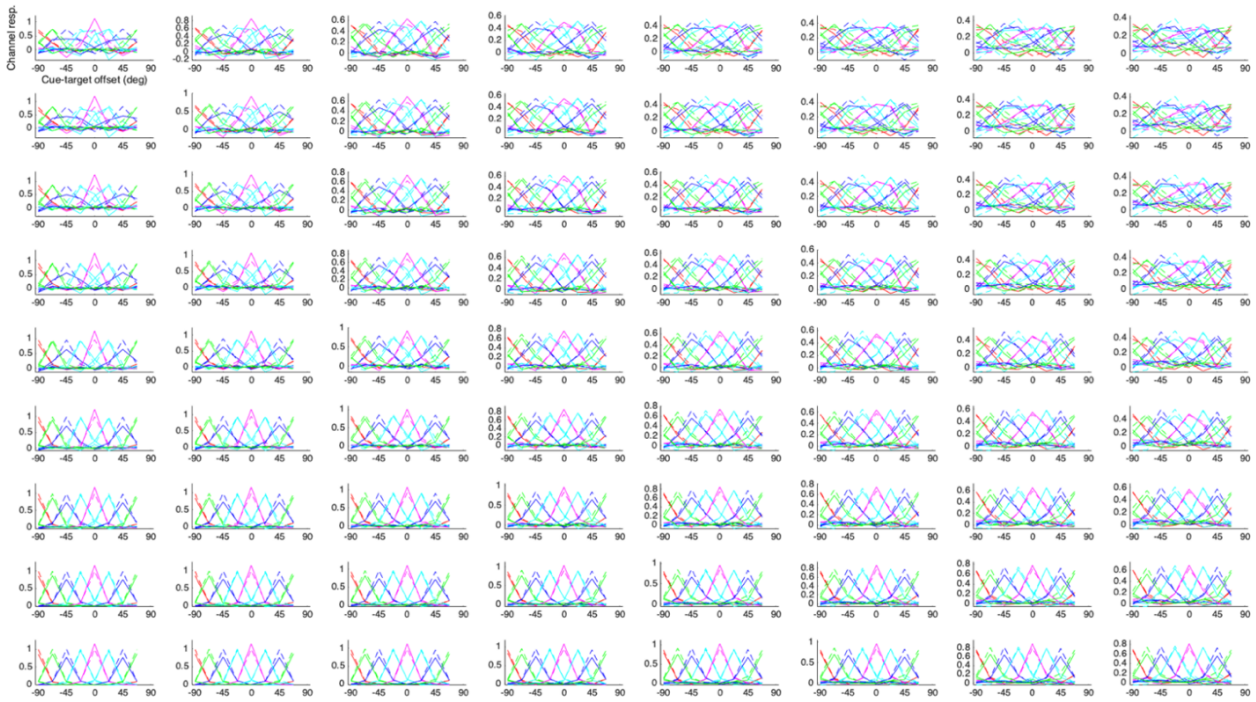
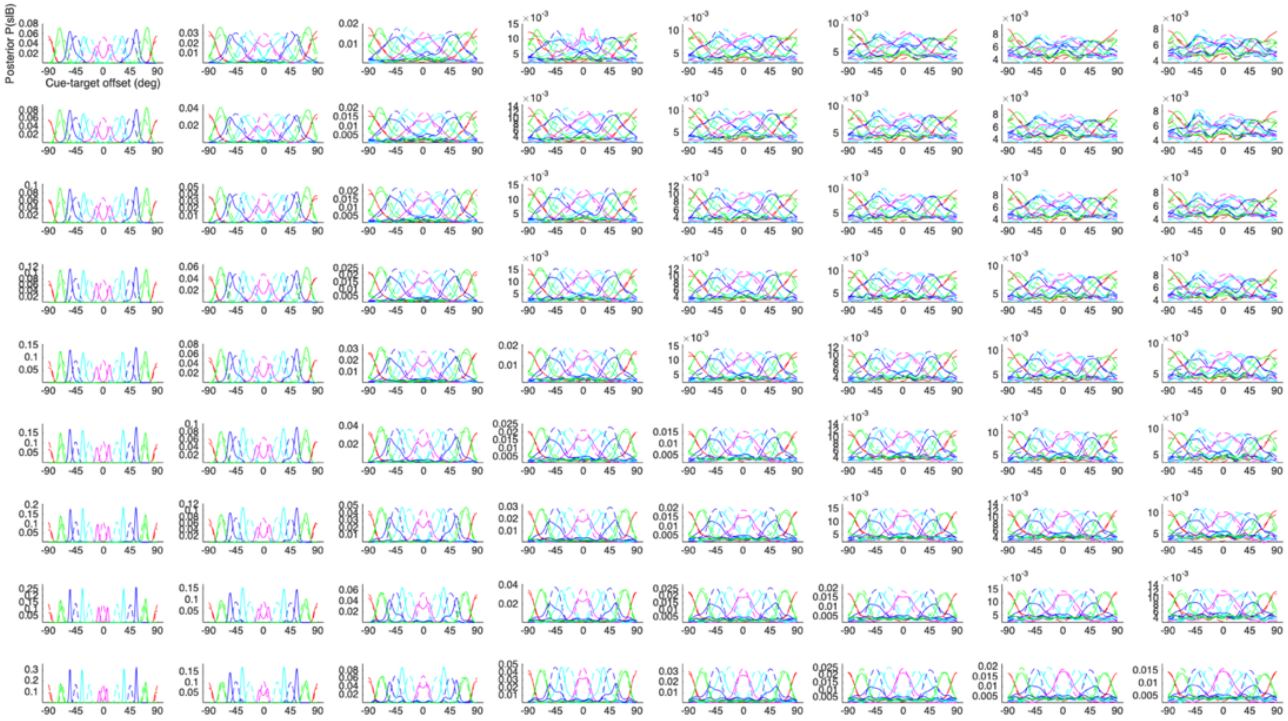


Figure 9. Full results for reconstructed CRF (Channel basis function: 25°). Color convention is the same as in Fig. 3. (a) Results for shift mechanism. (b) Results for gain mechanism. Within each panel, CRF was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Panels are shown on a 9 by 8 grid. Rows represents different neuronal tuning width parameters (25°, 30°, 35°, 40°, 45°, 50°, 55°, 60°, 65°), and columns represents different voxel variances (λ: 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, & 35%). See Method section for details.

