CAMBRIDGE UNIVERSITY

Mathematical Models of the Representation of Faces in Humans

Jonathan Neil O'Keeffe
St John's College

supervised by
Dr. Nikolaus KRIEGESKORTE

This dissertation is submitted for the degree of
Doctor of Philosophy

October, 2017

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit.

Abstract

Mathematical Models of the Representation of Faces in Humans

Jonathan Neil O'Keeffe

The representation of faces is a crucial function of the human CNS, as demonstrated by the severe social difficulties experienced when people lack this ability (prosopagnosia). However, the precise way in which faces are represented and differentiated from one another is not well understood. This work addresses two substantial issues.

Firstly, how is information about faces integrated over time? In chapter 2 a simple model of temporal integration is set forth, based on the statistical technique of exponential smoothing. In chapter 3 results of experiments testing this model are presented, demonstrating the model to be inadequate in certain respects. In particular a systematic bias towards the

origin of face space is observed, a phenomenon referred to as ”bowing”. Chapter 4 contains a further model, which aims to show how this bowing could arise from a Bayesian inferential process.

The second issue, addressed in chapter 5 of this thesis, is how well human judgements of facial similarity correspond to predictions made using Basel Face Space (BFS), a popular and widely used representation of faces from the field of computer vision. The degree of agreement is quantified using a novel experimental approach, and subsequently salient differences between the biological face space and BFS, including some original findings relating to isotropy or directionality, are demonstrated.

Acknowledgements

I would like to thank Dr Nikolaus Kriegeskorte for his advice and suggestions at multiple points over the course of my work, and in particular for his input to the work described in chapter 5, which was crucial to the design of the experiments. Many other persons at the MRC CBU have also been of assistance.

For reasons of my own making my time in Cambridge was relatively solitary, but of the several people with whom I did have contact my friend Nicky Ferguson was a particular stalwart and deserves special mention. My family and friends outside of Cambridge were also a great support.

Finally, I would like to acknowledge my Welsh terrier, Rosie, who has been my near constant companion since beginning my PhD. In addition to her company and affection, her agility and intelligence have been a forceful reminder to me of just how far we are from understanding, let alone recreating, the motor and perceptual systems of the living world.

Contents

1

Introduction

In this introductory chapter some of the fundamental issues concerning levels of explanation and analysis within neuroscience are first introduced. Leaning on the work of David Marr, it is argued that an understanding at the algorithmic/representational level is of great practical import in addition to its conceptual validity. The classical cognitive neuroscience paradigm is discussed, and in particular the central concept of a representation, citing some of the challenges which

have been made to this construct, particularly from within dynamical systems and robotics. There follows an account of Valentine's conception of face space, placing it within its historical and intellectual context. Finally I discuss the "Bayesian brain" within neuroscience reviewing some of the basic concepts. The chapter concludes with a summary that draws these threads together and relates them to the substance of this thesis.

1.1 Levels of Explanation in Neuroscience

There is in biology today an understandable, and in many respects salutary, bias towards molecular and cellular mechanisms, as pronounced in neuroscience as anywhere. However, as has been explicitly recognised, at least since Marr's seminal contributions to the field [Marr, 1983], a full understanding of the brain must comprise a representational and algorithmic account in addition to low level mechanistic accounts.

This is a matter with very practical implications, well beyond a purely academic concern for the full understanding of a system. A case in point is the field of neuromodulation or functional neurosurgery. This field began with efforts at using DBS (Deep Brain Stimulation) to treat the symptoms of motor disorders, such as Parkinsonian and essential tremor. Remarkably it turns out that essentially the same methods can be applied to treatment resistant depression and obsessive compulsive disorder [O'Keeffe, 2011, Dyster et al., 2016]. And yet, there is essentially no consensus on the mechanism of action by which this therapy works [Torres-Sanchez et al.,

2017]. This then is a clear example of where our ability to intervene has outstripped our scientific understanding of the phenomenon, analogous to the development of steam engines before the science of thermodynamics, which governs their operation, was understood. While early steam engines such as the Savery Engine were useful, they exhibited efficiencies on the order of a few percent, and the support of a sound body of scientific theory, provided by Carnot, Watt and many others, was required to unleash the full potential of this technology [Cardwell, 1971]. Accordingly in neuroscience, as attempts to read and write information within neural circuits grow in ambition [1], for therapeutic or other purposes, so it becomes increasingly imperative to understand neural communication and computation at the systems, representational and algorithmic level.

Of the many strata that make up the stack of modern neuroscience this thesis focuses at a relatively high level. Little will therefore be said about firing rates, ion channels or genes and most discussion will be confined to a computational examination of a particular hypothetical representation within the human brain, that of *face space*.

1.2 The Nature of Mental Representations

Lying at the heart of cognitive science is the postulate of a *representation* within the brain. An attractive view of the nervous system sees it implementing a mapping between sensory inputs to adaptive motor outputs.

[1] For example in the fascinating, and therapeutically extremely promising, field of machine-brain interfacing [Lebedev and Nicolelis, 2006]

Because this cannot practicably be done using a look-up-table approach it has been argued that the brain must construct abstract representations, or symbols, of the world with which to compute evolutionarily fit actions [Gallistel and King, 2009]. Indeed, some have generalised the thesis to include *all possible* intelligent systems, natural and artificial, as in the renowned PSSH (Physical Symbol System Hypothesis) of Simon and Newel which asserted that a "physical symbol system has the necessary and sufficient means for general intelligent action" [Newell and Simon, 1976]. Not everyone has found this conception of intelligence altogether persuasive as model for any intelligent system, least of all natural ones. Practising physiologists for example, acquainted with the reality of nervous systems might observe that, empirically, we do not find such physical symbols when we look within actual brains [Shadlen et al., 2008]. Equally certain philosophers, most notably Hubert Dreyfus, have argued that the conceptual basis of the PSSH, and thereby its psychological dependant cognitive science, is irretrievably misconceived [Dreyfus, 1992]. However the case against the PSHH (Physical Symbol System Hypothesis) was perhaps most trenchantly put in Rodney Brook's 1990 article "Elephants Don't Play Chess" in which he argued that the PSSH had not only failed at a practical level, yielding few if any tangible results in robotics and AI, but was also, echoing Dreyfus, flawed in principle since, far from being a good idea to explicitly represent or model the world,

"the world is its own best model. It is always exactly up to date. It always has every detail there is to be known. The trick is to sense it appro-

priately and often enough." [Brooks, 1990].

This document principally describes attempts to model and experimentally probe a hypothetical representation in the primate visual cortex, so-called face space [Valentine, 1991b]. As such it falls squarely within the compass of cognitive science, implying that the brain really does physically instantiate what is essentially a state space. At the beginning of the work for this PhD certain criticisms of the classical paradigm underlying cognitive science should have been accorded greater weight that was the case, in particular those from the field of dynamical systems neuroscience and contained within the so-called "computation vs dynamics" debate. Without dilating too greatly on the nuances of this debate, the basic issue concerns whether biological systems really need construct explicit representations or symbols in order to compute adaptive motor outputs. The confidence of the author in the computational approach had been reinforced early on in his studies by Gallistel and King's excellent book *Memory and the Computational Brain* in which they made, to my mind, a strong case for a representational view of cognition [Gallistel and King, 2009]. Since then however this confidence has been undermined by the success, and theoretical appeal, of a dynamical systems approach to understanding biological computation, paralleled in robotics by the SED (Situated Embodied Dynamic) framework [Pfeifer and Bongard, 2006]. Perhaps it is most likely that the resolution to this debate will contain components of both outlooks, as has been argued [Mitchell, 1995]. However, I am conscious that this research has been conducted as something of an *unrecon-*

structed computationalist, to coin a phrase. Were one to undertake this work again quite a different approach might be adopted. One more informed by developments in dynamical systems neuroscience and the SED framework within robotics.

1.3 The Origins of Valentine's Conception of Face Space

1.3.1 Precedents of Face Space

Face space has provided a general unifying, and perhaps dominant, framework for much of the research done in the study of face perception over the past two decades [Valentine et al., 2016]. Some light can be cast on the ascendancy of face space as a framework by considering what it replaced. In 1975 Hadyn Ellis published an influential review article in the British Journal of Psychology entitled *Recognizing Faces* [ELLIS, 1975]. In it he drew attention to the lack of theoretical underpinning then current work in face recognition. Partly in response to this influential review, attempts were made to furnish the field of face perception with a theoretically satisfying account of face perception. Rightly or wrongly, many of the subsequent efforts to provide such a theory were derivative, in the sense that they sought to bring concepts from other fields of psychology to bear on face perception. One such contribution was a popular model by Bruce and Young [Bruce and Young, 1986], based on theoretical work in word recognition [Morton et al., 1979]. However, evidence for this model centred on

neuropsychological studies and a single diary study [Young et al., 1985], complicating questions of empirical support. Added to this, Bruce and Young's model had nothing to say about the nature of the visual processing of faces at a computational level, and little by way of experimental prediction regarding the recognition of unfamiliar faces (as faces) [Valentine et al., 2016]. Similarly, efforts to provide an explanatory framework based on schema theory, a product originally of research into memory [Goldstein and Chance, 1980], could account for some of the apparent properties of face perception, such as the inversion effect, but failed to provide testable predictions [Valentine et al., 2016]. In summary, then, throughout the 1970s and 1980s even the most popular of the then-current frameworks for face recognition offered only fragmentary explanatory coverage and/or generated few testable predictions required to arbitrate between competing theories. Moreover, the absence of a canonical theoretical framework, naturally enough, prompted researchers to deploy in their experiments extremely simple stimuli, such as those shown in figure 1.1. Such abstractions may have been well intentioned, being motivated by something like a desire to isolate the essence of the phenomena, much as physics sought to understand the motion of pendulums before whirlpools. However, the effect of this heuristic in face perception was to throw out the baby with the bathwater. Indeed, it is the high-dimensional natural variation in human faces that constitutes the central computational challenge for a biological, or indeed artificial, face recognition system [Valentine et al., 2016].

Despite the lack of a widely accepted paradigm within the field empirical work continued over subsequent years, yielding several results

which prepared the ground for Valentine's conception of face space. Working within the context of schema theory Light, Kyra-Stuart and Hollander [Light et al., 1979] published a study in 1979 in which they demonstrated a distinctiveness effect across a wide range of exposure conditions. In particular they showed that recognition accuracy from memory was greater for faces rated as distinctive than for those rated as typical. In keeping with schema theory the authors presented the effect of distinctiveness in terms of deviation from a prototypical face. This result was later confirmed by Valentine and Bruce and supplemented with a further key finding, namely that the effect of distinctiveness (improved recognition) could be reversed (distinctive faces being more poorly recognised) by changing the task demands [Valentine and Bruce, 1986]. In particular, although consistent with Goldstein's and Chance they found that distinctive faces were indeed recognised faster than typical faces, yet when the task was changed to classifying faces from among jumbled faces (i.e. non-faces) distinctive faces actually took longer than typical faces to be classified. This result was at odds with schema theory-based accounts, since it suggests what what enables a face to be recognised rapidly is not intrinsic to the mental representation itself (e.g. deviation from a definitive schema), but highly dependant on the demands of the particular task.

To summarise, by the late 1980s one could describe the field as comprising a raft of suggestive experimental findings, though lacking a coherent and unifying theoretical framework. Schema theory was, evidently, not the answer, and a more radical departure was required to achieve something like a unified account. Moreover, the desire for basic insights had

led researchers to using increasingly simple and artificial stimuli, to such an extent that it was debatable whether their stimuli any relevance at all to face perception as a natural phenomenon.

Against this backdrop Valentine published in 1991 a paper wherein he presented a framework to account for most of the then current empirical findings as well as furnishing experimentalists with novel empirical predictions [Valentine, 1991b]. In Valentine's conception face space is viewed as a high dimensional space in which the dimensions of face space are hypothesised to correspond to a broad and fairly indeterminate notion of "feature". Both the precise dimensionality of face space (i.e. the number of orthogonal features) and the nature of the "features", to which the axes of face space correspond, are left undefined in Valentine's original conception. Having said that one influential interpretation is that the axes of face space correspond to something very close to the eigenvectors resulting from PCA (Principal Component Analysis) performed on faces distributed in pixel space [Turk and Pentland, 1991, Bartlett et al., 2002]. Interestingly there is increasingly good empirical and theoretical support that the brain does indeed perform computations like this in face recognition [Leibo et al., 2017].

1.3.2 Norm Vs Exemplar Face Space

An important consideration in the formulation of face space is the nature of the metric governing similarity/dissimilarity. Debate has largely centred around two categories of models, know as *exemplar* models and *norm*

models [Valentine et al., 2016, Lewis, 2004, Leopold et al., 2001]. In exemplar models of face space the dissimilarity metric is a function of the Euclidean distance within feature space between a particular face and faces which has been seen previously (exemplars). Notice that this metric is isotropic in the sense that the Euclidean distance is not dependent on the frame of reference [Bishop, 2006]. In contrast a norm model implies that the dissimilarity metric is dependent on the prototypical, average or norm face. Moreover, this metric is not isotropic, in that changes in the radial vector (i.e. distance from the norm) represents changes in distinctiveness, while changes in the tangential vector (orthogonal to the radial vector) represent changes in identity. Perhaps rather surprisingly a proposal for the actual form of this norm metric remains absent from the literature [Lewis, 2004], making its intrinsic coherence rather difficult to assess, quite apart from the question of empirical support.

Prominent in the norm-versus-exemplar debate have been results relating to the phenomenon of adaptation within facial perceptions. In general terms adaptation is a recalibration of a perceptual system following exposure to a particular stimulus characteristic [Blakemore et al., 1970]. One of the earliest descriptions of the phenomenon for face perception was provided by Lewis and Ellis who showed that displaying 30 different views of an identity would slow recognition, compared to showing just three views [Lewis and Ellis, 2000]. Subsequently adaptive effects were demonstrated across a wide range of facial dimensions including attractiveness [Rhodes G., 2003], personality [Buckingham, 2006] and identity [Leopold et al., 2001]. Leopold et al.'s 2001 paper is of particular inter-

est because the results were framed as supporting a norm based account of identity recognition. Using the same BFS model addressed in chapter 5, the authors showed that adaptation to a face produced the strongest perceptual change (measured via perceptual thresholds for identity) for identities lying on the same trajectory (i.e. on a line passing through the adapted identity and the norm or origin of face space), as compared to identities not on the same trajectory. Several points appear germane to this claim. Firstly, the study assumed that the BFS model was a sound model of human face space, the question addressed in chapter 5. From the data presented there we can concede that this assumption is borne out to a reasonable degree. A second point is that the inference from a psychophysical finding to an assertion about the tuning properties of single neurons relies on a chain of reasoning that can hardly be considered watertight. Leopold and colleagues did subsequently seek and find neurons "tuned" to the norm [Leopold et al., 2006], although a major question concerns whether this study was not subject to a strong confirmation bias.