a



b

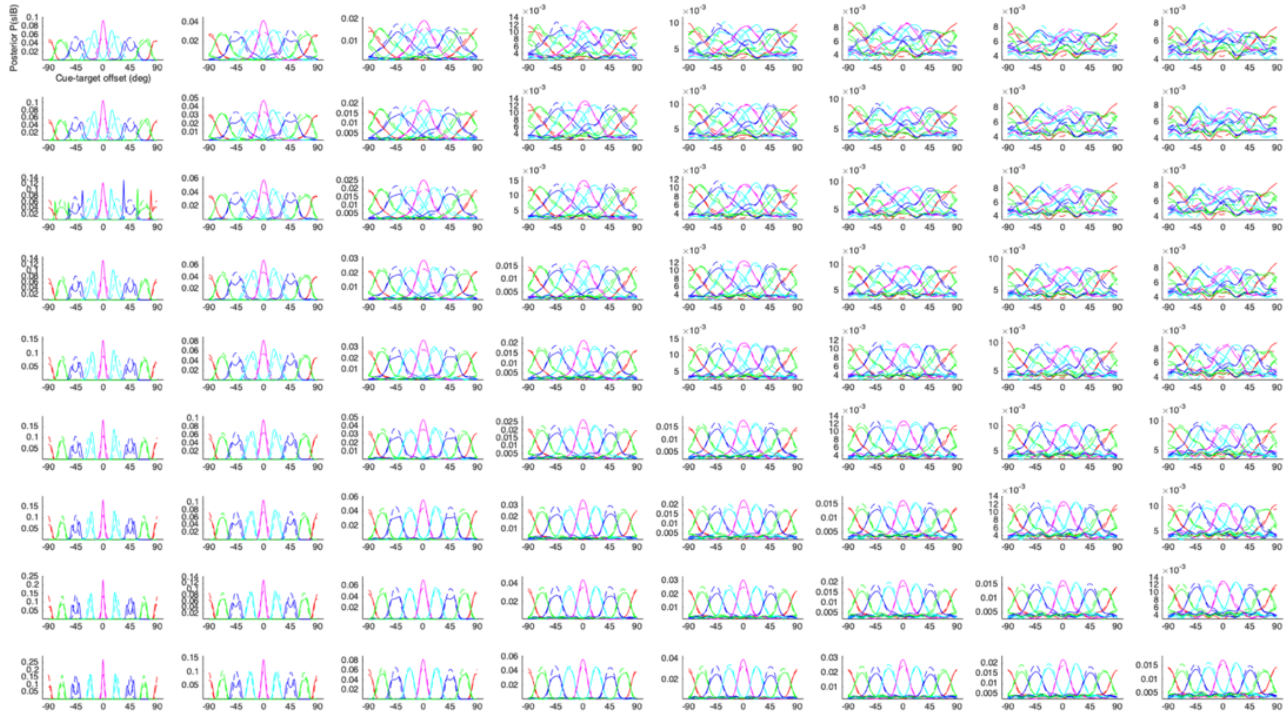
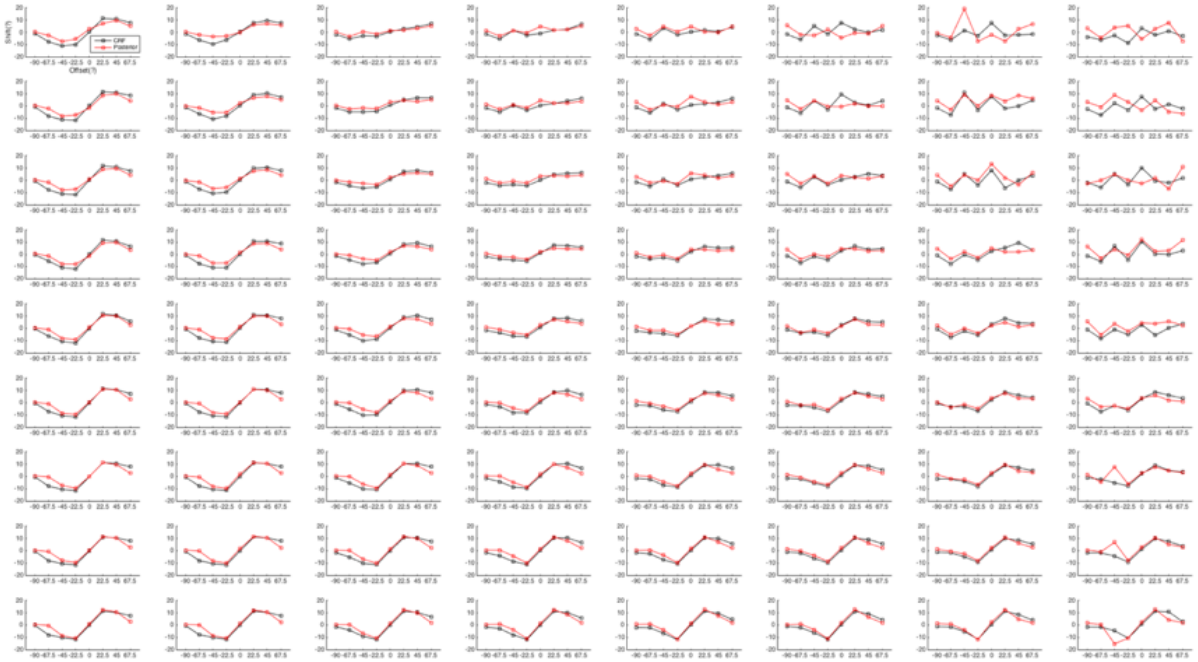


Figure 10. Full results for posterior probability distribution (Channel basis function: 25°). Color convention is the same as in Fig. 3. (a) Results for shift mechanism. (b) Results for gain mechanism. Within each panel, posterior probability was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Panels are organized on a 9 by 8 grid. Rows represents different neuronal tuning width parameters and columns represents different voxel variances.

a



b

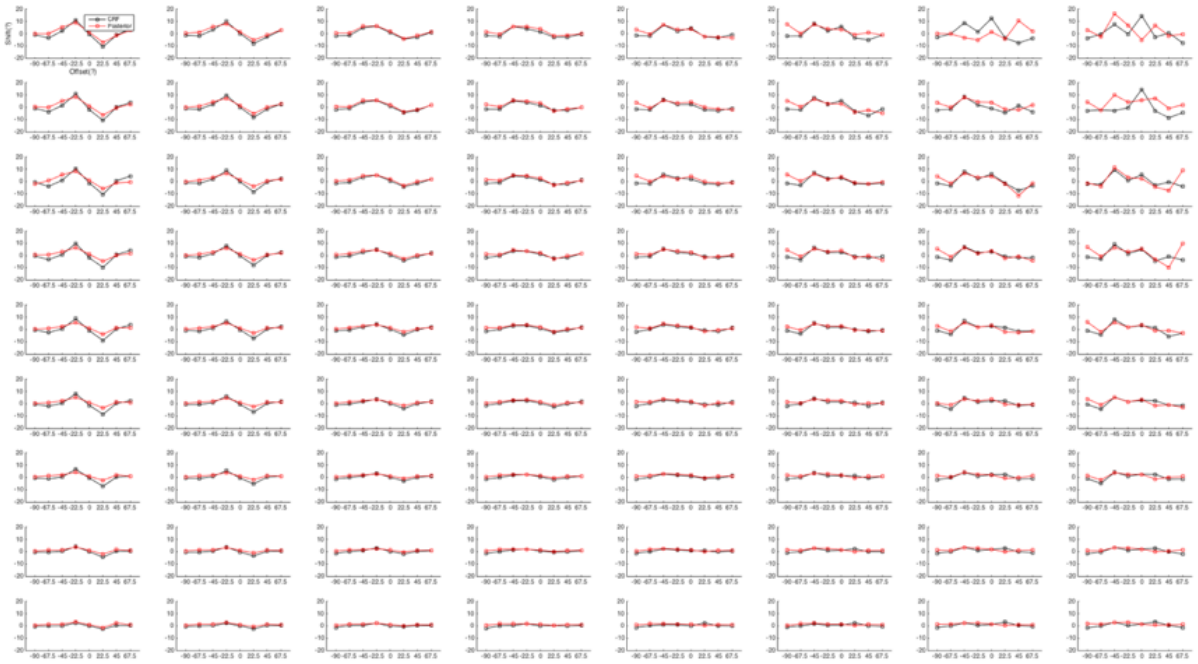


Figure 11. Full results for orientation shift (Channel basis function: 25°). (a) full results for shift mechanism. (b) full results for gain mechanism. Panels are plotted on a 6 by 8 grid. Rows represents different neuronal tuning width parameters and columns represents different voxel variances. Within each panel, orientation shift was plotted for each offset condition. Figure convention is same as in Fig. 5a & 5b. The amount of shift was computed by subtracting the estimated orientation (i.e., mean of fitted von Mises function) from the actual stimulus orientation after fitting the CRF (black) and posterior probability distribution (red).