On first sight it may seem that the norm account of face space provides an immediate explanation of the curious caricature effect. According to the norm-based account a caricature, corresponding to a vector with the same direction but greater magnitude than the veridical face, is simply a more distinctive version of the veridical face. Hence, the increased ease of recognition. However, a concomitant finding is that the caricature effect is only present up to an exaggeration of 16%, after which it declines and disappears [Lee et al., 2000]. As is discussed in subsection 1.3.3 Valentine's version of an exemplar based face space provides an elegant account

of why this should occur, but it is not at all clear why this would occur in a norm based account. On the contrary, it seems a clear prediction of the norm account that this should not occur [Lewis, 2004]. In addition, once one begins to consider other well established phenomena for which a norm based model must account, in particular the other race-effect, it becomes difficult to avoid positing multiple norms, making the norm model indistinguishable from exemplar models [Valentine and Endo, 1992,Lewis, 2004]. Furthermore, thanks to simulation studies, it is now clear that exemplar models *can* account for many well established psychophysical effects, in addition to the findings of Leopold et al. [Ross et al., 2014]. Thus, following the development of exemplar based accounts that can explain phenomena such as the other race effect, the caricature effect, and the results of certain adaptation experiments, there has been reduced interest in norm based accounts of face space. In subsection 1.3.3 Valentine's version of exemplar based face space is described and its implications for these and other experimental findings discussed. There will be no further discussion of norm based accounts in this or subsequent chapters, and references to Valentine's face space should be interpreted as referring to his exemplar based presentation. The explanatory implications of this account are fleshed out in section 1.3.3

1.3.3 Explanatory Properties of Valentine's Face Space

Valentine's general framework is agnostic about the exact nature of the features used in face perception. However, it makes definite assertions about

the distribution of faces within face space, namely that this distribution is (approximately) multivariate normal. That appears a very reasonable assumption, primarily because it is an almost trivial observation that continuous properties of biological objects, such as height, weight, I.Q. etc. tend to approximate a normal distribution. It is therefore a sensible null hypothesis that the features utilised for natural face recognition, be those what they may, are normally distributed.

Given then some distribution of faces within face space, Valentine maintains that the difficulty of recognition for some particular face depends on 1) the error of the encoding process, 2) the distance of the face from its nearest neighbour and 3) the distance from the second nearest neighbour. Later accounts, such as Byatt and Rhodes' *Absolute Coding Face Space* [Lee et al., 2000], argued that the distances from *all* other faces were relevant in determining the ease of recognition.

Developmentally one can suppose that the neural architecture learns adapts to the axes of variation one comes across in one's environment, resulting in a normal distribution in neural representational space that reflects one's developmental environment. This is a crucial element of the framework because it is known that children learn to recognise and distinguish those faces they encounter in their developmental environment, and not faces in general [Gruter et al., 2008]. In something like the same way children learn the features and statistics of the language in their environment, not languages in general.

A further aspect of Valentine's presentation is that dissimilarity in face space is governed by a Euclidean metric. There is no requirement that this

should be the case, and indeed solid evidence underpinning this assumption is conspicuously absent. However, one relevant observation is that over small distances many non-Euclidean metrics behave much the same as a Euclidean metric [Lewis, 2004].

From these assumption, concerning the normal distribution of faces and Euclidean metric, a number of experimental results can be explained, including distinctiveness effects, caricature effects, developmental dependencies, adaptation effects and, of particular note, the so-called "other race" effect.

The other race effect follows from the supposing that the axes of variation are different for different ethnicities. In that case a neural face space adapted to Asian faces would reflect variation in Caucasian faces poor, and all Caucasian faces would be essentially clustered together at an centripetal location. This would produce the impression of a distinctive ethnicity but would prevent good differentiation between individuals, exactly as is observed [Richler and Gauthier, 2014].

The inversion effect is the observation that rotating a face 180°results in a catastrophic failure in perception and recognition of faces [Thompson, 1980]. In keeping with the given account of the other race effect, just as one is adapted to the ethnically associated variation present in one's environment, one is also adapted to the statistics and features of generally upright faces, not inverted faces. Subsequent research has supported this account, in particular with regards to the sensitivity to the orientation of facial features [Psalta et al., 2014].

The distinctiveness effect described by Light et al. [Light et al., 1979],

whereby distinctive faces are better recognised from among other faces but more poorly recognised from among other non-face objects, can also be elegantly accounted for. Distinctive faces are by definition found at centripetal locations in face space, where there are fewer distractors than at central locations in terms of density. In contrast, distinctive faces are also, by construction, less typical than more typical faces, in that they are more eccentric than the average face, explaining why they are less easily identified as being faces from among non-face objects.

Related to distinctiveness effects is the caricature effect, a caricature being a more distinctive/eccentric version of a veridical face. In this framework caricatures are more distinctive (i.e. more centripetal) versions of veridical faces, and therefore also located in regions of lower distractor density, making them easier to recognise [Lee et al., 2000]. The trade-off is that as a caricature becomes ever more eccentric and enters regions of lower and lower distractor density, it becomes increasingly dissimilar to the veridical face and thus harder to identify. This jibes well with the finding by Rhodes et al. [S., 1987] that the caricature advantage increased only up to caricatures of 16%, after which the advantage declined.

1.3.4 Representation and Readout in Face Space

When considering face space as a functional representation one important question is how information about facial identity is represented in a population of neurons. This is addressed in subsection 1.3.2, which concerns the longstanding debate of Norm Vs Exemplar coding, and which is per-

haps only well known to those acquainted with the literature around facial perception and recognition. However, it is worth obtaining an overview of some more general approaches to categorisation and considering their application in face space. For example, although not originally developed with explicit reference to faces, Nosofsky's GCM (Generalised Context Model) [Nosofsky, 2011] has been applied to understanding face recognition, by considering it as categorisation of a face to a particular identity [Valentine, 1991a]. Briefly, given some probe, the probability of the probe belonging to some identity/category is computed by measuring its similarity to exemplars of that identity, divided by the sum of similarities over all other identities. A problem faced by this model, as for the case of the norm based model discussed in subsection 1.3.2, is that it is unable to account for the falling off of the caricature effect beyond a certain eccentricity, and instead predicts a monotonically increasing caricature effect. This cannot very well be correct for the simple reason that as eccentricity increases the face will eventually stop looking like a face at all, let alone a more recognisable caricature of a particular identity. A second problem is that a GCM based approach produces a distribution over all identities, that lacks the *winner-takes-all* quality of common experience, illustrated most vividly by bistable phenomena such as the Necker cube. An interesting question concerns whether different perceptual modalities ultimately employ the same algorithmic principles as one another, versus the possibility that uniquely adapted representational system. Might there not be then, in addition to face space, a "car" space, a "word" space and and "animal" space? Just as faces can morph and merge, apparently

seamlessly from one identity to another, many natural categories display similar properties. Experimentally, it can be shown that a statement such as *A robin is a bird* will be assented to more rapidly than the statement *A penguin is a bird*, because, it is argued, *robin* represents a more prototypical exemplar of the category *bird* [Rips, 1973]. This may be seen as analogous to the finding that more prototypical faces are recognised as faces more rapidly as faces than distinctive faces [Valentine and Bruce, 1986]. Some of the shortcomings of the GCM and other generalised approaches to categorisation applied to face perception, have been address by more recent modelling efforts. For example, Lewis's face-space-R [Lewis, 2004] builds on the basic assumptions of a Euclidean metric and a normal distribution of exemplars within face space. However, the model includes an activation function for each known identity and a subtractive term based on the activity of all other identities, effectively instantiating a competitive process whereby at most one identity can have a positive response. Thus, given some probe identity, if there is no overall positive response the model essentially returns an *unfamiliar face* response, whereas if there is an overall positive response to a known identity (and there can be no more than one positive response), then that identity is returned and the probe is recognised. Thanks to the explicit formulation of the face-space-R model it is possible to perform simple simulations, which demonstrate basic psychophysical phenomena such as distinctiveness effects and the caricature effect. Lewis also performed high dimensional simulations in face-space-R, with realistic numbers of exemplars, allowing him to obtain an (empirically constrained) estimate of the dimensionality of biological

face space of between 15 and 22 dimensions. It should be noted however that other approaches can produce estimates suggesting a much higher dimensionality (e.g. [Chang, 2017]). As will be seen in chapter 2 the model therein described is inspired in part by the class of drift diffusion models that have displayed such explanatory power, both regarding distributions of reaction times in psychophysical experiments and the electrical behaviour of neurones [Romo, 2012, Beck et al., 2008, Shadlen et al., 2008]. While much of this work has utilised unidimensional stimuli, such as the directional coherence of random dot stimuli, there has also been attention paid to the relationship between features in the context of multidimensional stimuli, which must surely be of relevance for a representational space whose dimensionality few researchers would estimate to be lower than double figures (e.g. [Lewis, 2004]). Of particular interest is the issue of dependancy between the dimensional features and whether the brain takes advantage of such dependencies when they are present. A broad basis of work in this field suggests that the brain does indeed take advantage of such dependencies, when they exist, demonstrated behaviourally by the distribution of reaction times among other things [Wenger, 2001]. In Valentine's conception of face space facial features are distributed in a jointly normal fashion. If the axes of face space are furthermore determined by something close to PCA [Turk and Pentland, 1991], as recent neurophysiological evidence strongly suggests [Chang, 2017] then the dependencies investigated by Wenger, Townsend and colleagues may not be relevant, since these conditions, strictly speaking, imply independence.

Given the preceding observations it seems clear that Valentine's con-

ception of face space has much in common with work done in other fields of psychology. However, it does appear that some modifications is required, as with face-space-R, if the basic findings relating to caricature effects, distinctiveness effects and adaptation are to be explained.

Stepping down the hierarchy of explanatory accounts one is forced to consider biologically plausible mechanisms that could instantiate formal descriptions of face space. The properties of individual or populations of neurons, such as spike adaptation, may be helpful in accounting mechanistically for dynamic phenomena such as psychophysical adaptation [Clifford et al., 2007]. Equally, whatever the neural substrate of face space, to be useful the information must be made available to other regions of the brain (e.g. motor regions), so another question concerns how that information is read out of one neuronal population and into another. Simplistically, one might suppose that the neuronal face state space is instantiated by a simple rate code, with specific neuronal populations representing the axes of face space. In this scenario axonal projections to an otherwise segregated neuronal population would make that information available elsewhere for use in computation. More sophisticated models, for example utilising oscillations and so-called *communication through coherence*, can also be adduced and are an active areas of experimental and theoretical work [Fries, 2015, Buzsaki, 2006]. Relating to the focus of this thesis chapter 4 presents a Bayesian account of the inferential process behind face perception, subsequent to which (section 4.7) some fascinating recent work in entomology is discussed. This work, by Seeley and colleagues [Seeley et al., 2012] suggests there may be deep commonalities

between swarm intelligence and neuronal computation, such as that underlying face perception. This possibility is discussed in chapter 4.

Despite the intrinsic interest and importance of connecting computational and neurophysiological accounts, this thesis, beginning with chapter 2, primarily concerns the question of what algorithm is used to update the representation of facial identity over time. To couch this in Marr's classic framework this model addresses the problem at the algorithmic or representational level, not the level of neurophysiology and mechanism [Marr, 1983]. The concept of face space is, for the most part, treated purely as a representation, without recourse or reference to adjacent structures and functions. Face space's function is accordingly assumed to be to *make available* information to the rest of the brain about the face being currently foveated. In chapter 2 a dynamical state space model is presented in which it is assumed that the current state constitutes an estimate of the face being foveated. That is to say that the state is a representation of the foveated face. It *tells* the rest of the brain which face is present in the environment. Of course a central question in neuroscience is *how* a representation in one region of the brain is accessed by another, but on this question the model itself is silent.

1.3.5 Face Space as a Flexible Framework

Although it is not a necessary implications of Valentine's work, the working assumption of most scientists in this field is that face space is a high dimensional space, certainly well beyond the three dimensions of Euclidean

space within which visualisation and intuition may be relied upon to some extent [Burton and Vokey, 1998]. As already discussed, empirical estimates of the dimensionality of face space can only be made in the context of a model of the computations supporting face recognition, but such studies generally support the idea that dimensionality cannot be less than double figures [Sirovich and Meytlis, 2009, Lewis, 2004].

Given this vagueness about the number and indeed nature of the dimensions of face space, a durable aspect of Valentine's conception is its agnosticism concerning the precise dimensionality and specification of the feature space relevant to face perception. This means that it can accommodate a wide array of feature models. Valentine's 1991 paper asserts that many of the properties of face perception, and in particular those concerning distinctiveness, inversion, and recognition, could be parsimoniously accounted for by simply supposing such a high dimensional representation is utilised, irrespective of the exact nature or number of the axes within it. As such, Valentine's face space can be construed as nothing more than a general framework, lacking in details. Alternatively it can be viewed as a judicious and limited claim about the nature of face space, given some very reasonable sounding assumptions about the task within which it performs a role (face perception and recognition) and the nature of the data it must work with (faces under variation in background, pose, lighting etc.).

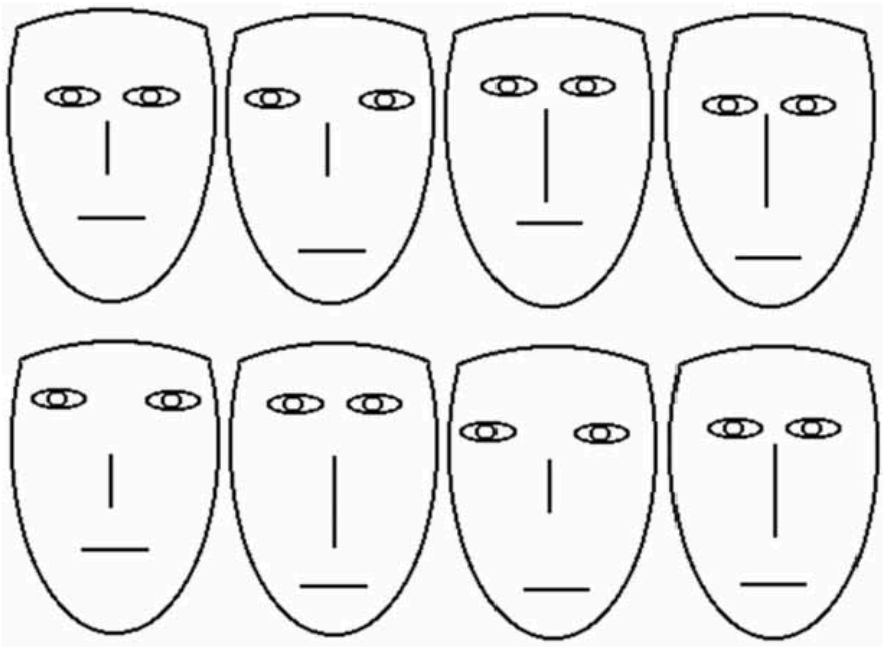


Figure 1.1: The absence of a coherent and comprehensive theoretical framework prior to the advent of Valentine's conception of face space [Valentine et al., 2016] in combination with the evident complexity of natural variation in faces, both between and within individuals (due to pose and lighting for example), led to the use of highly simplified schematic stimuli in experiments. Yet arguably, such stimuli simplified the phenomena so much that they no longer represented one of the central computational challenges of face recognition, the inherent high-dimensional complexity of natural variation.

1.4 The Bayesian Brain

A Bayesian account of perception typically begins with a nod to Reverend Bayes and a quotation from the nineteenth century polymath Helmholtz who held that the brain performs Bayesian reasoning by a process of "unconscious inference". While a pleasing if hackneyed trope, careful examination suggests the attribution to be rather tenuous [Westheimer, 2008]. A more prosaic and plausible account sees probabilistic ideas percolating into the mainstream of psychological and neuroscientific thinking throughout the latter half of the 20th century, from within and without neuroscience, through the pioneering work of figures such as Horace Barlow [Barlow, 1961] and E. T. Jaynes [Jaynes, 1988]. In any case, this approach to the study of mental processes has only increased in prevalence over the past 30 years and nowadays represents a major theme in modern brain science.