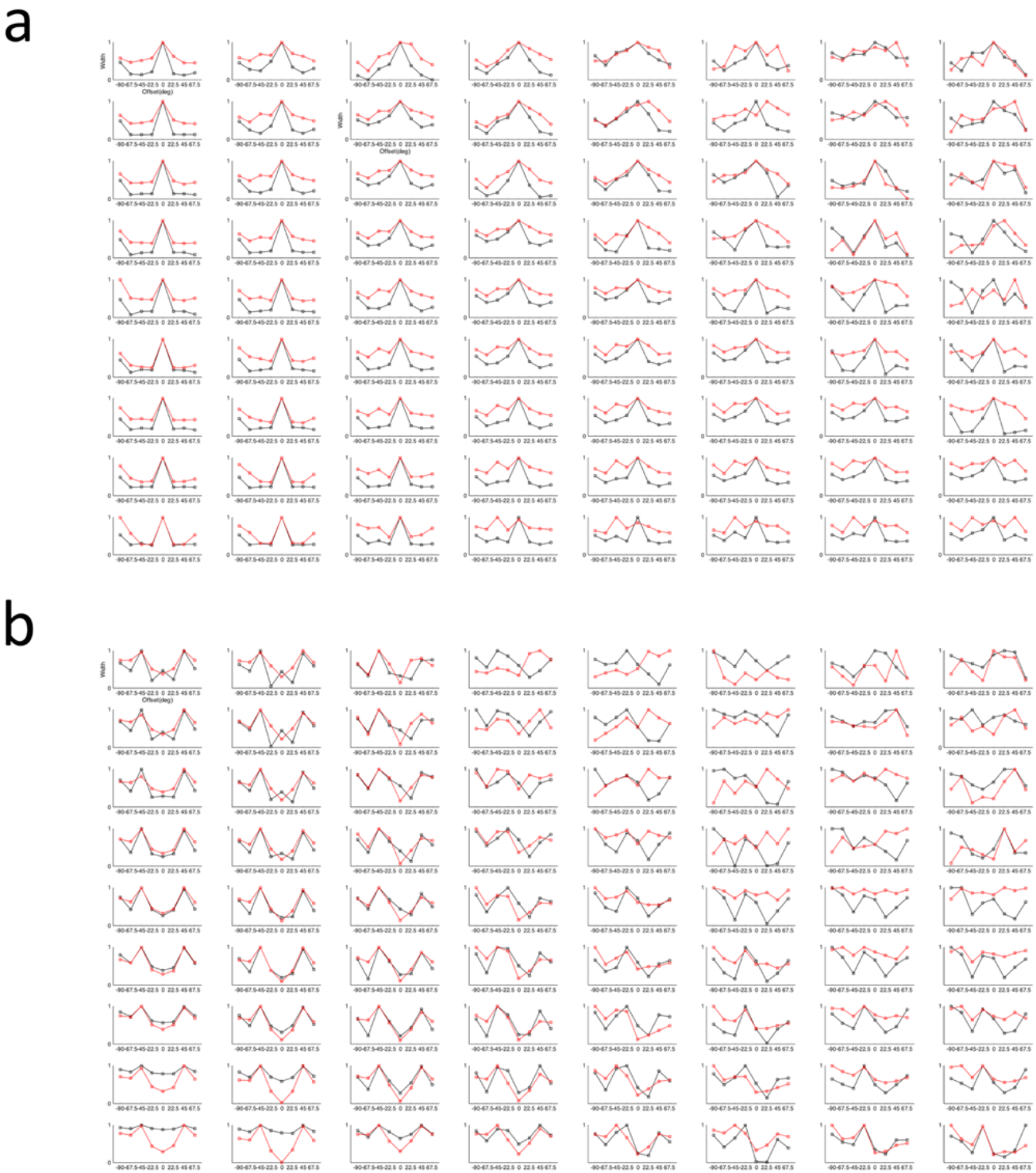
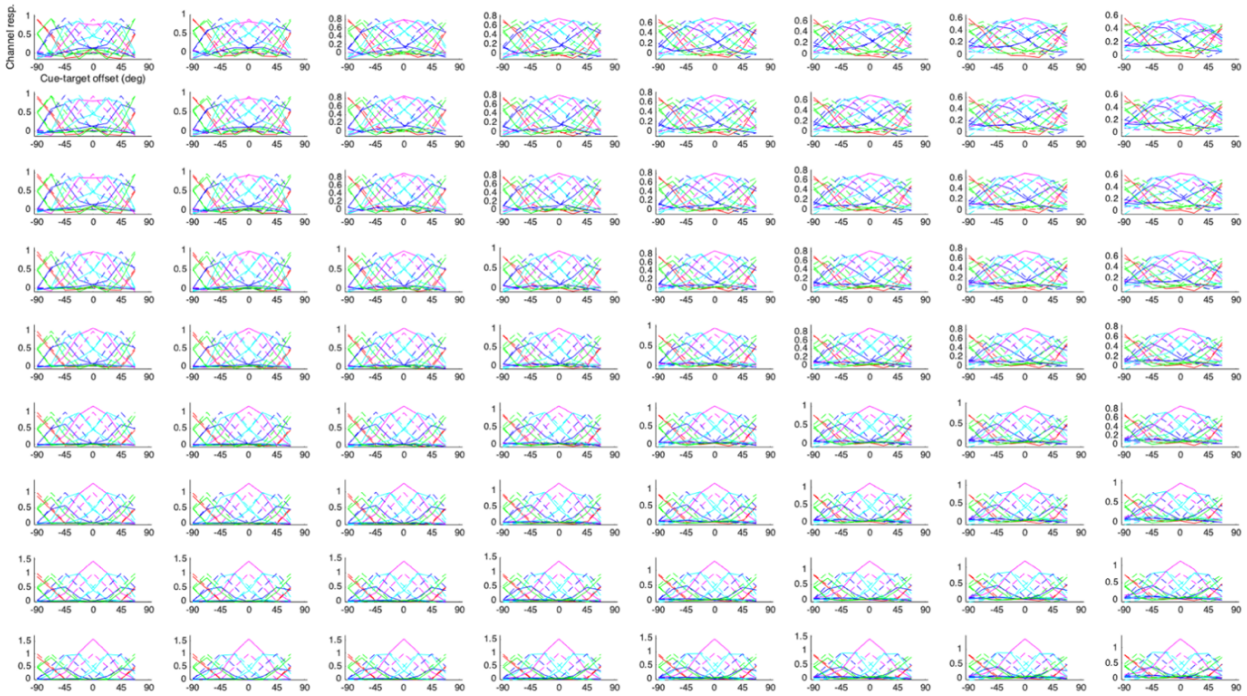


Figure 12. Full results for normalized width (Channel basis function: 25°). (a) full results for shift mechanism. (b) full results for gain mechanism. Figure convention is same as Fig. 5c & 5d. Rows represents different neuronal tuning width parameters and columns represents different voxel variances. Within each panel, estimated width was plotted for each offset condition after fitting CRF (black) and posterior probability (red). For each combination of parameters, estimated width was normalized relative to maximum value of the 8 offset values.

a



b

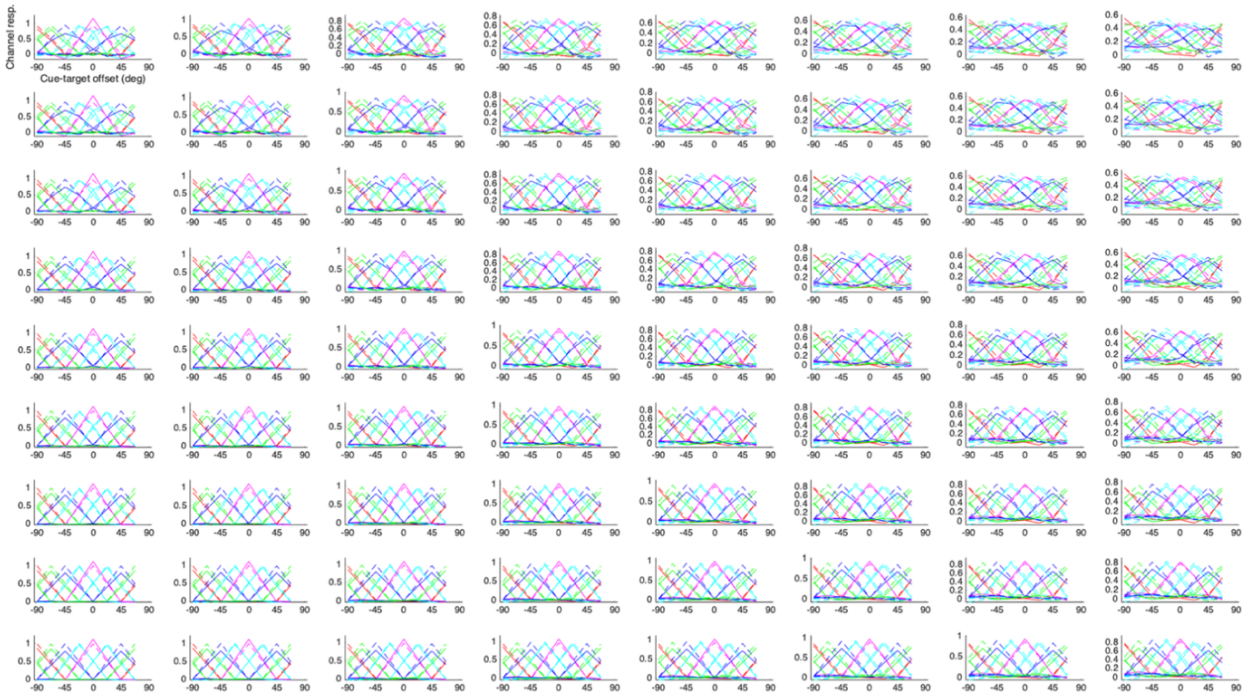


Figure 13. Full results for reconstructed CRF (Channel basis function: 45°). Color convention is the same as in Fig. 3. (a) Results for shift mechanism. (b) Results for gain mechanism. Within each panel, CRF was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Rows represents different neuronal tuning width parameters (25°, 30°, 35°, 40°, 45°, 50°, 55°, 60°, 65°), and columns represents different voxel variances (λ: 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, & 35%). See Method section for details.