The essence of Bayesian inference is that it provides a principled way to combine disparate sources of information. Using it, one can incorporate new information into what was already believed about some aspect of the world, say the bias of a coin. More biologically, a predator might obtain both visual and subsequent auditory information about the location of its prey, which it has to combine to achieve the optimal estimate of the prey's location. Mathematically the information is encoded in the form of a prior probability distribution and a likelihood function, the latter of which represents something like the "plausibility" of the data considered in the context of a model of its generation. In the terminology of the field

one obtains a posterior distribution (updated beliefs) about some aspect of the world by combining the prior distribution (previous beliefs) with the likelihood function (plausibility of the new information under some model). If we denote the new information, or data, by D and the relevant aspect of the world by θ this general concept can be expressed in the form of the canonical Bayes' theorem

$$\underbrace{P(\theta \mid D)}_{\text{posterior}} \propto \underbrace{P(D \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}} \tag{1.1}$$

An appealing aspect of this formalism is that since both the posterior and the prior are well defined probability distributions the output of one application (a posterior) can be inputted to a second application (in the form of a prior). Thus in an algorithmic sense this formalism is perfectly suited to integrating the stream of data that a behaving organism requires to obtain up-to-date estimates of the state of the world, and thereby select adaptive motor plans. This approach, whereby incoming data is recursively integrated with previous knowledge using Bayes' theorem, is known as recursive Bayesian estimation or Bayesian filtering and finds applications across many disparate fields from engineering to linguistics [Simo, 2013].

How much evidence is there that something like this actually happens in behaving organisms? There is now a very significant body of evidence that nervous systems do, at least in certain circumstances, approximate Bayesian inference. Figure 1.2 is reproduced from a paper published by

a leading lab in this area and illustrates the kind of non-uniform natural statistics for which a Bayesian approach to a task, in this case edge orientation estimation, would be well suited. A feature of work in this area is the use of perceptual illusions and biases which have been described and studied in the psychological literature over the past century or more. Part of the elegance and appeal of a Bayesian account of perception is that many perceptual biases can be understood as ”features” rather than ”bugs”. Indeed it turns out that *bias* is typically the price that must be paid in reducing the variance of an estimator, the so-called bias-variance trade-off. Such a bias is found in the estimation of edge orientation where people tend to estimate edges to be closer to vertical or horizontal than they in fact are. By supposing that a Bayesian process of inference underlies this task performance, and the biases therein, Girshick and colleagues were able to show that the prior required to explain such a bias corresponds extremely well with that implied by the statistics of natural scenes [Girshick et al., 2011]. This correspondence is displayed in figure 1.3. In section 1.3.4 the issue of neural mechanisms was touched upon and this is no less an issue for Bayesian theories of the brain. For it would seem that the brain must represent probability distributions, either explicitly or implicitly (e.g. being able to sample from them), if Bayesian inference is to be at all possible [Fiser, 2010]. There is now a large body of theoretical and increasingly experiemtal work showing how a Bayesian inferential process could be implemented, for example by utilising local lateral inhibition [Bill et al., 2015, Berkes P, 2011].

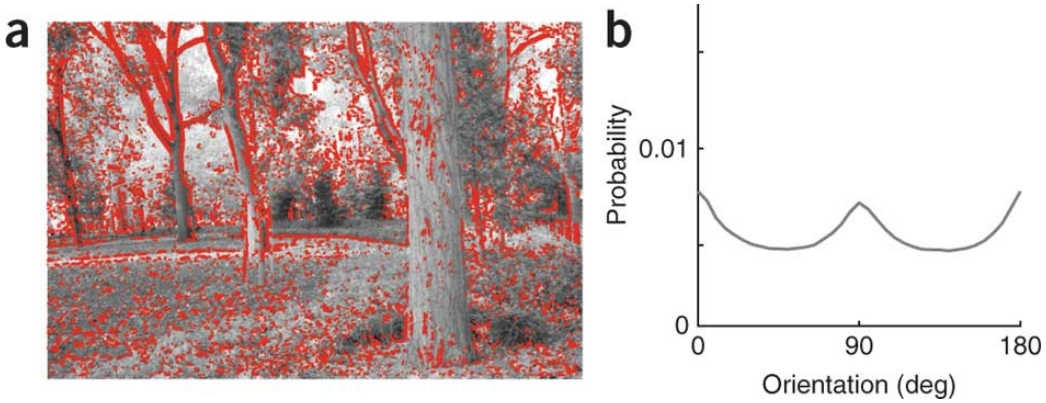


Figure 1.2: The non-uniform distribution of edge orientation in natural scenes. Left panel: a natural scene. Red dots mark the points where edges have been detected and their orientation extracted algorithmically using standard techniques from machine vision. Right panel: the distribution of orientations is far from uniform, showing modes at the so-called cardinal orientations (vertical and horizontal). From [Girshick et al., 2011].

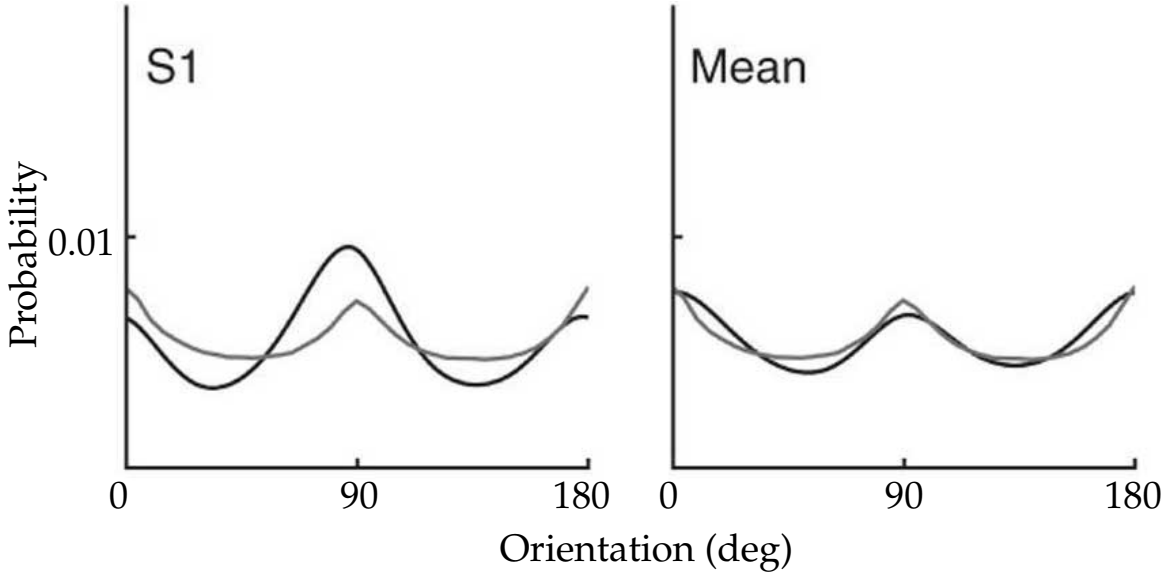


Figure 1.3: The priors extracted from a Bayesian model of the experimental data obtained from subjects making orientation judgements concerning Gabor patches. The left panel shows the extracted prior for a single subject (S1) in black and the environmental prior in grey (obtained by analysis of natural scenes). In the right panel the extracted priors across all 5 subjects have been pooled to obtain a prior for the mean subject (black), which as can be appreciated corresponds well to the environmental prior (grey). From [Girshick et al., 2011].

Having established that the brain does indeed use, at least approximate, Bayesian inference for low level perceptual task, such as edge orientation, a natural question is whether the brain is *in general* a Bayesian machine. Does the brain for example use a form of Bayesian inference for performing identification at the level of whole objects such as cups, computers, phones and indeed faces? And do nervous systems deploy a Bayesian mechanism for selecting the next action that should be taken? Increasingly the field is coalescing around an affirmative answer to these questions in which action and perception are views as complementary limbs of a single Bayesian process, aptly described as *Bayesian action and perception* [Fishel and Loeb, 2012]. This account implies that the the next action is selected so as to provide the maximum amount of sensory information, so that for example the pattern of saccadic eye movements that a person naturally performs when perceiving a face should, in the context of an identification task, maximise the amount of identity information extracted. Interestingly, this prediction is borne our by experiment [Or and Eckstein., 2015], suggesting that the identification of faces represents a particular instance of a more general strategy. Notwithstanding, it deserves mention that several respected detractors have argued against this thesis, for example arguing that Bayesian decision making is fundamentally flawed as an approach to general intelligence (aka "stong" AI) because it is limited to induction and extrapolation, as opposed to creative hypothesis formation [Deutsch, 2012].

In chapter 4 a model of face perception is presented, which supposes that the brain utilises the prior distribution of faces within face space in in-

ferring facial identity. The model is rather coarse in the sense that both information and the effect of noise are encoded in the form of (self-conjugate) Gaussian pdfs, which are combined with one another in order to obtain a (Gaussian) posterior distribution. This assumption simplifies the mathematics involved considerably since it ensures that we need only deal with Gaussians, which can be combined relatively simply. This model is of course not intended to be a realistic representation of the computations involved in face perception. Indeed many of the assumptions implicit in it, such as explicit representation of probability distributions, are fairly implausible. It is rather meant as a demonstration that the oddities within the experimental data presented in chapter 3 can be understood as a natural consequence of probabilistic inference within face space. That is, deviations from what we might consider "correct" perception can instead be understood as functional consequences of a system tuned to perform probabilistic inference by utilising informative priors defined over the relevant state space (face space).

1.5 Summary

Face space has then, as intended, provided a 'useful heuristic framework' over the past two decades [Valentine, 1991b] . However the model itself has changed little, and has not been fully exploited. In particular there has been little attempt to model its temporal dynamics, despite the fact that many standard experimental approaches to face perception rely on inherently temporal phenomena, most conspicuously that of adaptation. This

is the subject of chapter 2 in which a model of temporal integration is developed, based on the method of exponential smoothing, a workhorse of the statistics and signal processing communities since the mid 20th century [Chatfield, 2003]. This method encapsulates the intuition that more recent evidence should be weighted more heavily than older evidence and allows us to conceptualise face space as a simple dynamical system where the current perceptual 'state' corresponds to a point estimate obtained by exponential smoothing. Surprisingly, this elementary model is capable of accounting for canonical and apparently complex data, such as the skewed distributions of reaction times found in 2-alternative forced choice experiments (section 2.6 and reference [Ratcliff and Smith, 2004]). Applying this simple model to face space one obtains clear predictions about how percepts should behave when a particular sequence of faces is presented. Chapter 3 describes how key predictions of this model have been scrutinised experimentally, showing results which contradict it at a fairly fundamental level, but revealing a novel psychophysical effect in so doing.

In designing and performing experiments the work of Thomas Vetter and colleagues in Basel has been invaluable [Paysan et al., 2009, Blanz and Vetter, 1999]. Their 3D generative model of face shape and texture offers one a high dimensional face space akin in many respects to the psychological construct proposed by Valentine [Valentine, 1991b], except that by implication they specify the computations defining the axes of face space, whereas Valentine is agnostic about their precise nature. This Basel Face Space (BFS) model has been leveraged throughout my investigations and has been of utility in both designing and conducting experi-

ments. It has also been pivotal to the work of many others [Houlsby et al., 2013, Learned-Miller et al., 2006, Wang and Lai, 2011]. Like this preceding work, the conclusions of chapter 3 depend on the assumption that BFS is a satisfactory approximation of human face space. In chapter 5 this assumption is itself examined, by comparing human similarity judgements to Euclidean distance within BFS, and found to be wanting in certain respects. Of course it comes as no great surprise to find that BFS is an imperfect approximation of biological face space, which after all exhibits a considerable degree of variability attributable to individual difference [Herzmann et al., 2010]. Of perhaps more interests is the finding, described in chapter 5, that the assumption of isotropy is substantially incorrect. Consistent with this, recent work has suggested some interesting possibilities for *why* face space may be anisotropic, beyond the contention that Basel Face Space is a 'bad' model of human face space. One interesting suggestion is that individuals build up complex, task-specific prior distributions, which may be highly non-Gaussian and/or anisotropic in form [Houlsby et al., 2013]. Such a possibility is attractive also since it provides a natural account of certain phenomena such as the "other race effect" [Behrman and Davey, 2001].

Finally, this thesis falls into two main sections which, while related, can be considered in relative isolation from one another. The first, comprising chapters 2 to 4, presents attempts to build a simple model of temporal integration within face space, to test this model using a psychophysical approach, and to then reconcile the data with a new model centred on a Bayesian framework. The second, comprising chapter 5, is more straight forward, in that it seeks to compare and contrast the widely used Basel

Face Model (BFM) with human face space by comparing ”BFS-distances” (measured as Euclidean distance in parameter space) with human dissimilarity judgements (measured psychophysically). Insofar as there is anything of worth contained within this document, this second section may be felt to be of more use to other researchers. However, the first section is more my own in both conception and execution and should be considered the kernel of my submission, warts and all.

2

A Normative State Space Model of Temporal Integration in Primate Face Space

In this chapter I conceptualise face space as a state space and develop a simple model of temporal integration based on exponential-smoothing, arguing that this is consistent with physiological data as

well as being supported, in a normative sense, by a sound body of statistical theory. The next section deals with some empirical prediction of the model showing, firstly, that the ”long tail” of human reaction-time distributions falls out naturally and, secondly, that the model makes firm predictions about what percept subjects should have when a series of identities are presented in quick succession. These predictions form the basis of the experiments described in chapter 3

2.1 Preliminary Comments on Modelling Objectives

Prior to describing the state space model, with which this chapter is chiefly concerned, it is important to motivate its construction in terms of what it does and does not seek to explain. In this case the principal objective is to address the question of the temporal integration of identity information in face space. The world is dynamic, and facial identity changes as one foveates from one face to another. Does face space *reset* somehow following each saccade, so that the new task is approached, as it were, with a clean slate? Or does information from one inter-saccadic epoch bleed into adjacent epochs?

It is far from obvious what the optimal strategy would be in the context of face perception, or indeed for almost any other domain, as attested to the large body of theoretical and applied work coming out of the academic

field known as *Time Series Analysis* [Chatfield, 2003]. Later in this chapter I will draw out a simple prediction of the model, namely that when two face stimuli are alternated according to some duty cycle, the percept experienced by the subject will be a linear interpolation between the two stimuli. It may be objected that this is "obvious", because for example, this is precisely what takes place in a movie. To this there are at least three relevant observations. The first is that in chapter 3 results are presented of experiments testing this very prediction of linear interpolation, yet the results do not support the linear prediction and we observe a marked bias towards the origin of face space. So, far from being obviously true, the prediction is demonstrably false. The second is that the literature on ambiguous or conflicting stimuli, such as bistable stimuli like the well known Necker cube, demonstrates that the brain typically does not interpolate linearly, but in contrast switches in a highly non-linear, stepwise and stochastic fashion between interpretations [Wilson, 2001]. The third is that the state space model presented in this chapter is capable of dealing with *any* sequence of stimuli. Suppose we have several faces presented in some random order, sampled sequentially and with replacement, say, from a multinomial distribution. What now is the obvious percept? It is very hard to say. In contrast, the state space model advanced in this chapter is capable of producing a precise prediction for any sequence of inputs.

Accordingly, the model described in this chapter is conceived as a speculative hypothesis, to be tested empirically. As alluded to in chapter 3, the model will be shown to offer a poor account of the experimental data. However, the model is not conjured out of thin air, and is based on a

widely utilised statistical approach to temporal integration, exponential smoothing [Chatfield, 2003]. This technique captures the intuitive idea that as information becomes older and older it contains less and less information about the current state of affairs. Note that this need not be the case, for example in the case of a periodic signal, but it seems a reasonable assumption for the majority of ecologically valid situations.

In focusing on the question of how information is integrated over time one is of course neglecting many other worthy questions. For example, what is the readout mechanism? Or to put it another way, how is the representation used to infer which of all possible identities is present? This question could form the basis of another thesis entirely. Notwithstanding, it is important to demonstrate that the model could in principle accommodate at least *some* readout mechanism. Therefore, following Shadlen and others [Palmer, 2005], in section 2.6, it is shown how a log-odds race model of decision making can be very naturally superimposed onto the proposed model of temporal integration [Fetsch et al., 2014]. Moreover, the model produces the classical skewed distribution seen in two alternative forced choice data [Ratcliff and McKoon, 2008]. As far as can be determined from extensive literature searches, two-alternative forced-choice reaction-time distributions in face recognition have not been explored experimentally, but there seems to be no obvious *prima facie* reason to expect facial stimuli should yield different results from the many other categories of stimuli in which the models have been tested. Within this race-to-threshold framework for decision making one sees all the complexity of decision making displayed, demonstrating how ”false identifications” can and will be

made (see figure 2.7), depending on the degree of intrinsic and perceptual noise. Additionally, by setting the decision thresholds relatively high (or equivalently, the strength of evidence low) one can easily imagine a situation in which no recognition threshold is passed, corresponding to the situation in which a new face/identity is observed.

An equally important property of any model of face space is its ability to model the process of learning, and by doing so replicate the many well established effects associated with this process in face perception. Prominent examples include distinctiveness effects, the so-called other-race effect, the caricature effect, and adaptation. Although they have not in every case been modelled out and simulated, it is clear that a race model could easily capture many of these properties, such as those distinctiveness effects discussed in some detail in section 1.3.3. It has however been shown that the model can display the properties of adaptation and priming, which are prominent in the face space literature and probably mediated by a common mechanism [Walther et al., 2013]. These effects emerge given the simple assumption that the origin, or *norm*, is determined by essentially averaging the actual faces encountered in the environment, a process referred to as *stimulus matching*. Further explication and simulation are presented in subsection 2.6.3.

To summarise, while the focus of this modelling exercise is upon the temporal integration of information, it is nevertheless important to demonstrate that the model can be extended to core experimental phenomena, such as priming and adaptation. Thus, there are many directions in which this model could be enriched and extended, but its focus, and the primary

phenomenon probed through simulation and subsequently experiment, is the temporal integration of information in face space.

2.2 Facespace as a State Space

Suppose that there exists a space within which the particular characteristics of an individuals face allow it to be positioned uniquely, known in the literature as face-space [Valentine, 1991b]. Each facial identity occupies some position in this face space and the axes of face-space correspond to those attributes or features used to differentiate faces from one another. There is debate about what these axes correspond to but for the immediate purposes it matters only that some normed vector space exists [Callier and Desoer, 1991].

Within this space it is supposed that each point corresponds to a particular identity [1], this being what is commonly meant by face space. At the origin of these axes, is the norm or average face. The scheme is illustrated for 2D in figure 2.2.

[1] And the set of points within some neighbourhood all correspond to a single identity.

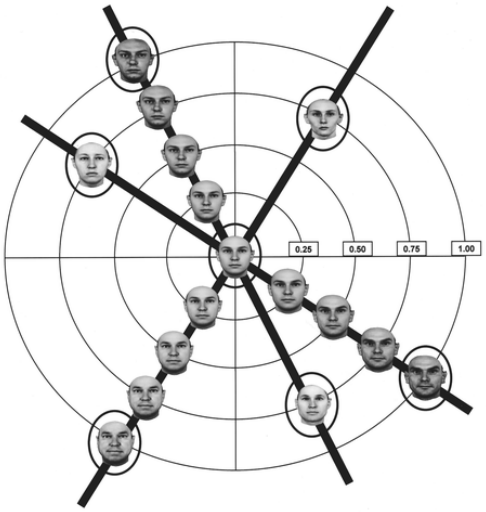


Figure 2.1: Face space in 2D. Taken from [Leopold et al., 2001]

2.3 State Space Model of Face Representation in the Primate Brain

It is likely that face space, as utilised in the brain, is very high dimensional. We suppose that this face space can be thought of as a state space for the apparatus used by the brain for face perception. To be in the condition of perceiving a face is, by assumption, to be in some state or position, P, within this space. The position, P, can equally be thought of as a particle, which can move through face space. P then denotes 'position', 'particle' and 'percept', as within this psychological model these are just different ways of looking at the same thing.

If P is required to move to some point in space corresponding to an identity when a face is seen, where should it optimally be located as a starting point? The answer depends on the prior distribution of faces.

Supposing that facial characteristics are distributed normally along each axis with a mean of $\vec{0}$, then the optimal position (i.e. that facilitating the shorted possible route to an identity-point on average) is at the mean, $\vec{0}$ in this case. In statistical terms, if we think of P as a random variable, this is the expectation of P, which is identical to the mean. The particle P should then be located at the origin in face-space, and we may note in passing that this location also corresponds to the average face, the so-called norm. Thus the norm occupies no special status in this scheme beyond its statistical significance (the mean of a multivariate Gaussian $\mathcal{N}(\vec{0}, \Sigma)$).

We can now adduce a simple scheme in which the particle P's position in space at time t is described by a vector $\mathbf{p}$. We can move P around this space by performing linear operations such as addition and multiplication on this vector.

Suppose that when a face, say face A, is present in the field of vision, at each time step a perceptual vector containing information about A's location in facespace, $\mathbf{f}_t^A$, is computed by the brain[2]. Previous perceptual vectors $\mathbf{f}_{t-1,\ldots,T}^A$ are combined so as to determine a new position for particle P. Naively, this might simply be an average over some temporal window. So,

$$P_t = \mathbf{p}_t = \frac{1}{T}\left(\mathbf{f}_{t-1}^A + \mathbf{f}_{t-2}^A + \ldots + \mathbf{f}_t^A + \ldots + \mathbf{f}_{t-T}^A\right) \tag{2.1}$$

Where t indexes the timestep, and T is the size of the temporal window,

[2]The superscript here denotes the "true" identity, or noiseless perceptual input, of a particular face, and is not an exponent. lower-case superscripts, e.g. x^n, denote exponents, while vector notation such as transpose is written in upper-case bold, e.g.$\mathbf{f}^\mathsf{T}$

in terms of timesteps, over which we average. In this formulation, when face A is presented, $\mathbf{p}$ will evolve over T timesteps to $\mathbf{f}^A$, which we might call an identity-point, as illustrated in figure 2.2.

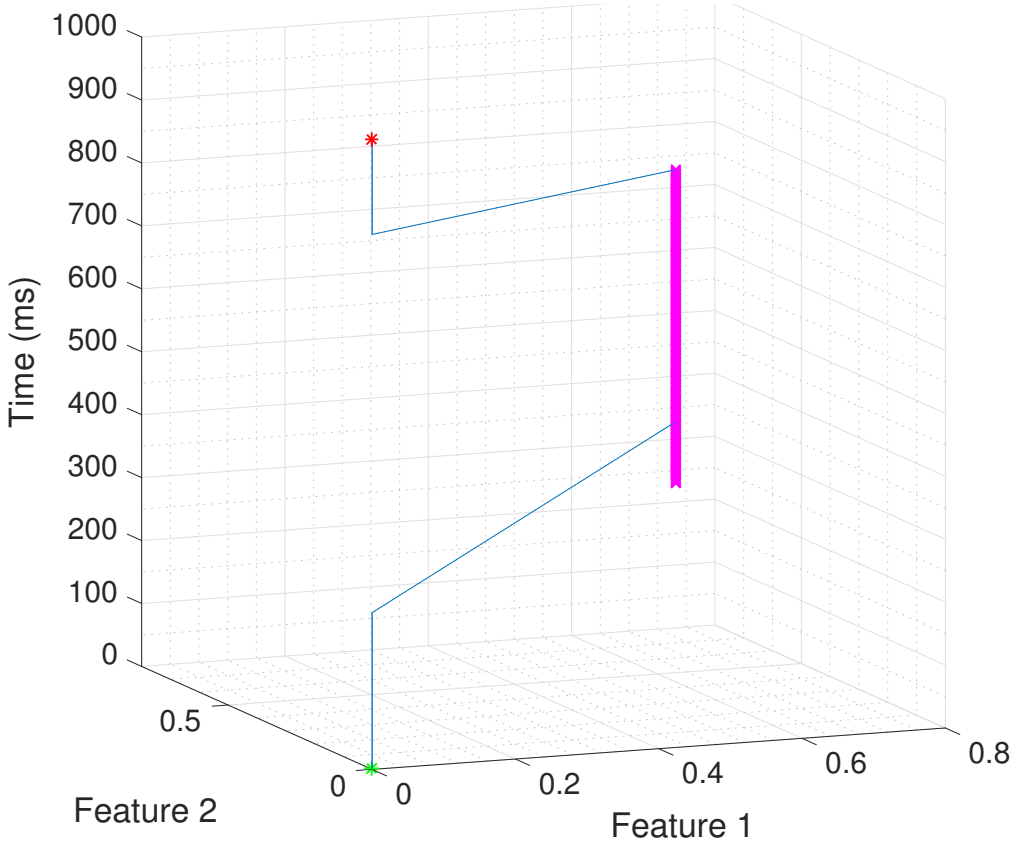


Figure 2.2: Trajectory of particle P within 2D face space over 1 second, using a moving average window of 100ms. P begins (green asterisk) at the origin, stimulus onset occurs at 250ms, lasts for 500ms after which P evolves back to the origin (in the absence of input). The attractor corresponding to face A, $\mathbf{f}^A$, is shown for the duration of the input (from 250ms to 750ms) in magenta.

So far we have supposed that information is uniformly weighted across a temporal window, known as a moving average in statistics, and reflected in the uniform $\frac{1}{T}$ factor in 2.1 weighting the current perceptual input's contribution to the estimate $\mathbf{p}_t$. A natural thought, however, is that more recent information may typically be more useful that temporally distant information, since it may correlate more strongly with the current state of the (changing) world. We can accordingly define a parameter τ, where $0 < \tau < 1$, whose effect is to weight more recent inputs to a greater or lesser degree. Using this parameter consider the following series s

$$s = \tau + (1-\tau)\tau + (1-\tau)^2\tau + (1-\tau)^3\tau + (1-\tau)^4\tau 4 + \dots \tag{2.2}$$

Observing that this is a geometric series with first term $a = \tau$ and a common ratio $r = (1-\tau)$ it follows that

$$s = \frac{a}{(1-r)} = \frac{\tau}{(1-(1-\tau))} = 1 \tag{2.3}$$

Applying these weights to a series of inputs then yields

$$\mathbf{p}_t = \tau\mathbf{f}_{t-1} + (1-\tau)\tau\mathbf{f}_{t-2} + (1-\tau)^2\tau\mathbf{f}_{t-2} + (1-\tau)^3\tau\mathbf{f}_{t-3} + \dots \tag{2.4}$$

And evidently

$$\mathbf{p}_{t-1} = \tau\mathbf{f}_{t-2} + (1-\tau)\tau\mathbf{f}_{t-2} + (1-\tau)^2\tau\mathbf{f}_{t-3} + \dots \tag{2.5}$$

So multiplying 2.5 by $(1-\tau)$ and subtracting the product from 2.4 yields

$$\mathbf{p}_t - (1-\tau)\mathbf{p}_{t-1} = \tau\mathbf{f}_{t-1} \tag{2.6}$$

From which a recursive formulation follows straight forwardly i.e.

$$\mathbf{p}_t = (1-\tau)\mathbf{p}_{t-1} + \tau\mathbf{f}_{t-1} \tag{2.7}$$

It can be appreciated that when τ is close to 0 $\mathbf{p}$ will be relatively unaffected by the most recent input, whereas when τ is close to 1 $\mathbf{p}$ is largely determined by the most recent input. Values in between naturally correspond to intermediate weightings of recent inputs versus more temporally distant inputs. Notice that this formulation requires a neural mechanism to store only one variable, which is recursively updated, a significant consideration in terms of biological plausibility. Moreover, it will become clear in section 2.3.3 why this formulation is particularly appropriate from a normative perspective.

2.3.1 Individual Faces Define Attractors and Basins of Attraction

It is an important feature of the dynamics of a proposed model of face space that there be basins of attraction corresponding to any individual face. An appealing feature of the proposed framework is that, given some input, the whole of face space becomes a basic of attraction for that identity. There remains a degree of history dependence within the model, for

example in the sense that if an input, say A, is preceded by an input, say B, which is relatively remote in face space, such as an anti-face, then it will take longer to evolve to the corresponding attractor than if B were closer to A. However there cannot arise a situation in which preceding input "locks" the future dynamics, analogous to an absorbing state in a Markov chain. This can be seen formally as follows.

For some given constant input $\mathbf{f}$, $\mathbf{p}_t$ will always be closer to $\mathbf{f}$ than $\mathbf{p}_{t-1}$ since

$$\mathbf{p}_t = \tau\mathbf{p}_{t-1} + \mathbf{f}_{t-1} - \tau\mathbf{f}_{t-1} = \tau(\mathbf{p}_{t-1} - \mathbf{f}_{t-1}) + \mathbf{f}_{t-1} \tag{2.8}$$

That is to say, $\mathbf{p}_t$ is a point falling $\tau * 100\%$ along the line connecting $\mathbf{p}_{t-1}$ and $\mathbf{f}_{t-1}$. In addition it can be seen that $\mathbf{f}$ itself is a fixed point attractor since

$$\mathbf{p}_{t-1} = \mathbf{f}_{t-1} \rightarrow \mathbf{p}_t = \tau\mathbf{f}_{t-1} + (1-\tau)\mathbf{f}_{t-1} = \mathbf{f}_{t-1} \tag{2.9}$$

2.3.2 Noise and Variability

How does the model behave if, instead of a steady input, the model receives noisy input? Suppose that at each timestep $\mathbf{f}_t \sim \mathcal{N}(\vec{0}, \Sigma_e)$

Then by equation 2.7

$$\mathbf{p}_t = (1-\tau)\mathbf{f}_t + \tau(1-\tau)\mathbf{f}_{-1} + \tau^2(1-\tau)\mathbf{f}_{-2} + \dots \tag{2.10}$$

So the covariance is given by

$$\mathrm{Cov}[\mathbf{p}_0] = \Sigma_P = (1-\tau)^2\Sigma_e + \tau^2(1-\tau)^2\Sigma_e + \tau^4(1-\tau)^2\Sigma_e + \dots \quad (2.11)$$

Σ_e can then be removed from the summation leaving a geometric series with the first term $a = (1-\tau)^2$ and common ratio τ^2, and yielding the following expression, [3]

$$\Sigma_P = \Sigma_e \cdot \sum_{t=0}^{\infty}(1-\tau)^2\tau^{2t} = \Sigma_e \cdot \frac{(1-\tau)^2}{1-\tau^2} \quad (2.12)$$

It should now be possible to appreciate some of the underlying rationale for this scheme. Should the value of τ be set close to 0 the current estimate is largely determined by the most recent input, the payoff being sensitivity and meaning that the current estimate will track changes in the target variable with very little lag. It will however be very sensitive to noise in the input (i.e. by equation 2.12 the variance will be relatively large). We can combat the sensitivity to noise by setting the value of τ be close to 1, so that temporally distant inputs are weighted more heavily, but the (relatively low variance) estimate will then lag behind the true value of the tracked variable. Thus there is an inevitable tension between the desire for sensitivity and for noise resistance in our estimate, which is illustrated in simulation in figure 2.4.