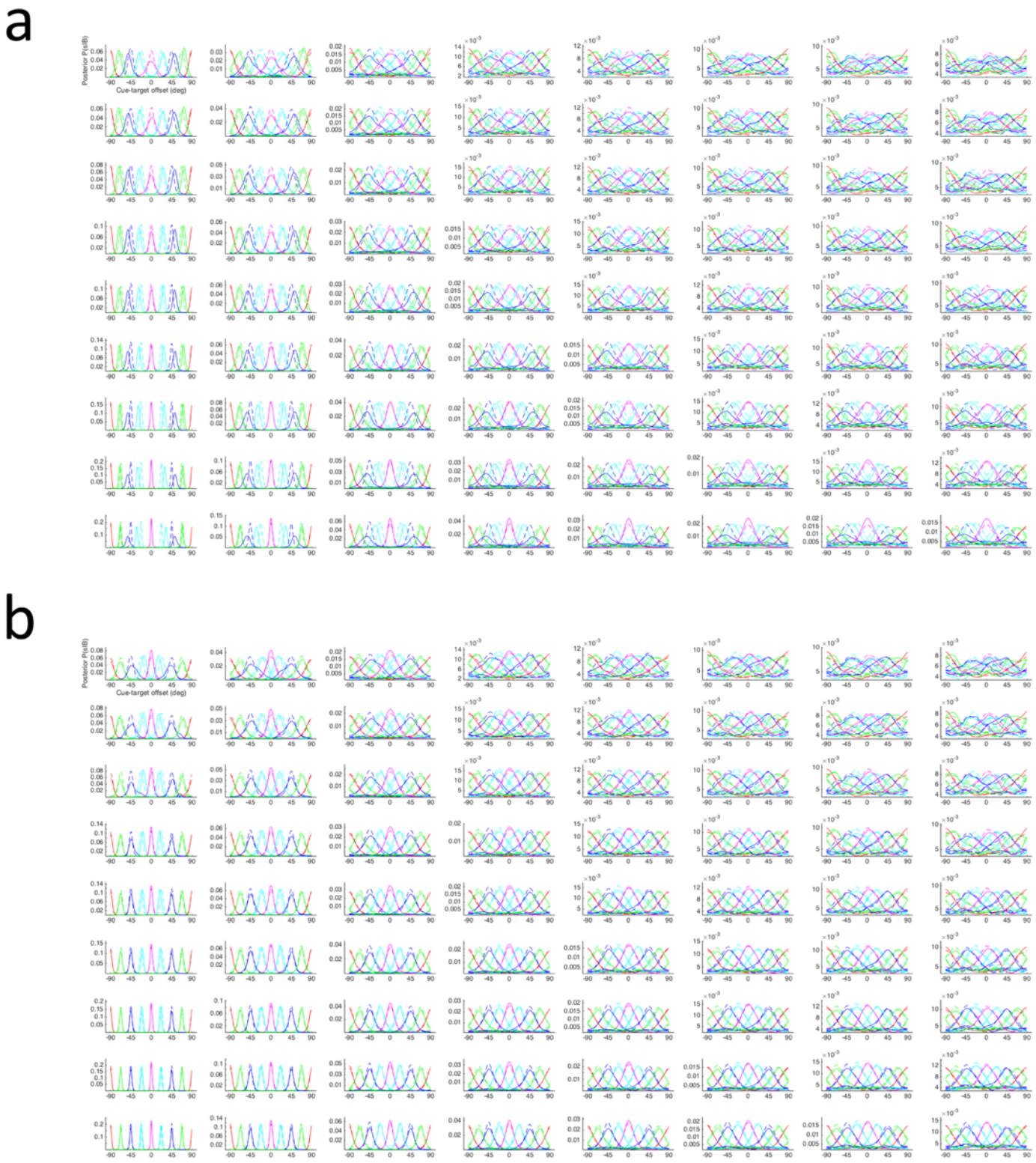
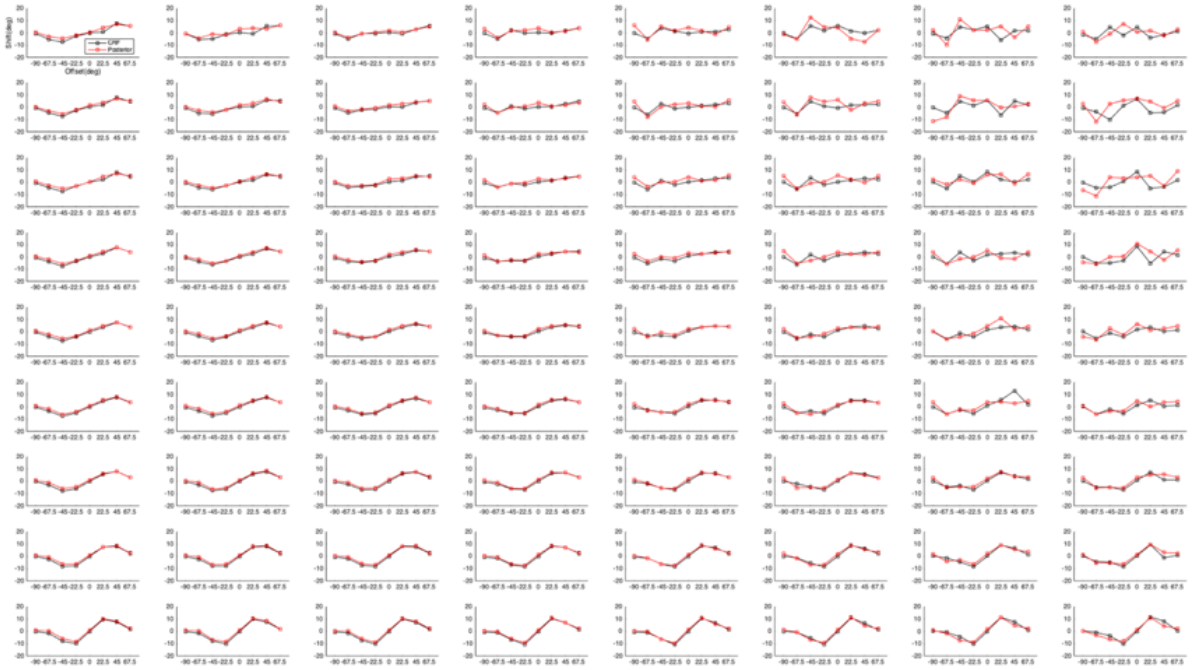


Figure 14. Full results for posterior probability distribution (Channel basis function: 45°). Color convention is the same as in Fig. 3. (a) Results for shift mechanism. (b) Results for gain mechanism. Within each panel, posterior probability was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Rows represents different neuronal tuning width parameters and columns represents different voxel variances.

a



b

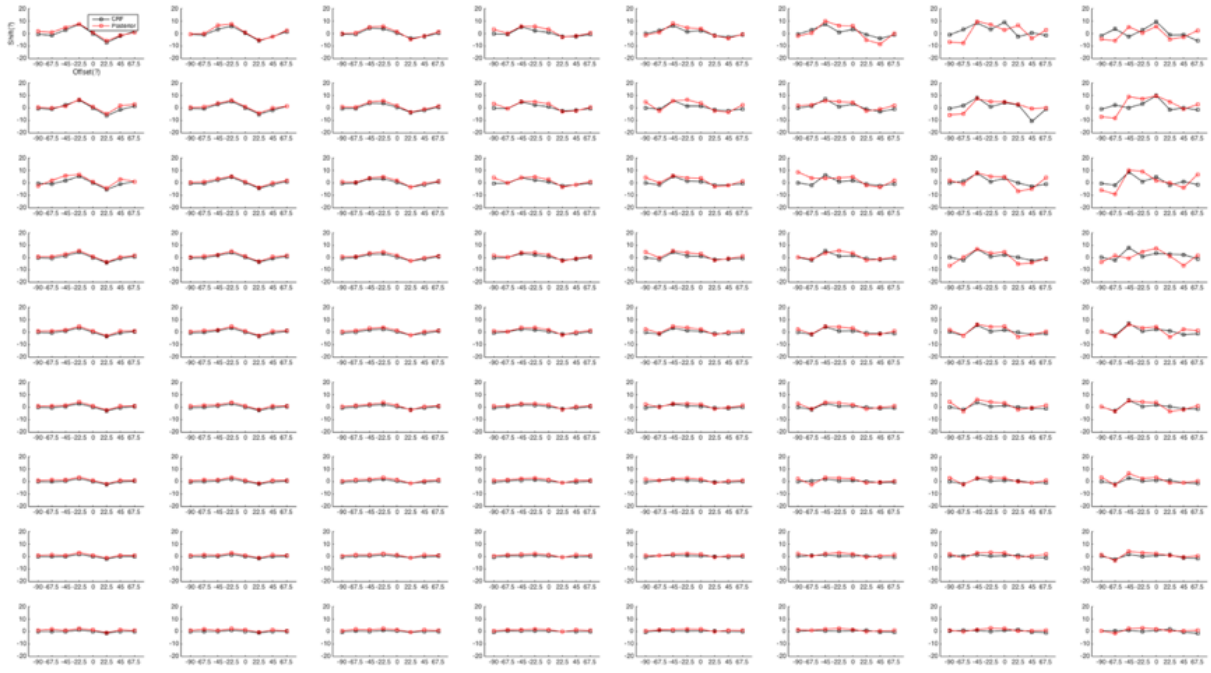
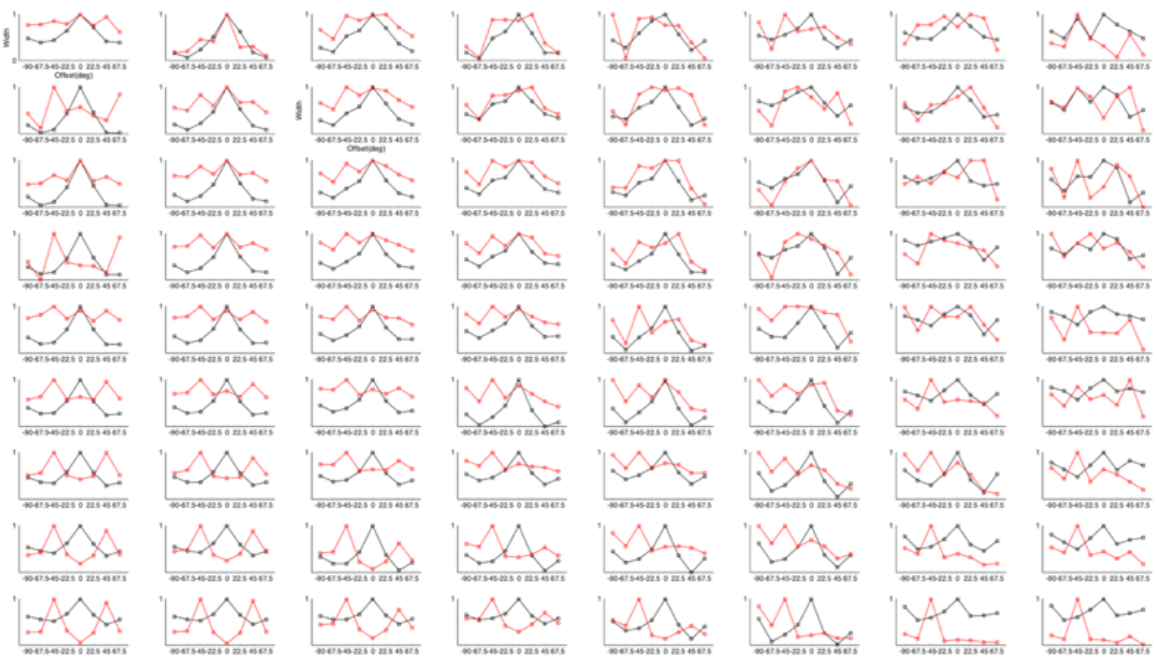