[3] Note that the sigma notation is overloaded in that it is being used to denote both a covariance matrix (the smaller) and summation (the larger).

2.3.3 Statistical Motivation of the State Space Model

Although we have so far developed the SSM in a relatively intuitive fashion, we will now provide a statistical motivation by showing that it is equivalent to a widely used technique from time series analysis called exponential smoothing. Suppose we have a time series for a variable x, from time $t = 1, \ldots, T$.

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T \tag{2.13}$$

In order to obtain an estimate of the current value of x simple exponential smoothing [Chatfield, 2003] mandates the followsing recursive relation.

$$\hat{\mathbf{x}}_{t+1} = \alpha \mathbf{x}_t + (1 - \alpha)\hat{\mathbf{x}}_t \tag{2.14}$$

Comparison of equations 2.14 and 2.7 will make clear that the model we have proposed for the integration of temporal information in facespace is completely equivalent to exponential smoothing, where $\tau = 1 - \alpha$. It may also be noted that exponential smoothing is a special case of a Kalman filter [Chatfield, 2003] placing this approach within the broader context of a sound probabilistic framework. This makes a good deal of sense since the brain is here challenged with a similar task to a Kalman filter, namely to estimate and track the value of a potentially dynamic variable using a noisy time series.

2.4 Continuous Dynamics for Face Space

The preceding description of the state space model of face space dealt in discrete time steps, yet biological face space presumably possesses continuous dynamics. It is straightforward to extract a continuous model from the discrete one so far described, as follows. The vector describing the change in the position $\mathbf{p}$, for some constant input $\mathbf{f}$, over a time step of size δt is,

$$\mathbf{p}_t - \mathbf{p}_{t-\delta t} = \tau \mathbf{p}_{t-\delta t} + (1-\tau)\mathbf{f} - \mathbf{p}_{t-\delta t} \tag{2.15}$$

So in the limit as $t \to 0$,

$$\nabla \mathbf{p} = \lim_{\delta t \to 0} \mathbf{p}_t - \mathbf{p}_{t-\delta t} = (1-\tau)(\mathbf{f} - \mathbf{p}_t) \tag{2.16}$$

Which can then straight forwardly be integrated to yield the general solution

$$\mathbf{p}_t = \mathbf{f} + \mathbf{C}\exp(-t(1-\tau)) \tag{2.17}$$

Where $\mathbf{C}$ is a constant vector.

If we then suppose that the position at time $t = 0$ is $\mathbf{p}_0$ (initial conditions) then

$$\mathbf{C} = \mathbf{p}_0 - \mathbf{f} \tag{2.18}$$

So that the particular solution describing the continuous dynamics within

face space for a given input $\mathbf{f}$ and a starting position $\mathbf{p}_0$ is given as follows.

$$\mathbf{p}_t = \mathbf{f} + (\mathbf{p}_0 - \mathbf{f}) \exp(-t(1-\tau)) \tag{2.19}$$

This equation describes the continuous dynamics of a particle in face space. From our derivation it can also be appreciated that a physical system with these dynamics, such as a population of face selective neurons, would be in essence computing the continuous, exponentially smoothed estimate of face identity. As has been observed previously this provides a normative rationale for the observation that electrical recordings of neurons do indeed display exponential dynamics [Yu and Cohen, 2009].

2.5 Illustrations of the State Space Model

Some illustrations of the model should help clarify its operation. Figure 2.3 shows the trajectory of $\mathbf{p}$ from the origin, the default initialisation point, to an attractor, corresponding to a particular face.

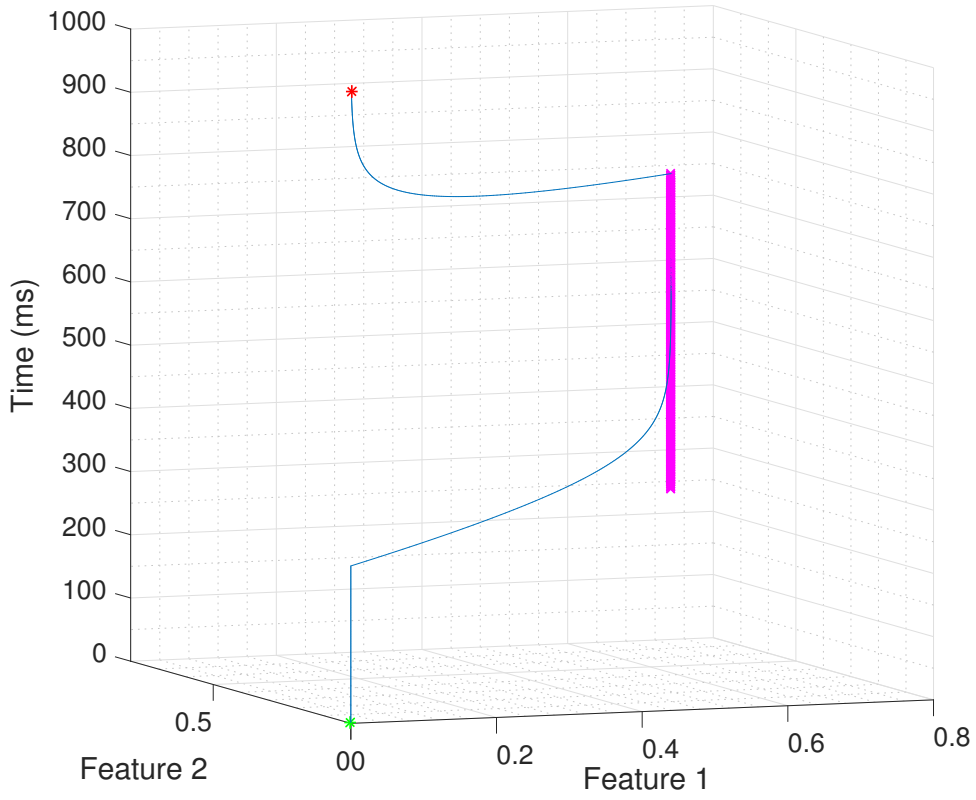


Figure 2.3: Illustrates the temporal evolution over time in 2D state-space without noise. $\mathbf{p}$ begins at the origin (the norm) and evolves towards an attractor, marked in magenta at each timestep, according to equation 2.7. The characteristic shape of the trajectory derives from the fact that this is essentially a form of exponential smoothing, so that the particle $\mathbf{p}$ "decays" exponentially and asymptotically to the attractor and then back to the origin at the offset of the stimulus. For the purposes of illustrating the form of the dynamics τ is set rather high, at 0.975. cf. figure 2.2

Figure 2.4 illustrates a similar scenario but with the addition of noise and six different values of τ. In the noisy condition with τ set to a relatively low value P does not quite converge to the fixed point but instead ″rolls″ noisily around in the basin of attraction. This effect becomes comparatively smaller as the value of τ increases towards 1, going from top left to bottom right in the subplots of figure 2.4.

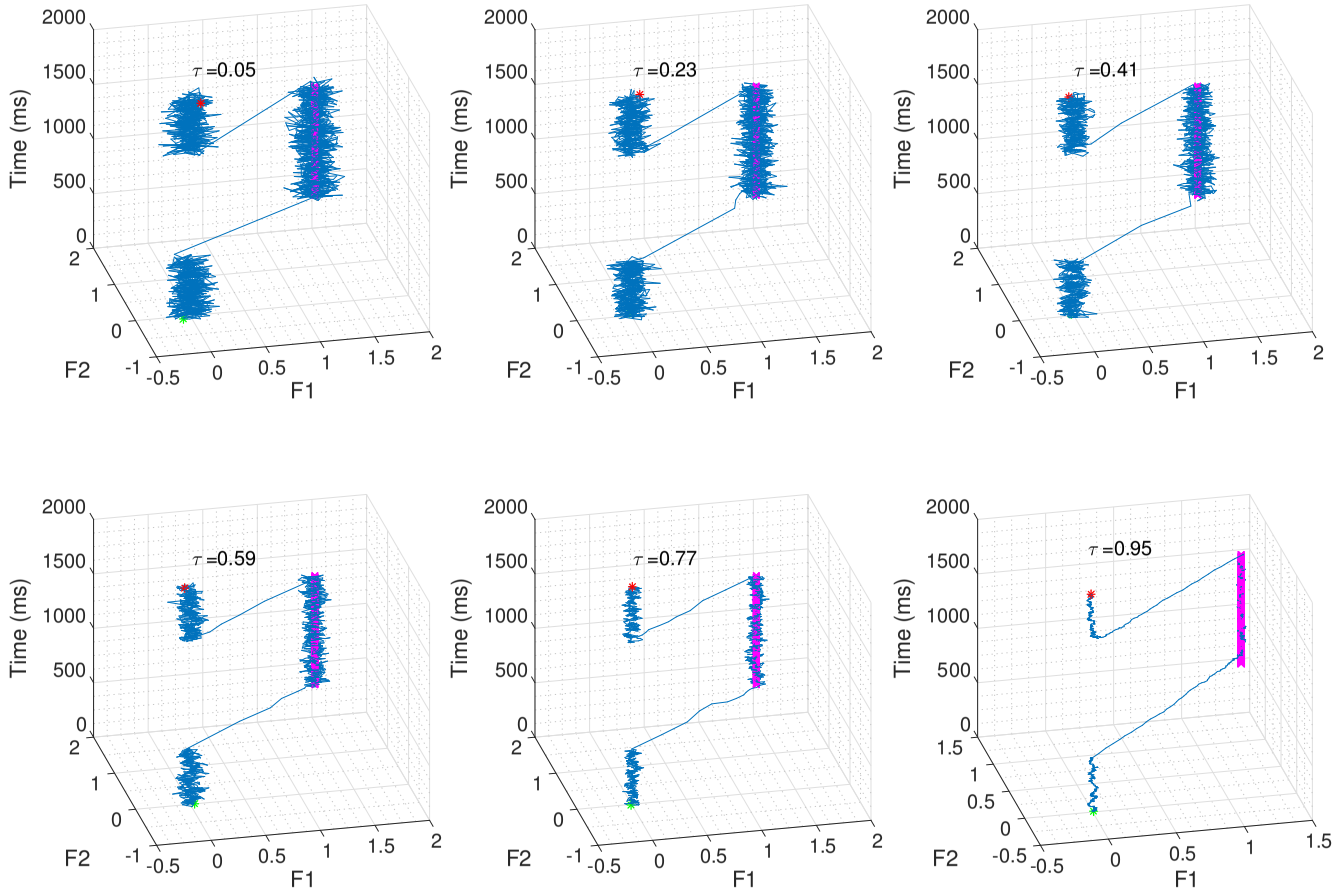


Figure 2.4: Illustrates the noise cancelling properties of exponential smoothing. From top left to bottom right increasing values of τ (the smoothing parameter) result in a smoother trajectory to and from an attractor as well as better approximation to the attractor itself over the duration of stimulus presentation.

In figure 2.5 we now have two stimuli presented sequentially, switching after 1000 timesteps, illustarting how the system re-converges on presentation of a new input to the system.

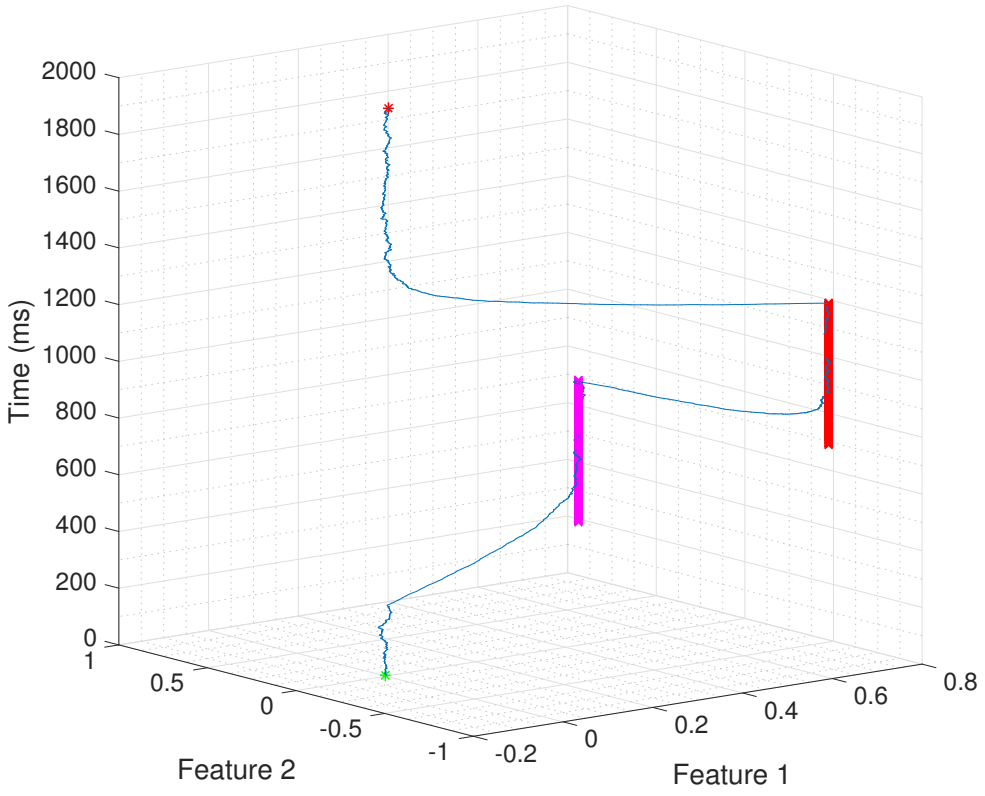


Figure 2.5: Illustrates the temporal evolution of the model with a small amount of noise and two stimuli (marked by magenta and red) presented successively. P begins at the origin (the norm) and evolves towards an attractor corresponding to the first stimulus (magenta), remains in its neighbourhood until time point 750, when the second stimulus is presented, to which the system then evolves. At the offset of the second stimulus (red) P then returns to the origin where, notwithstanding ongoing noise, it remains.

2.6 Experimental Predictions for the State Space Model

2.6.1 Distribution of Reaction Times

One of the principal experimental validations of the class of models known as drift diffusion models is their ability to predict the positively skewed distribution of reaction times in human subjects during, for example, forced choice experiments [Romo, 2012]. This is illustrated in panel b of figure 2.6.

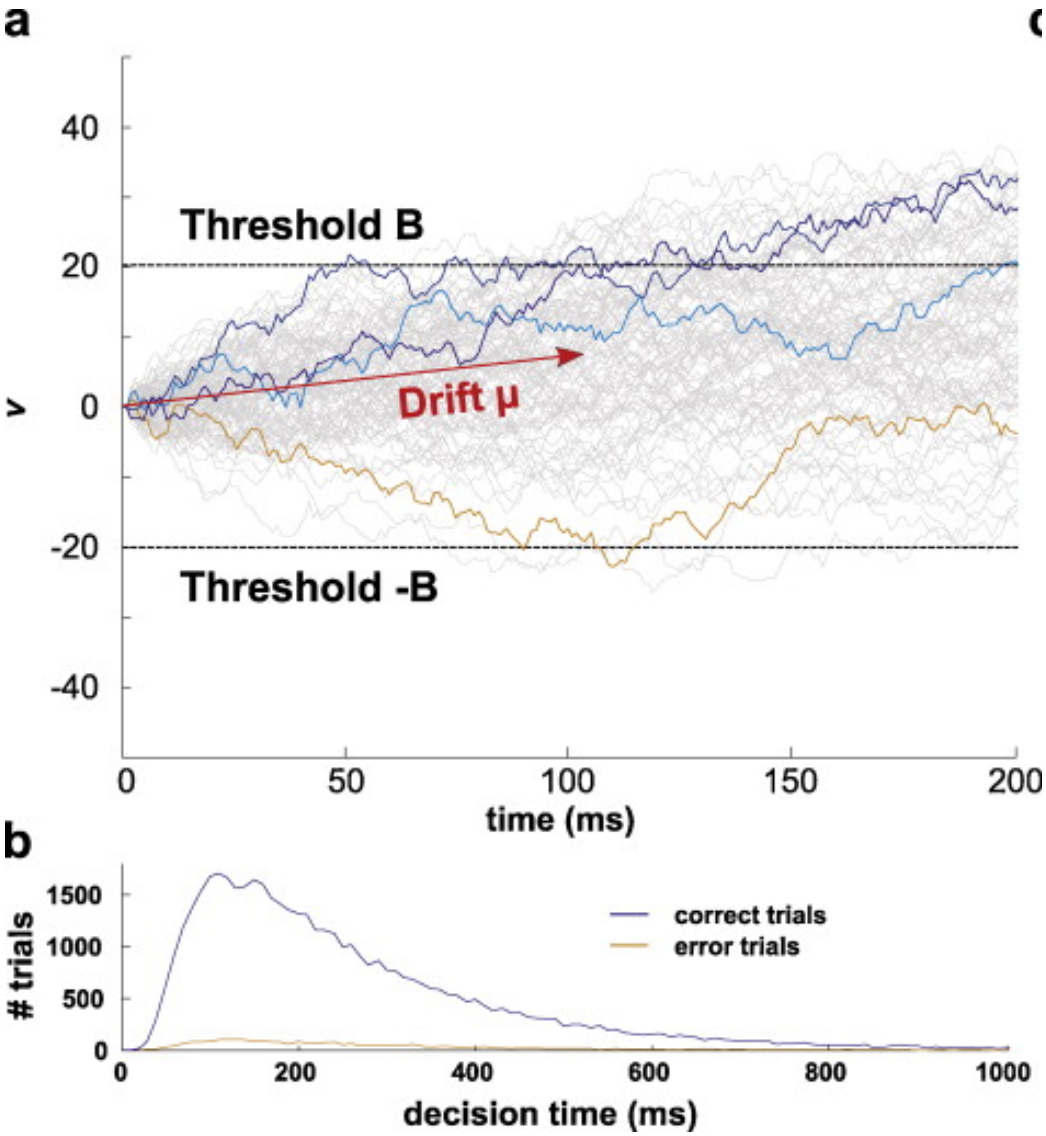


Figure 2.6: The sDDM model and the distribution of reaction times. Panel a shows the trajectory of evidence accumulation for many independent trials. Several are depicted in colour for illustration. The signal is depicted by the red vector marked *Drift* μ and its strength corresponds to the steepness of its gradient. The threshold at which a decision is taken is marked by a pair of black lines equidistant from the origin. Note that the level at which the thresholds are set is arbitrary in the absence of an objective function and indeed the positive and negative threshold might conceivably be of different magnitudes. Panel b shows a frequency plot of correct and incorrect trials, i.e. trials in which threshold B (correct) or threshold -B (incorrect) was crossed first, versus reaction time. Of particular note is the long tail, or right skew, present in both distributions. The fact that drift diffusion models of decision making reproduce this signature form, a right skew, as a natural consequence of their dynamics is considered by many to be a cogent piece of evidence in their favour. From [Romo, 2012]

The state space model can likewise reproduce these features of human decision making. We integrate evidence across time by calculating an odds ratio in the following way. If we suppose Gaussian probability distributions centred on the attractors, $\mathbf{f}_A$ and $\mathbf{f}_B$, then we can calculate a conditional probability (density) for $\mathbf{p}_t$ at each time-step t. i.e.

$$p(\mathbf{p}_t|\mathbf{f}_A) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\mathbf{p}_t - \mathbf{f}_A)^2/2\sigma^2} \tag{2.20}$$

And likewise, *mutatis mutandis*, for $p(\mathbf{p}_t|\mathbf{f}_B)$. The log of the ratio of these two probabilities

$$\ln\left(\frac{P(\mathbf{p}_t|\mathbf{f}_A)}{P(\mathbf{p}_t|\mathbf{f}_B)}\right) \tag{2.21}$$

at time t will therefore yield a positive or negative value depending on P's relative proximity to the two attractors. This quantity, call it d, can then be summed across time-steps and a 'decision' triggered when some arbitrary decision threshold, $\pm b$, is crossed.

$$d_t = \sum_0^t \ln\left(\frac{P(\mathbf{p}_t|\mathbf{f}_A)}{P(\mathbf{p}_t|\mathbf{f}_B)}\right) \tag{2.22}$$

This process is illustrated in figure 2.7 for 100,000 trials within the model.

We can now reproduce the skewed distribution within our simulation. Figure 2.8 illustrates a situation in which face A is presented in the presence of much noise. In this situation the noise has a significant impact on P's trajectory, one of which is traced for illustration. It is worth noting that the critical computation here, determining when the decision criterion has

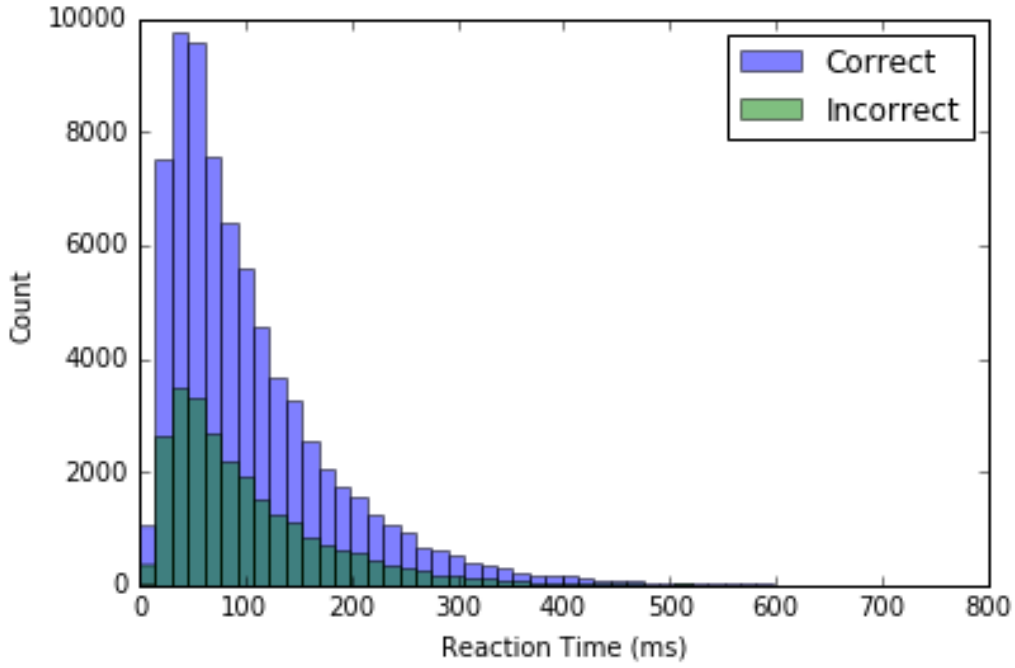


Figure 2.7: Reaction time distribution produced by the state space model for a forced choice between a face (A) and an anti-face (B). The true signal, i.e the input to the model, is A but this is corrupted by significant noise, so there is the possibility that the system is driven past the decision threshold for B before that of A. Thus over the 100,000 simulated trials we obtain a distribution of correct and incorrect trials, which of note displays the characteristic long tail (right skew) found in human psychophysical data across multiple tasks [Usher and McClelland, 2001].

been surpassed in one direction or another, could be implemented in a biologically plausible way by a competitive neural network in which two populations of neurons are driven by the two stimuli (faces) but mutually inhibit one another [Usher and McClelland, 2001].

The distributions of reaction times, for correct and incorrect decisions, are plotted in figure 2.7. While the results in figure 2.7 represent a simulation with various hand-set parameters, the general form of the distribution is invariant. This prediction can be mapped directly onto psychophysics,

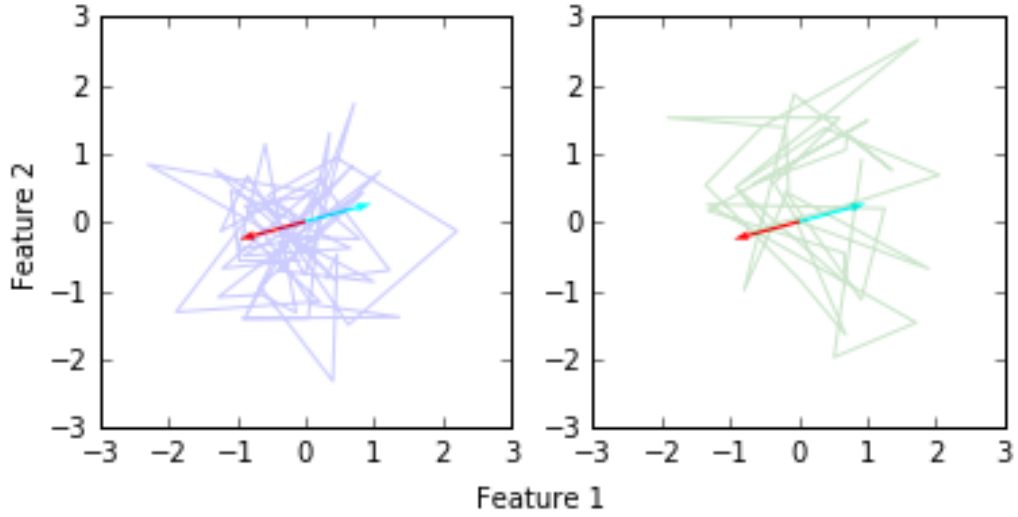


Figure 2.8: Two example trajectories for a simulate two-alternative forced choice task in the exponential state space model. In the presence of significant noise, as in this case, the trajectory of p is dominated by noise. Combined with the noise there is however a weak signal in the direction of the "true" input face (red), which needs to differentiated from the corresponding anti-face (cyan). cf. figure 2.9

and to my knowledge has not been tested in the literature. In particular, it is not adequate to merely present degraded faces and measure reaction times. Since the level of internal noise in the neuronal accumulator is unknown, and may even be close to 0 [Brunton et al., 2013], it is essential that the stimulus be corrupted with significant noise, where significant means of a comparable magnitude to the signal.

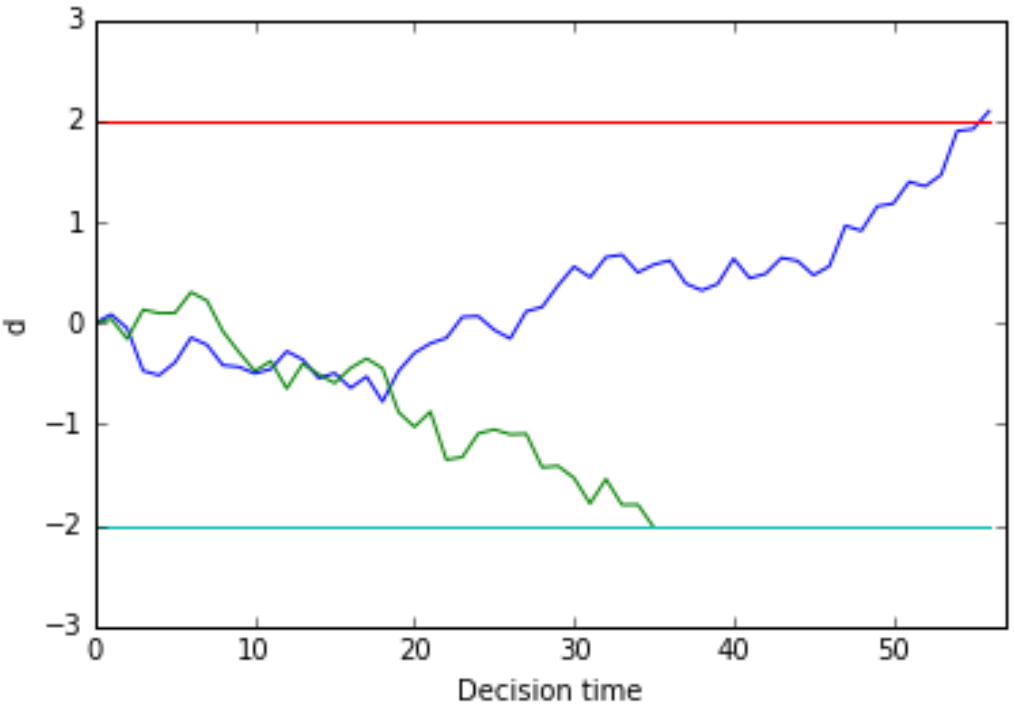


Figure 2.9: Shown is the process of evidence accumulation for two simulated trials, terminating when decision boundary $\pm b$ is exceeded (marked by the red and cyan lines), for the SSM using a sequential probability ratio as a decision criterion. Compare the two runs of the simulation shown here to the two drift-diffusion random walks shown in figure 2.8

2.6.2 Interleaved Presentation of Multiple Faces

What should be the effect of alternating between two different faces? If the interval between the two is sufficiently long then it is clear that the model will simply evolve to one attractor and then the other sequentially. Likewise, perceptually one simply sees one face and then the other if the period is sufficiently long. However, if the alternations are rapid should one perceive a face intermediate to those two presented, or alternatively should the percept alternate stochastically between the two as is found with bistable stimuli such as the Necker cube? The predictions of the exponential state space model can be drawn out straightforwardly. Consider some arbitrary sequence of two faces, $\mathbf{f}^A$ and $\mathbf{f}^B$, beginning as say $B, A, A, B, \ldots$ and continuing in a random fashion *ad infinitum*. In the model this can represented as follows,

$$\mathbf{p}_t = (1-\tau)\mathbf{f}_t^B + \tau(1-\tau)\mathbf{f}_{-1}^A + \tau^2(1-\tau)\mathbf{f}_{-2}^A + \tau^3(1-\tau)\mathbf{f}_{-3}^A + \ldots \quad (2.23)$$

It can be seen that whatever the sequence, the weights determining the contribution of each element of the sum to the particle position at time t, $\mathbf{p_t}$, must sum to 1 since,

$$s = (1-\tau) + \tau(1-\tau) + \tau^2(1-\tau) + \tau^3(1-\tau) + \ldots = \frac{1-\tau}{1-\tau} = 1 \quad (2.24)$$

I.e. s is a geometric series with $a = 1-\tau$ and common ratio $r = \tau$.