Figure 15. Full results for orientation shift (Channel basis function: 45°). (a) full results for shift mechanism. (b) full results for gain mechanism. Panels are plotted on a 6 by 8 grid. Rows represents different neuronal tuning width parameters and columns represents different voxel variances. Within each panel, orientation shift was plotted for each offset condition. Figure convention is same as in Fig. 5a & 5b. The amount of shift was computed by subtracting the estimated orientation (i.e., mean of fitted von Mises function) from the actual stimulus orientation after fitting the CRF (black) and posterior probability distribution (red).

a



b

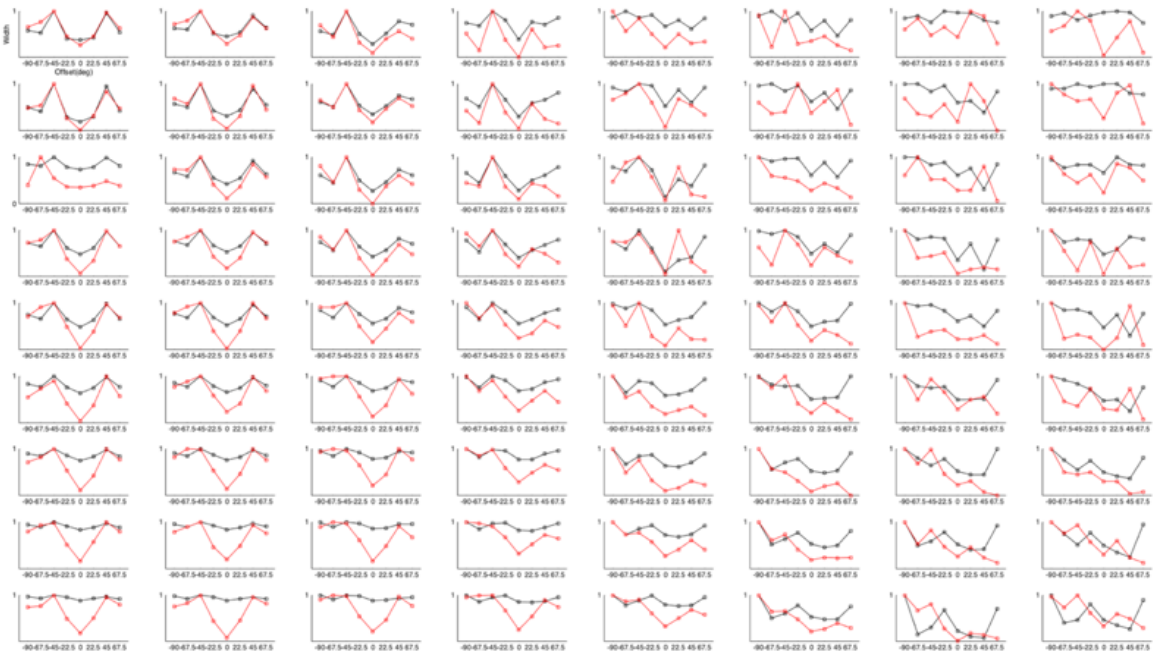


Figure 16. Full results for normalized width (Channel basis function: 45°). (a) full results for shift mechanism. (b) full results for gain mechanism. Figure convention is same as Fig. 5c & 5d. Rows represents different neuronal tuning width parameters and columns represents different voxel variances. Within each panel, estimated width was plotted for each offset condition after fitting CRF (black) and posterior probability (red). For each combination of parameters, estimated width was normalized relative to maximum value of the 8 offset values.

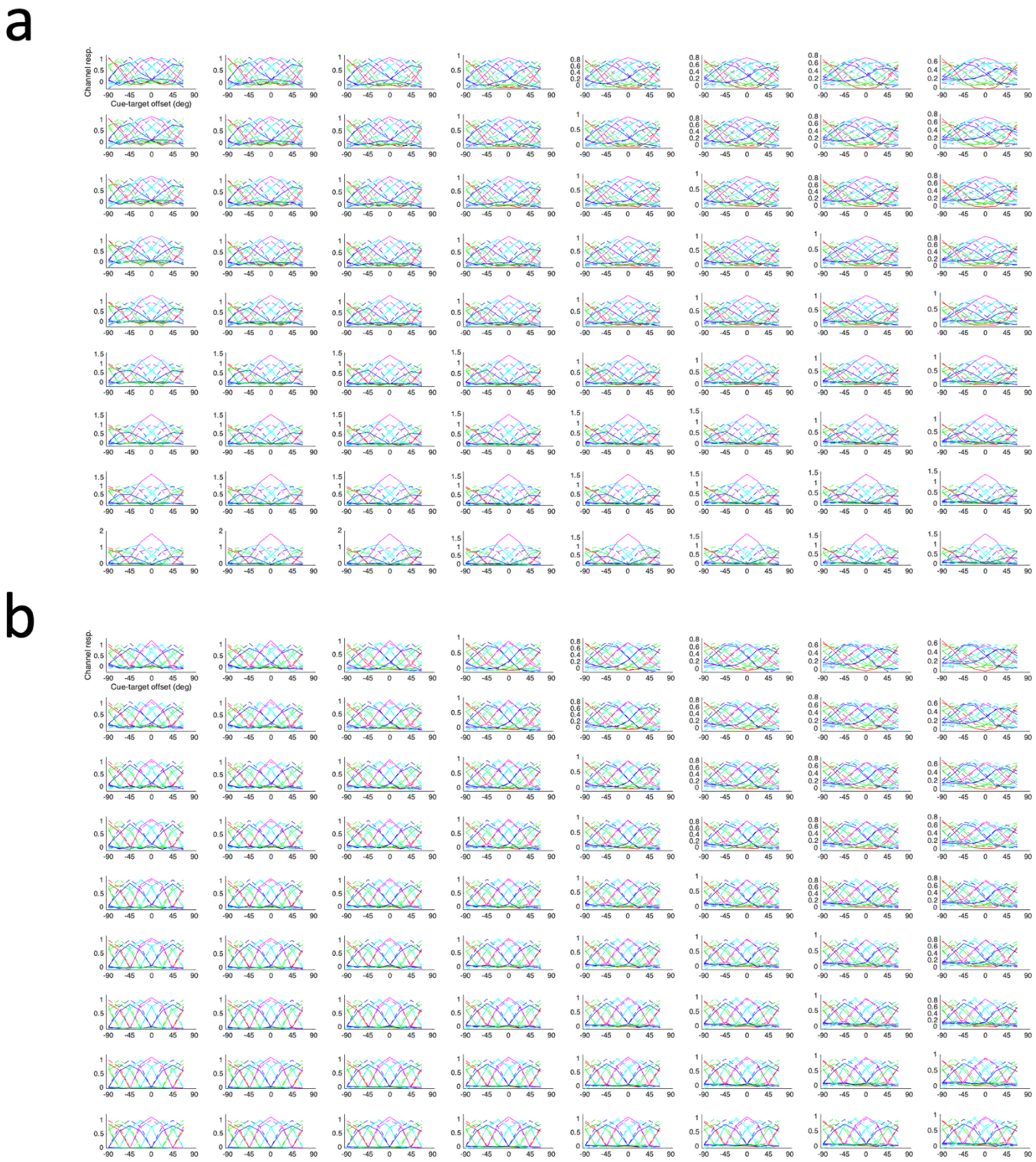


Figure 17. Full results for reconstructed CRF (Channel basis function: 65°). Color convention is the same as in Fig. 3. (a) Results for shift mechanism. (b) Results for gain mechanism. Within each panel, CRF was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Rows represents different neuronal tuning width parameters (25°, 30°, 35°, 40°, 45°, 50°, 55°, 60°, 65°), and columns represents different voxel variances (λ: 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, & 35%). See Method section for details.