We can then decompose s into a sum of weights for A and a sum of weights for B,

$$s = s^A + s^B = 1 \tag{2.25}$$

And finally, using linearity,

$$s\mathbf{p}_t = s^A\mathbf{f}^A + s^B\mathbf{f}^B = \mathbf{p}_t \tag{2.26}$$

This will necessarily be a point on the line connecting $\mathbf{f}^A$ and $\mathbf{f}^B$, since in general for any vectors $\mathbf{a}$ and $\mathbf{b}$, letting $\mathbf{c} = \mathbf{b} - \mathbf{a}$ and $0 < w < 1$, a point $\mathbf{x}$ falls on the line connecting $\mathbf{a}$ and $\mathbf{b}$ i.i.f.

$$\mathbf{x} = \mathbf{a} + w\mathbf{c} = \mathbf{a} + w\mathbf{b} - w\mathbf{a} = (1 - w)\mathbf{a} + w\mathbf{b} \tag{2.27}$$

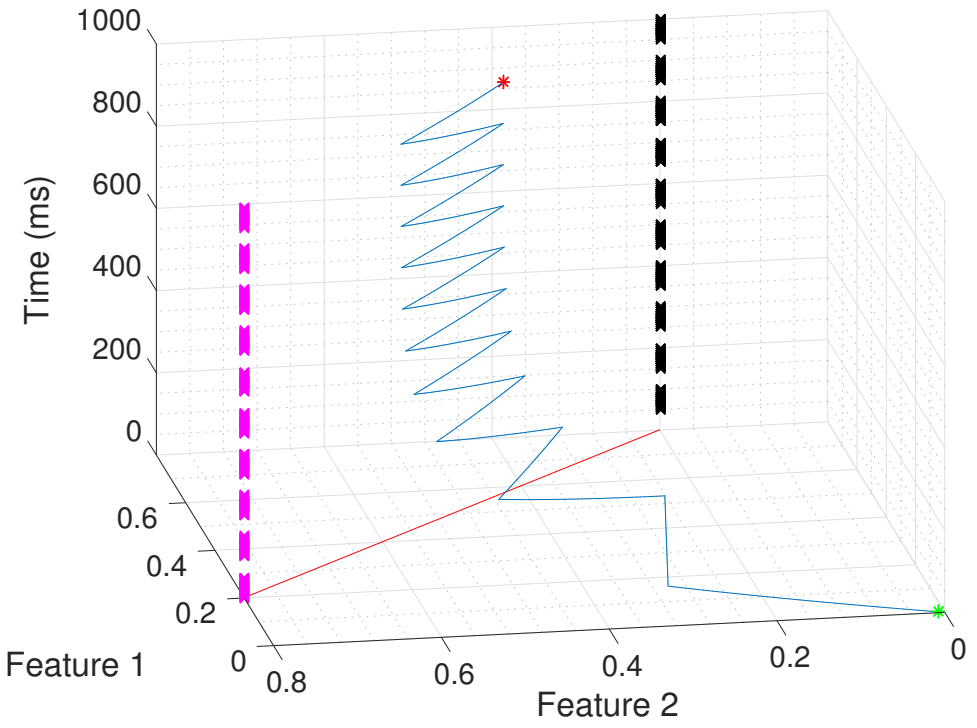


Figure 2.10: Where two faces (correspondong attractors marked in black and magenta) are interleaved so that they are presented in an alternating pattern. $\mathbf{p}_t$ begins at the origin (green asterisk) and rapidly evolved towards a 2-cycle in which it is 'suspended' between the two attractors, situated at a point on the morph line (marked in red for t=0) between the two attractors A and B. In this example the duty cycle for face A is 50%, meaning that $\mathbf{p}_t$'s average position is halfway between the two attractors. Varying the duty cycle of A between 0 and 100% results in a linear interpolation between A and B with respect to the average position over time.

It is a trivial extension of this result to demonstrate that for any sequence consisting of members of a set of M (perceptual) vectors, $\mathbf{p}_t$ will always occupy a point in an $M-1$ hyperplane within $\mathbb{R}^N$ (perceptual) space.

We can now adduce a method for "supsending" $\mathbf{p}$ at some arbitrary point on the line between two faces as follows. Suppose at each timestep we present either face A or B according to a Bernoulli distribution with parameter p, i.e. $\mathbf{f}_t \sim \mathrm{Bern}\,(p)$. Then the expectation of $\mathbf{p}$ is

$$\mathbb{E}[\mathbf{p}] = \sum_{t=0}^{\infty} \left(p\mathbf{f}^A(1-\tau)\tau^t + (1-p)\mathbf{f}^B(1-\tau)\tau^t \right) = p\mathbf{f}^A + (1-p)\mathbf{f}^B \quad (2.28)$$

the RHS of which, by 2.25, is formally equivalent to the middle expression in equation 2.26. So, setting p to some value between 0 and 1 will 'suspend' $\mathbf{p}$ a certain proportion, p, along the line from face A to B, although for any actual sampled sequence $\mathbf{p}$ will typically shuttle stochastically back and forth across the point $\mathbb{E}[\mathbf{p}]$. Again, it is straight-forward to generalise this result to the case of multiple faces, where the multinomial distribution plays the role of the Bernoulli distribution. Nor need we know from which probability distribution a sequence has been generated. For any given sequence for $t = 0, \ldots T$ we can simply compute the expected value of $\mathbf{p}_0$ empirically as

$$\mathbf{p}_0 = \sum_{t=0}^{T} \mathbf{f}_t^i (1-\tau)\tau^t \quad (2.29)$$

Where i indexes over different perceptual vectors and i denotes the

value at time t.

2.6.3 Stimulus Matching Accounts for Adaptation and Priming in State Space Model

It has already been observed in section 2.3 that the optimal position for P to be located in face space is the origin, corresponding to the norm or average face. How could a brain compute this location? One possibility is that the brain moves the origin of the approximate representation to this location by taking a weighted average of the faces it is exposed to, where the weighting is determined by some function of the duration and/or frequency of presentation. That could be accomplished by a simple rule such as the following: if a face is present, move the origin of neural face-space towards the location of the current stimulus. This could of course be implemented at many time-scales, but for now we consider a single timescale.

In an influential adaptation study by Leopold and colleagues [Leopold et al., 2001] subjects were tasked with identifying presented faces as a particular face, or its anti-face. After adaptation to a face subjects were presented with the norm, and were found to be more likely to classify it as the anti-face, despite the norm being by definition equidistant in face-space from both the face and anti-face. Adaptation to the anti-face resulted in the opposite identification bias. In this sense adaptation can be said to have a repulsive effect, since identification is biased away from the adapted stimulus. In contrast, when a subject is primed to a stimulus such as a face,

they are subsequently quicker to identify the face and do so more accurately than they otherwise would. In this sense priming can be seen as an attractive effect, since the stimulus is recognised faster and more accurately post-exposure. It can seem therefore *prima facie* that priming and adaptation are antagonistic to one another in that exposure to a stimulus apparently both inhibits *and* facilitates future identification.

Figure 2.11 illustrates how stimulus-matching can explain both priming and adaptation simultaneously, consistent with evidence supporting a common mechanism underlying both phenomena [Walther et al., 2013]. The simulation, illustrated in figure 2.11, mimics the design of Leopold and colleagues [Leopold et al., 2001] and shows that the average face, the norm, is found on the opposite side of the origin once adaptation with stimulus-matching has occurred, explaining why the norm is more likely to be identified as the anti-face post adaptation since, in the representational space, the norm is no longer at the origin, but effectively constitutes a version of the anti-face. The priming effect, meanwhile, is explained by the fact that the adapted stimulus now occupies a position closer to the origin than pre-adaptation, and so the system evolves to the corresponding attractor in fewer time steps when initialised from the norm (i.e. the stimulus is recognised more rapidly).

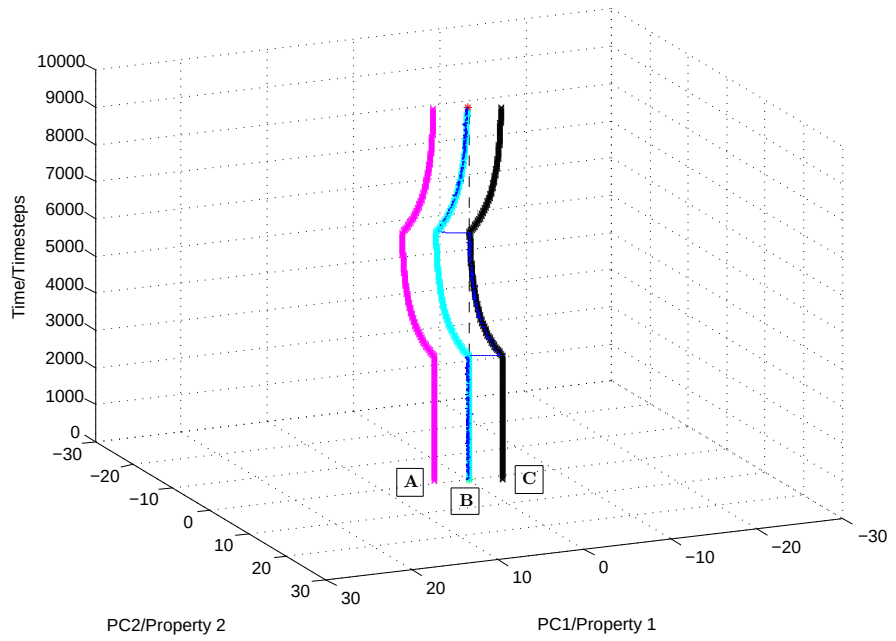


Figure 2.11: Illustration of adaptation to stimulus C (attractor in black). With no stimulus present P remains (noisily) at the origin, which initially coincides with the average face (attractor B in cyan). When stimulus C is presented (at t=3,000) P evolves towards $\mathbf{p}_{eq}^{C}$ and remains there for the duration of the presentation (i.e. from t=3,000 to t=5,999), during which time the origin of the representation shifts towards $\mathbf{p}_{eq}^{C}$ at a constant rate. Note that subsequent to this the average face (B, in cyan) is now on the opposite side of the norm from C, and on the same side of the norm as the anti-face (A). Thus B is more likely to be identified as a version of A than of B, at least under a norm based interpretation (i.e. a repulsive effect of adaptation). In contrast C is now closer to the norm than previously, and so is recognised more rapidly since the system evolve to the corresponding attractor in fewer time steps. (n.b. A shift in the origin described by vector $\mathbf{s}$ is equivalent to adding the vector $-1 \times \mathbf{s}$ to all the data points and keeping the origin stationary, which is how the shift is depicted here).

Thus, to summarise this section, by making the simple assumption that the origin or norm in face space is determined by sampling from those faces actually experienced and taking the smoothed mean of that sample, the model can immediately reproduce the central features of priming and adaptation. Naturally, a very similar approach could be adopted to account for the so-called *other-race effect* and many other psychophysical phenomena.

2.7 Conclusion

I believe this simple normative model has much to offer, by way of explanatory accounts of current results, but also in making novel predictions for new experiments. However, there are certain phenomena, such as the FFDE (Flashed Faces Distortion Effect) [Tangen et al., 2011] and my own results described in chapter 3 which cannot, I believe even in principle, be accounted for by this model. Perhaps then the most generous construction that can be put on it is that it represents a clear example of Popperian science [Popper, 2002], yielding strong experimental predictions, which (typically) turn out to be wrong.

3

Experimental tests of the state space model of facespace

This chapter concerns the design, methods and results of experiments aimed at testing the predictions of the exponential model of temporal integration, arrived at in chapter 2. The two experiments performed are referred to as the DE (Dynamic Experiment) and CE (Contrast Experiment). The results obtained in the DE contradict the predictions

of exponential model and the results of the CE argue strongly that this deviation from predictions is somehow a consequence of the dynamic nature of the stimuli used. These results set the scene for an alternative Bayesian model of inference developed in chapter 4

3.1 Introduction

The purpose of the experiments described in this chapter was to test the predictions of the normative model of face space dynamics developed in chapter 2. To briefly recap, the model predicted that a stimulus consisting of two rapidly alternating faces would yield an intermediate percept, representing an interpolation along the morph line connecting the two faces in face space. Moreover, it was shown that the position of the interpolation on the morph line should be dictated by the duty cycle of the stimulus. As discussed in section 2.1, it is not "obvious" that the model should yield a linear interpolation because (1) it turns out (as will be seen) that a linear interpolation is falsified experimentally (i.e. far from being obviously true, it is demonstrably false)(2) the model can cater for input of arbitrary complexity and (3) the literature on ambiguous stimuli provides ample examples of profoundly non-linear percepts, such as the bistable Necker cube. Finally, note that the choice of exponential smoothing is not arbitrary, versus some other form of smoothing say. On the contrary, as outlined in sections 2.4 and 2.3.3 there are sound normative [Chatfield, 2003] and neurophysiological [Yu and Cohen, 2009] grounds for this hypothesis.

The fundamental hypothesis at stake then is as follows: does psychophys-

ical data support a basic prediction yielded by supposing that face spaces behaves like a state space in which an estimate of the current state is made using exponential smoothing? In Results, section 3.3.2, it will be shown convincingly that the actual findings were quite different from those predicted, and require the rejection or revision of the exponentially smoothed state space model. Indeed, In the next chapter an alternative model is developed which does indeed account for some salient features of the experimental findings, although at a considerable cost in terms of parsimony. Notwithstanding, this supports the possibility that inference in face space is, like much lower level perceptual inference, Bayesian in character.

3.2 Methods: General Approach

Experiments were performed on 20 subjects with normal vision and normal face perception. Experiments were written using the programming language Matlab (R2010) and the package PsychToolBox [Brainard, 1997], and were conducted in the MRC CBU in Cambridge, UK. Subjects, recruited from the MRC CBU volunteer panel and ranging in age between 18 and 65, were paid £7 per hour, apart from four who were members of the Kriegeskorte lab and were paid nothing. Experiments were run on the Unit's dedicated experimental desktop computers, using 25" by 43" monitors (2560 x 1440 resolution), and running Windows 7. Subjects performed the experiments sat at a normal desk, which the monitor at eye level in front of them and at a distance of approximately 60cm, with the keyboard on the desk between the monitor and the subject. There were two distinct

experiments run for each subject in separate sessions, which I will refer to as 1. the Dynamic Experiment (DE) (see subsection 3.2.1) and 2. the Contrast Experiment (CE) (see subsection 3.2.2). In both the DE and the CE experiments followed an interleaved design, whereby on each trial a stimulus was presented according to parameters pseudorandomly sampled from tables 3.2.1 and 3.2 respectively. Each trial parameter setting, corresponding to a row in tables 3.2.1 and 3.2, was presented 10 times per subject, resulting in 220 trials per session for the DE and 110 trials per session for the CE. In both experiments the subject was required to manipulate a 12cm x13cm "matcher" face (which he could do with the keyboard 'o' (left), 'p' (right), 'q'(more centripetal), 'a' (less centripetal)) on the right so as to resemble a 12cm x13cm "target" face on the left, which on any particular trial consisted of two faces "merged" by one of two methods, corresponding to the DE and CE. For both the DE and the CE participants were simply instructed to do their best to make the two stimuli (on the left and right of the display) match as closely as possible. Subjects ere shown the four keys 'o' (left), 'p' (right), 'q'(more centripetal), 'a' (less centripetal) used to manipulate the matcher stimulus, and a fifth key 'd', which terminated the trial. There was no time limit placed on subjects, but subjects typically took 20-30 seconds per trial. Further details about the DE and CE are given in section 3.2.1 and 3.2.2. An example screen shot of the experimental display is shown in figure 3.1.