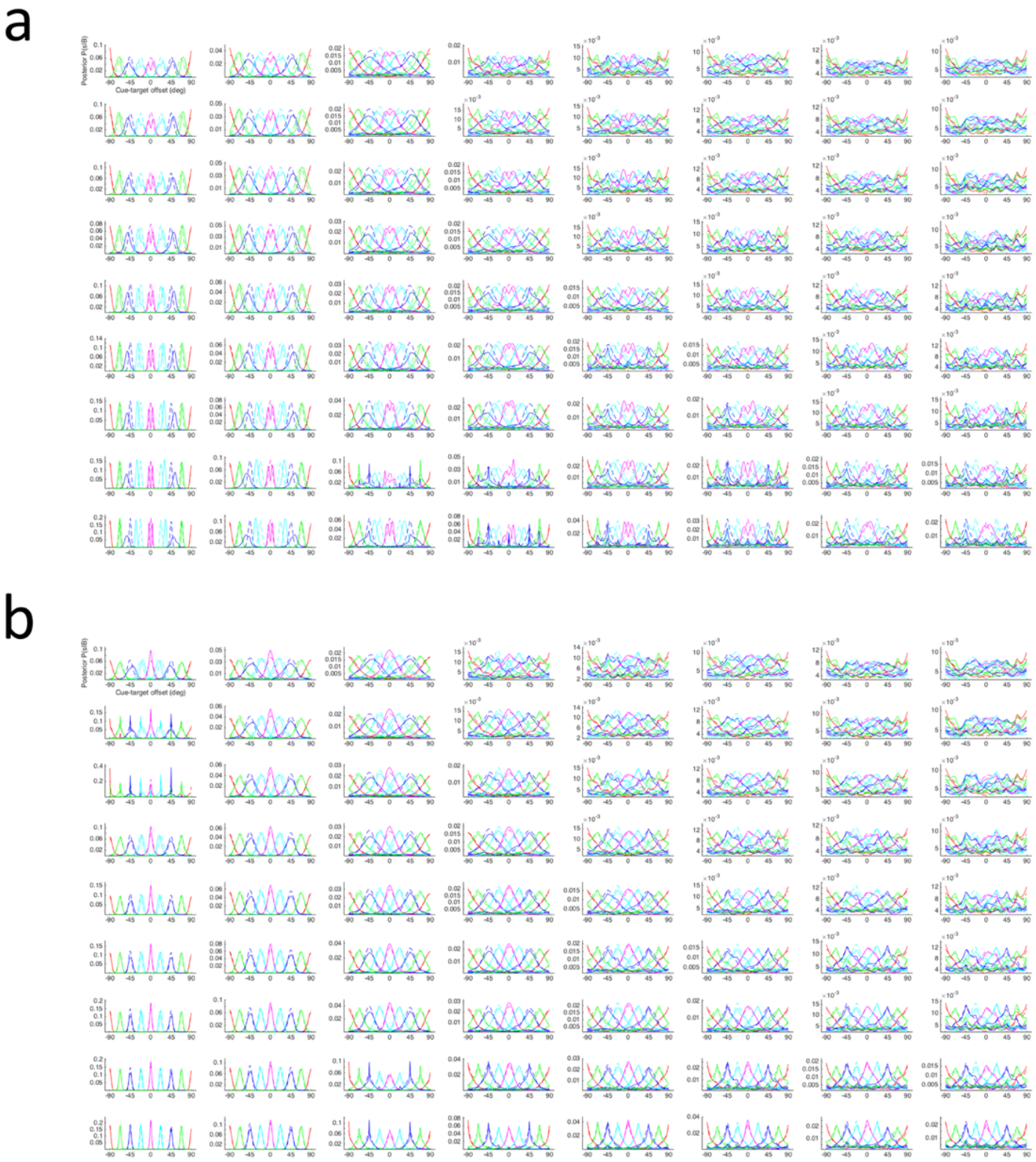
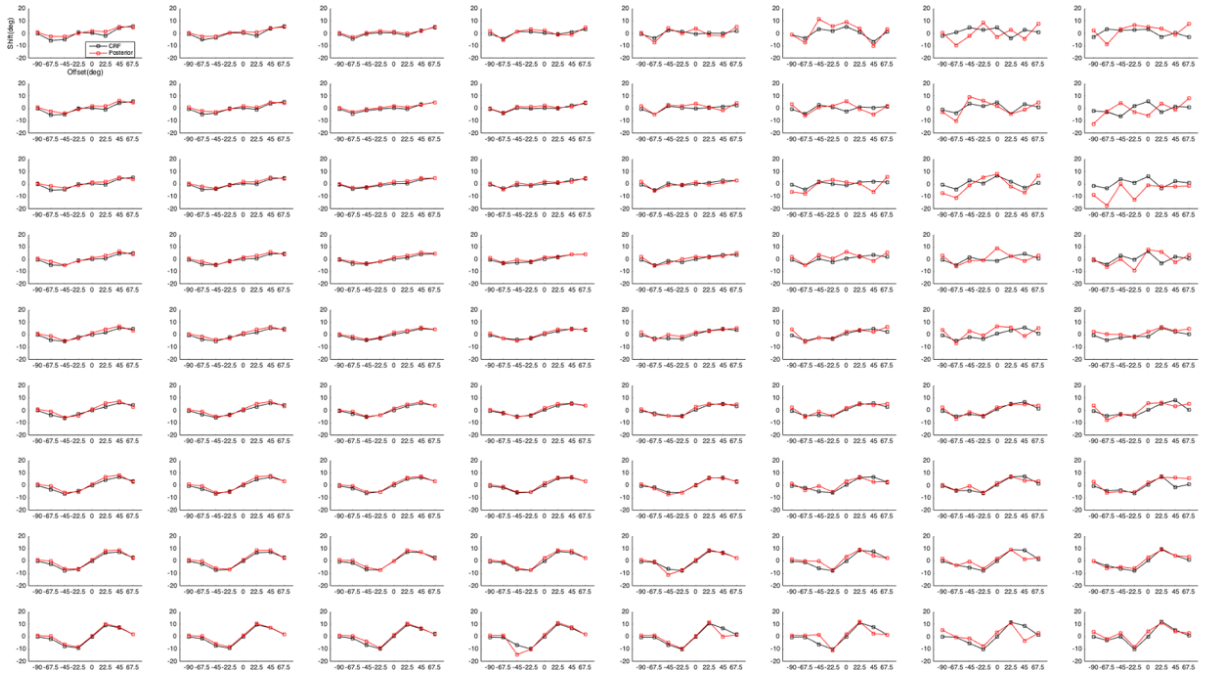


Figure 18. Full results for posterior probability distribution (Channel basis function: 65°). Color convention is the same as in Fig. 3. (a) Results for shift mechanism. (b) Results for gain mechanism. Within each panel, posterior probability was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Rows represents different neuronal tuning width parameters and columns represents different voxel variances.

a



b

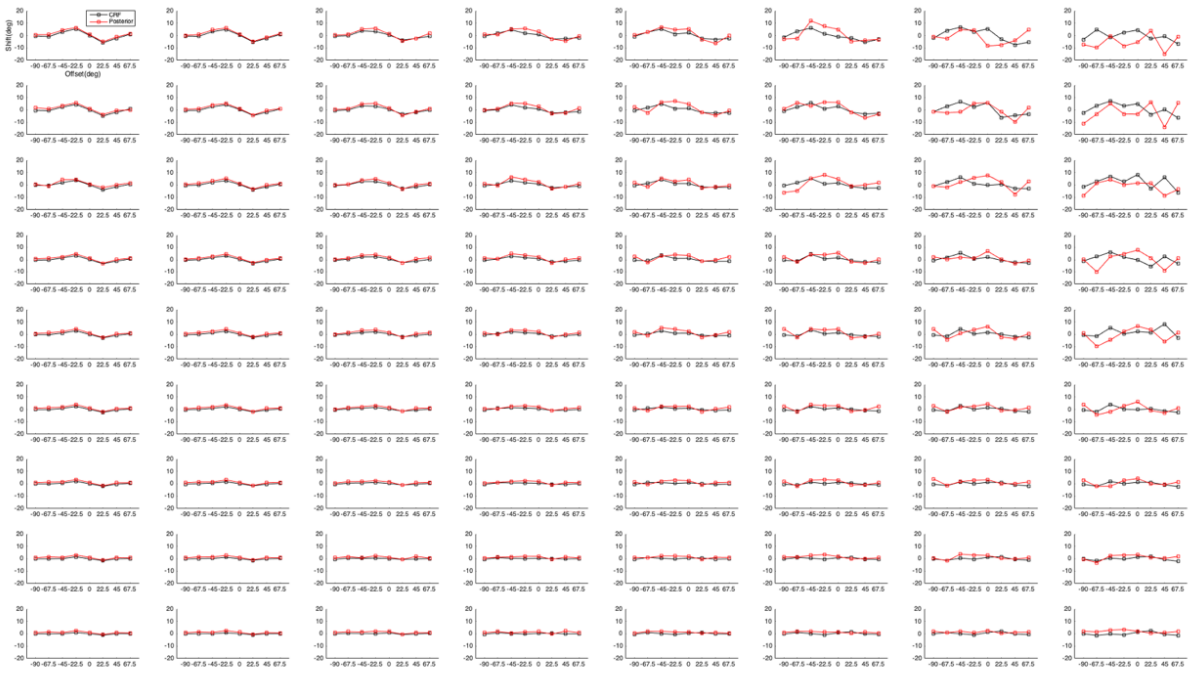


Figure 19. Full results for orientation shift (Channel basis function: 65°). (a) full results for shift mechanism. (b) full results for gain mechanism. Panels are plotted on a 6 by 8 grid. Rows represents different neuronal tuning width parameters and columns represents different voxel variances. Within each panel, orientation shift was plotted for each offset condition. Figure convention is same as in Fig. 5a & 5b. The amount of shift was computed by subtracting the estimated orientation (i.e., mean of fitted von Mises function) from the actual stimulus orientation after fitting the CRF (black) and posterior probability distribution (red).

Figure 20. Full results for normalized width (Channel basis function: 65°). (a) full results for shift mechanism. (b) full results for gain mechanism. Figure convention is same as Fig. 5c & 5d. Rows represents different neuronal tuning width parameters and columns represents different voxel variances. Within each panel, estimated width was plotted for each offset condition after fitting CRF (black) and posterior probability (red). For each combination of parameters, estimated width was normalized relative to maximum value of the 8 offset values.

Figure 21. Full results for a pure feature-similarity gain modulation. (a) reconstructed CRF. (b) Estimated posterior probability distributions. Within each panel, CRF/posterior probability was plotted at each individual offset (attentional condition: solid, neutral condition: dashed). Rows represents different neuronal tuning width parameters, and columns represents different voxel variances.

REFERENCES

REFERENCES

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology, 52*(6), 1106– 1130.

Anton-Erxleben, K., & Carrasco, M. (2013). Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, *14*(3), 188–200.

Anton-Erxleben, K., Stephan, V. M., & Treue, S. (2009). Attention Reshapes Center-Surround Receptive Field Structure in Macaque Cortical Area MT. *Cerebral Cortex*, *19*(10), 2466–2478.

Averbeck, B. B., Latham, P. E. & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5), 358–366.

Bartsch, M. V., Loewe, K., Merkel, C., Heinze, H.-J., Schoenfeld, M. A., Tsotsos, J. K., & Hopf, J.-M. (2017). Attention to Color Sharpens Neural Population Tuning via Feedback Processing in the Human Visual Cortex Hierarchy. *The Journal of Neuroscience*, *37*(43), 10346–10357.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Visual search for colour targets that are or are not linearly separable from distractors. *Vision Research*, *36*(10), 1439–1465.

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science (New York, N.Y.)*, *308*(5721), 529–534.

Brouwer, G. J. & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *The Journal of Neuroscience*, *29*(44), 13992–14003.

Brouwer, G. J. & Heeger, D. J. (2011). Cross-orientation suppression in human visual cortex. *Journal of Neurophysiology*, 106(5), 2108–2119.

Brouwer, G. J. & Heeger, D. J. (2013). Categorical Clustering of the Neural Representation of Color. *The Journal of Neuroscience*, *33*(39), 15454–15465.

Buracas, G. T. & Boynton, G. M. (2007). The Effect of Spatial Attention on Contrast Response Functions in Human Visual Cortex. *Journal of Neuroscience*, *27*(1), 93–97.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525.

Carrasco, M., Ling, S. & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, *7*(3), 308–313.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I. & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, *487*(7405), 51–56.

Cohen, M. R. & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 14(7), 811–819.

Connor, C. E., Preddie, D. C., Gallant, J. L., & Van Essen, D. C. (1997). Spatial attention effects in macaque area V4. *Journal of Neuroscience*, 17(9), 3201–3214.

Conway, B. R., & Livingstone, M. S. (2003). Spacetime maps and two-bar interactions of different classes of direction-selective cells in macaque V1. *Journal of Neurophysiology, 89*(5), 2726–2742.

Daoutis, C. A., Pilling, M., & Davies, I. R. L. (2006). Categorical effects in visual search for colour. *Visual Cognition*, *14*(2), 217–240.

David, S. V., Hayden, B. Y., Mazer, J. A., & Gallant, J. L. (2008). Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. *Neuron*, *59*(3), 509–521.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Dugué, L., Merriam, E. P., Heeger, D. J., & Carrasco, M. (2020). Differential impact of endogenous and exogenous attention on activity in human visual cortex. *Scientific Reports*, 10(1), 21274.

D'Zmura, M. (1991). Color in visual search. *Vision Research* 31, 951-966.

Ester, E. F., Anderson, D. E., Serences, J. T. & Awh, E. (2013). A Neural Measure of Precision in Visual Working Memory. *Journal of Cognitive Neuroscience*, 25(5), 754–761.

Ester, E. F., Sprague, T. C. & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus- Specific Mnemonic Representations during Visual Working Memory. *Neuron*, 87(4), 893–905.

Ester, E. F., Sprague, T. C. & Serences, J. T. (2020). Categorical Biases in Human Occipitoparietal Cortex. *Journal of Neuroscience*, 40(4), 917–931.

Ester, E. F., Sutterer, D. W., Serences, J. T. & Awh, E. (2016). Feature-Selective Attentional Modulations in Human Frontoparietal Cortex. *Journal of Neuroscience*, *36*(31), 8188–8199.

Fang, M. W. H., Becker, M. W., & Liu, T. (2019). Attention to colors induces surround suppression at category boundaries. *Scientific Reports*, *9*(1), 1443.

Fang, M. W. H., & Liu, T. (2019). The profile of attentional modulation to visual features. *Journal of Vision*, *19*(13), 13–16.

Found, A., & Müller, H. J. (1996). Searching for unknown feature targets on more than one dimension: investigating a "dimension-weighting" account. *Perception & Psychophysics*, *58*(1), 88–101.

Fusi, S., Miller, E. K. & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37, 66–74.

Gardner, J. L. & Liu, T. (2019). Inverted encoding models reconstruct an arbitrary model response, not the stimulus. *ENeuro*, *6*(2), ENEURO.0363-18.2019-11.

Garcia, J. O., Srinivasan, R. & Serences, J. T. (2013). Near-Real-Time Feature-Selective Modulations in Human Cortex. *Current Biology*, *23*(6), 515–522.

Hara, Y., Pestilli, F. & Gardner, J. L. (2014). Differing effects of attention in single-units and populations are well predicted by heterogeneous tuning and the normalization model of attention. *Frontiers in Computational Neuroscience*, *8*, 12.

Herrmann, K., Montaser-Kouhsari, L., Carrasco, M. & Heeger, D. J. (2010). When size matters: attention affects performance by contrast or response gain. *Nature*, *13*(12), 1554–1559.

Herrmann, K., Heeger, D. J., & Carrasco, M. (2012). Feature-based attention enhances performance by increasing response gain. *Vision Research*, *74*(C), 10–20.

Ho, T. C., Brown, S., Abuyo, N. A., Ku, E.-H. J. & Serences, J. T. (2012). Perceptual consequences of feature-based attentional enhancement and suppression. *Journal of Vision 12,* 1–17.

Hodsoll, J. P., & Humphreys, G. W. (2005). The effect of target foreknowledge on visual search for categorically separable orientation targets. *Vision Research*, *45*(18), 2346–2351.

Hopf, J. M., Boehler, C. N., Luck, S. J., Tsotsos, J. K., Heinze, H. J., & Schoenfeld, M. A. (2006). Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proceedings of the National Academy of Sciences*, *103*(4), 1053–1058.

Ibos, G. & Freedman, D. J. (2014). Dynamic integration of task-relevant visual features in posterior parietal cortex. *Neuron 83(6),* 1468–1480.

Jigo, M., Gong, M. & Liu, T. (2018). Neural Determinants of Task Performance during Feature-Based Attention in Human Cortex. *Eneuro*, 5(1), ENEURO.0375-17.2018-15.

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

Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352–355.

Klein, B. P., Harvey, B. M. & Dumoulin, S. O. (2014). Attraction of Position Preference by Spatial Attention throughout Human Visual Cortex. *Neuron*, *84*(1), 227–237.

Kohn, A., Coen-Cagli, R., Kanitscheider, I. & Pouget, A. (2016). Correlations and Neuronal Population Information. *Annual Review of Neuroscience,* 39(1), 237–256.

LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(3), 371–379.

Li, X., Lu, Z.-L., Tjan, B. S., Dosher, B. A. & Chu, W. (2008). Blood oxygenation level-dependent contrast response functions identify mechanisms of covert attention in early visual areas. *Proceedings of the National Academy of Sciences*, *105*(16), 6202–6207.

Ling, S., Liu, T., & Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Research*, *49*(10), 1194–1204.

Liu, T. (2019). Feature-based attention: effects and control. *Current Opinion in Psychology*, *29*(Trends Neurosci 29 2006), 187–192.

Liu, T., Abrams, J. & Carrasco, M. (2009). Voluntary Attention Enhances Contrast Appearance. *Psychological Science*, *20*(3), 354–362.

Liu, T., Cable, D. & Gardner, J. L. (2018). Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width. *Journal of Neuroscience*, *38*(2), 398–408.

Liu, T., Fuller, S. & Carrasco, M. (2006). Attention alters the appearance of motion coherence. *Psychonomic Bulletin & Review*, *13*(6), 1091–1096.

Liu, T., Hospadaruk, L., Zhu, D. C. & Gardner, J. L. (2011). Feature-specific attentional priority signals in human cortex. *Journal of Neuroscience*, 31(12), 4484–4495.

Liu, T. & Hou, Y. (2013). A hierarchy of attentional priority signals in human frontoparietal cortex. *Journal of Neuroscience*, 33(42), 16606–16616.

Liu, T., Larsson, J., & Carrasco, M. (2007). Feature-Based Attention Modulates Orientation-Selective Responses in Human Visual Cortex. *Neuron*, *55*(2), 313–323.

Liu, T., & Jigo, M. (2017). Limits in feature-based attention to multiple colors. *Attention, Perception, & Psychophysics*, *79*(8), 1–11.

Liu, T., Pestilli, F. & Carrasco, M. (2005). Transient Attention Enhances Perceptual Performance and fMRI Response in Human Visual Cortex. *Neuron*, *45*(3), 469–477.

Livingstone, M. S., & Conway, B. R. (2003). Substructure of direction-selective receptive fields in macaque V1. *Journal of Neurophysiology, 89(5),* 2743–2759.

Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience 9*, 1432–1438.

Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature, 503(7474), 78–84.

Martinez-Trujillo, J. C. & Treue, S. (2002). Attentional Modulation Strength in Cortical Area MT Depends on Stimulus Contrast. *Neuron*, *35*(2), 365–370.

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. *Current Biology*, *14*(9), 744–751.

Maunsell, J. H. R. & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322.

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience*, *19*(1), 431–441.

McAdams, C. J., & Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*(3), 1751–1755.

Moore, C. M., & Egeth, H. (1998). How does feature-based attention affect visual processing? *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1296–1310.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784.

Mounts, J. R. W. (2000a). Evidence for suppressive mechanisms in attentional selection: Feature singletons produce inhibitory surrounds. *Perception & Psychophysics*, *62*(5), 969–983.

Mounts, J. (2000b). Attentional capture by abrupt onsets and feature singletons produces inhibitory surrounds. *Attention*, *62*(7), 1485–1493.

Müller, H. J., Heller, D., & Ziegler, J. (1995). Visual search for singleton feature targets within and across feature dimensions. *Perception & Psychophysics*, *57*(1), 1–17.

Muller, N. G., & Kleinschmidt, A. (2004). The attentional 'spotlight's' penumbra: Center-surround modulation in striate cortex. *Neuroreport*, 15(6), 977–980.

Murray, S. O. (2008). The effects of spatial attention in early human visual cortex are stimulus independent. *Journal of Vision*, *8*(10), 2–2.

Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Navalpakkam, V., & Itti, L. (2007). Search Goal Tunes Visual Features Optimally. *Neuron*, *53*(4), 605–617.

Newsome, W. T., & Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201–2211.

Paltoglou, A. E., & Neri, P. (2012). Attentional control of sensory tuning in human visual perception. *Journal of Neurophysiology*, *107*(5), 1260–1274.

Pestilli, F., & Carrasco, M. (2005). Attention enhances contrast sensitivity at cued and impairs it at uncued locations. *Vision Research*, *45*(14), 1867–1875.

Pestilli, F., Carrasco, M., Heeger, D. J. & Gardner, J. L. (2011). Attentional Enhancement via Selection and Pooling of Early Sensory Responses in Human Visual Cortex. *Neuron*, *72*(5), 832–846.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Pouget, A., Dayan, P. & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience 1*(2), 125–132.

Pouget, A., Dayan, P. & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Reviews Neuroscience 26*, 381–410.

Reynolds, J. H. & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185.

Reynolds, J. H., Pasternak, T. & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–714.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*(7), 631–632.

Sàenz, M., Buraĉas, G. T., & Boynton, G. M. (2003). Global feature-based attention for motion and color. *Vision Research*, *43*(6), 629–637.

Saproo, S. & Serences, J. T. (2014). Attention improves transfer of motion information between V1 and MT. *Journal of Neuroscienc*e, 34(10), 3586–3596.

Sclar, G. & Freeman, R. D. (1982). Orientation selectivity in the cat's striate cortex is invariant with stimulus contrast. *Experimental Brain Research*, *46*(3), 457–461.

Scolari, M., Seidl-Rathkopf, K. N. & Kastner, S. (2015). Functions of the human frontoparietal attention network: Evidence from neuroimaging. *Current Opinion in Behavioral Sciences*, 1, 32–39.

Scolari, M., & Serences, J. T. (2009). Adaptive allocation of attentional gain. *Journal of Neuroscience*, *29*(38), 11933–11942.

Scolari, M., & Serences, J. T. (2010). Basing Perceptual Decisions on the Most Informative Sensory Neurons. *Journal of Neurophysiology*, *104*(4), 2266–2273.

Scolari, M., Byers, A. & Serences, J. T. (2012). Optimal deployment of attentional gain during fine discriminations. *Journal of Neuroscience*, *32*(22), 7723–7733.

Serences, J. T., & Boynton, G. M. (2007). Feature-Based Attentional Modulations in the Absence of Direct Visual Stimulation. *Neuron*, *55*(2), 301–312.

Serences, J. T. & Saproo, S. (2012). Computational advances towards linking BOLD and behavior. *Neuropsychologia*, *50*(4), 435–446.

Shih, S. I., & Sperling, G. (1996). Is there feature-based attentional selection in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 758–779.

Sprague, T. C., Boynton, G. M. & Serences, J. T. (2019). The importance of considering model choices when interpreting results in computational neuroimaging. eNeuro, ENEURO.0196-19.2019-21.

Sprague, T. C., Saproo, S., & Serences, J. T. (2015). Visual attention mitigates information loss in small- and large-scale neural codes. *Trends in Cognitive Sciences*, 19(4), 215–226.

Störmer, V. S., & Alvarez, G. A. (2014). Feature-Based Attention Elicits Surround Suppression in Feature Space. *Current Biology*, *24*(17), 1985–1988.

Suzuki, S. & Cavanagh, P. (1997). Focused attention distorts visual space: An attentional repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 443–463.

Tombu, M., & Tsotsos, J. K. (2008). Attending to orientation results in an inhibitory surround in orientation space. *Perception & Psychophysics*, *70*(1), 30–35.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treue, S. (2014). Attentional Selection: Mexican Hats Everywhere. *Current Biology*, *24*(18), 1–2.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual atention via selective tuning. *Artificial Intelligence,* 78(1–2), 507–545.

Tsotsos J.K. (2011) *A Computational Perspective on Visual Attention.* Cambridge, MA: MIT press.

van Bergen, R. S., Ma, W. J., Pratte, M. S. & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730.

van Bergen, R. S. & Jehee, J. F. M. (2018). Modeling correlated noise is necessary to decode uncertainty. *NeuroImage*, *180*(Part A), 78–87.

Wang, Y., Miller, J., & Liu, T. (2015). Suppression effects in feature-based attention. *Journal of Vision*, *15*(5), 15–15.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2), 202–238.

Womelsdorf, T., Anton-Erxleben, K., Pieper, F., & Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature Neuroscience*, *9*(9), 1156–1160.

Womelsdorf, T., Anton-Erxleben, K., & Treue, S. (2008). Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. *Journal of Neuroscience*, *28*(36), 8934–8944.

Yoo, S.-A., Martinez-Trujillo, J., Treue, S., Tsotsos, J. K. & Fallah, M. Feature-based attention induces surround suppression during the perception of visual motion.

Zhang, W., & Luck, S. J. (2008). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience*, *12*(1), 24–25.