Figure 3.1: A (static) screenshot of the experimental display. In any condition both the left and the right stimulus are variable. The stimulus on the left, the target stimulus, "merges" two identities, either by rapidly alternating between two identities (the *dynamic* experiment, DE) or by superimposing two identities with differential contrast (the *contrast* experiment, CE). The dominance of one identity over the other is varied systematically according to the parameters of the condition (see table 3.2.1). In either case the stimulus on the right, the matcher, can be controlled by the subject so as to match the target stimulus. This amounts to stepping around a grid-sampled 2D slice through face space which contains the norm face as well as the two faces used in the target stimulus (see figures 3.2 and 3.3)

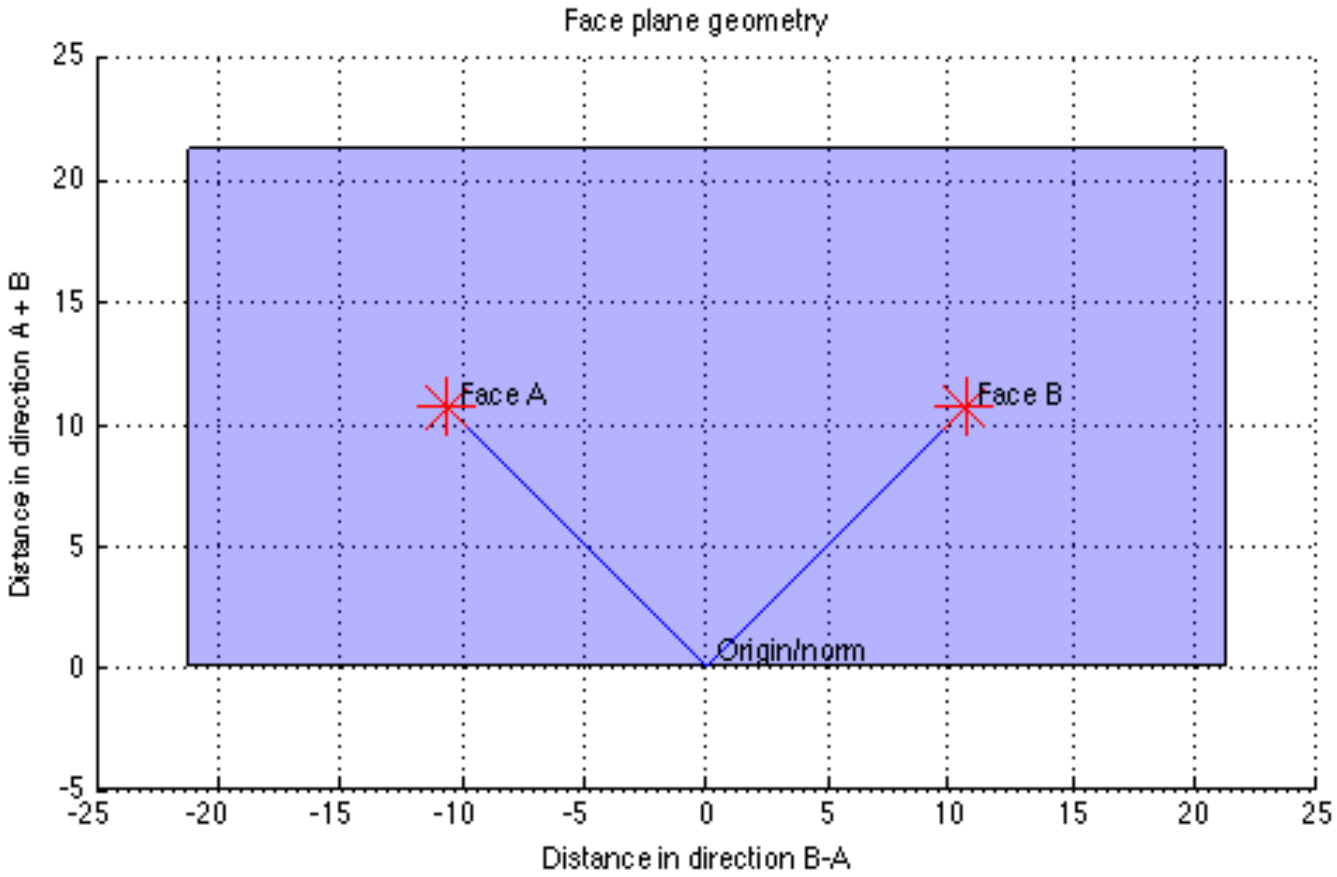


Figure 3.2: A representation of the geometry of the two dimensional slice through face space around which subjects navigate in attempting to match the target stimulus. The two faces corresponding to the target stimulus are marked with red asterisks, and relation to the origin, "norm" or zero-vector face is illustrated by two blue lines emanating from the same. Compare this to figure 3.3

3.2.1 Methods: Dynamic Experiment

As described in the general approach, section 3.2, in this experiment, the CE, the subject was on each trial presented with a display on which two stimuli (faces) were displayed side by side. One stimulus consisted of a rapidly alternating pair of faces, with a period of either 100ms or 200ms. The other "stimulus" consisted of a single face which could be altered by the subject by pressing keys 'o' (left), 'p' (right), 'q'(centrifugal), 'a' (centripetal). By means of these 4 keys and beginning from a randomly chosen initial point the subject was able to navigate around a randomly sampled rectangular, 2D grid of faces. The grid was constrained to contain the average or norm face and the pair of faces being rapidly alternated in the first stimulus. When the subject felt that he had found the best match from among this grid he terminated the trial by pressing 'd' (for *done*) and progressed to the next trial. The parameters of the next trial were then drawn pseudo-randomly from table 3.2.1, so that this was an interleaved design (as opposed to block design).

The temporal separation of the two faces was an important aspect of the experimental design, and one that originates from the modelling described in subsection 2.6.2. In the dynamic condition the two faces were rapidly alternated on a timescale of $10 - 180$ms. To unpack this, at duty cycles of 0 and 100% a single face was present, since the "second face" (at duty cycle 0%) was presented for a duration of 0ms on each cycle. The overall period of the alternation (i.e. the duration from the onset of A until the offset of B) was either 100ms or 200ms, but for each period the actual

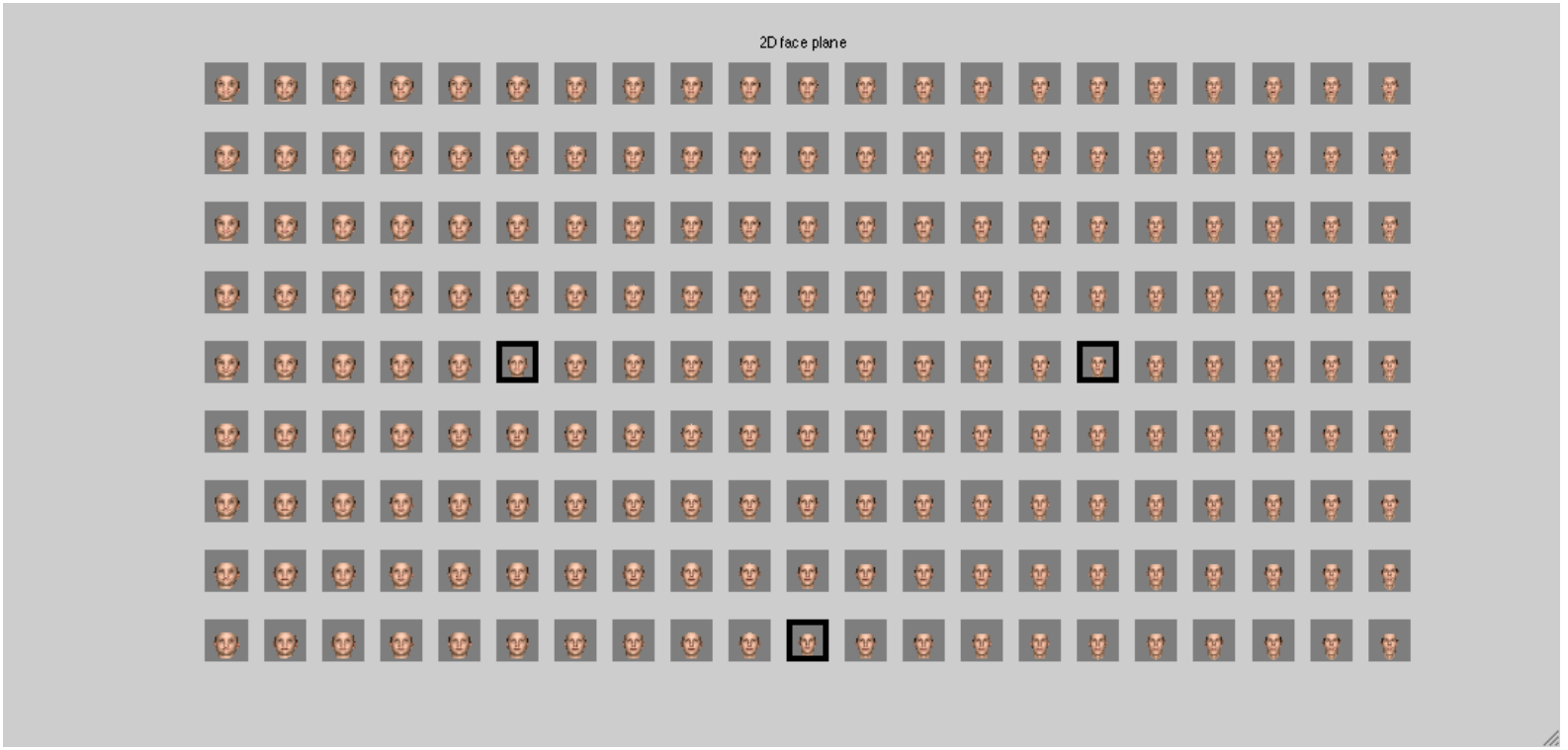


Figure 3.3: An example 2D-slice around which subjects can navigate using the keyboard. The "norm", or zero-vector, face (bottom row) along with the two stimuli for the target stimulus (middle row) are picked out with bold black borders. Compare this to figure 3.2

duration of presentation was determined as a proportion of the overall period. Within the relevant period the relative duration of presentation of A versus B was varied in the proportions 0, 0.1, 0.2, ... , 0.9, 1, resulting in 22 different conditions considered from an abstract perspective, but in fact only 18 distinct conditions, since conditions in which the relative duration was either 1 or 0 were equivalent (i.e. a single face presented continuously). The parameter values corresponding to the different experimental conditions are tabulated in table 3.2.1. In those conditions where both alternating faces were presented for a positive fraction of the duty cycle the maximum duration any face was continuously present was 180ms and the minimum 10ms. This is rapid enough that the faces do not appear distinct, but coalesce into a single identity, albeit a slightly spectral and flickering one with some of the subjective qualities of a stroboscopic lamp.

Period (ms)	Duration of A (ms)	Duration of B (ms)	Duty Cycle (%)
100	0	100	0
100	10	90	10
100	20	80	20
100	30	70	30
100	40	60	40
100	50	50	50
100	60	40	60
100	70	30	70
100	80	20	80
100	90	10	90
100	100	0	100
200	0	200	0
200	20	180	10
200	40	160	20
200	60	140	30
200	80	120	40
200	100	100	50
200	120	80	60
200	140	60	70
200	160	40	80
200	180	20	90
200	200	0	100

Table 3.1: A tabulation of the parameter values for each condition in the dynamic experiment. The duty cycle in the last column is defined with respect to A. There are 22 rows but conditions in which the duty cycle of stimulus A is 0% or 100% are equivalent since in these cases the target stimulus is a static face, B or A respectively.

3.2.2 Methods: Contrast Experiment

This experiment was designed to be the same as the DE, described in subsection 3.2.1, except for the manner in which two faces were merged to form the target stimulus on any trial. Rather than rapidly switching between two faces at the same spatial location, as in the DE, in the CE the two faces were superimposed each with a contrast of between 0 and 100%, such that the sum of the two values was always 100%. The contrast was varied at increments of 10% resulting the parameter values in table 3.2. Just as in the DE the subject was presented with a target stimulus on the left and was required to navigate around a 2D grid of faces which contained the norm face as well as the two faces used to construct the target stimulus. Having terminated a trial by pressing the key 'd' the parameters of the next trial were likewise selected in a pseudorandom fashion from those listed in table 3.2, so this was an interleaved design. Due to there being fewer parameter permutations for the CE (see table 3.2) than for the DE (cf. table 3.2.1) the sessions were shorter. Whereas subjects performed 220 trails in the DE (10 trials per row in table 3.2.1), subjects performed only 110 trials for the CE (10 trails per row in table 3.2). However, as the designs were in both cases pseudorandomly interleaved, there is little reason to suspect fatigue, or other duration related factors, could have systematically biased the results in one experiment but not the other.

Contrast of A (%)	Contrast of B (%)	Contrast weighting
0	100	0
10	90	0.1
20	80	0.2
30	70	0.3
40	60	0.4
50	50	0.5
60	40	0.6
70	30	0.7
80	20	0.8
90	10	0.9
100	0	1

Table 3.2: A tabulation of the parameter values for each condition in the static experiment. The contrast weighting in the last column is defined with respect to A. There are 11 rows corresponding to 11 different weightings between two faces, using increments of 0.1 and constrained to sum to 1. Again conditions in which the weighting is 0 or 1 correspond to a single face, B or A respectively.

3.3 Results

3.3.1 Summations, averaging and leveraging of the symmetry of the experimental condition

For the analyses contained in this section data was pooled across all 20 subjects, who individually displayed systematic, but extremely noisy, variations between the DE and CE experimental conditions. As described in detail in Methods (section 3.2) the geometry of the experimental design is inherently symmetric, with duty cycles of 0%, 10%, 20%, 30% and 40% being equivalent to duty cycles of 100%, 90%, 80%, 70% and 60% respectively. By dint of this inherent symmetry the task of estimating judgments was reduced to 6 point estimates, versus 11 naively. This manipulation essentially magnifies the quality of data available for each point estimate by a factor of 1.83 (on average) and explains the symmetry displayed the figures, very evident for example in figure 3.8.

For many of the statistical tests performed in this chapter it will be stated that a Bonferroni correction has been applied, with a correction factor of 1/6. Of course, this numerical value results from the fact that there are typically 6 independent conditions and 6 statistical tests performed, t-tests for example. This again is a consequence of the inherent symmetry of the experimental design.

3.3.2 Dynamic Experiment (DE) Results

In chapter 2 an exponentially smoothed state space model was developed, and a prediction of linear interpolation for rapidly alternated stimuli was extracted. As will be seen, experimental results do not bear this prediction out, neither in the radial ($\hat{\mathbf{r}}$) nor the tangential ($\hat{\mathbf{r}}^{\perp}$) axes. The basic numerical results are shown in table 3.3 in the form of % deviation [1] from the null hypothesis of linearity. Inspection of the table reveals that in almost every condition the deviation is statistically significant, in both the radial and tangential directions. Evidently, the supposition of linear interpolation along the morph line as a function of duty cycle is not supported by this data, at least not directly. A natural question following from this is whether there is a systematic pattern of deviation to be discerned.

[1] 100% = the norm of the vector of the midpoint of the morph line. See caption of table 3.3 for further detail.

% Duty cycle of A	% Radial deviation (s.e.)	% Tangential deviation (s.e.)
0	+8.5 (± 2.0)***	-5.2 (± 1.7)***
10	+5.5 (± 2.3)**	+6.7 (± 2.0)***
20	-1.7 (± 2.3)	+10.5 (± 2.5)***
30	-10.6 (± 2.3)***	+11.47 (± 3.0)***
40	-9.2 (± 2.4)***	+4.2 (± 3.4)***
50	-14.8 (± 2.3)***	+1.7 (± 3.5)
60	-9.2 (± 2.4)***	-0.8 (± 3.4)***
70	-10.6 (± 2.3)***	-8.0 (± 3.0)***
80	-1.7 (± 2.3)	-7.0 (± 2.5)***
90	+5.5 (± 2.5)**	-3.2 (± 2.0)***
100	+8.5 (± 2.0)***	+8.6 (± 1.7)***

Table 3.3: Radial and tangential deviations from the predictions made by the state space model (i.e. essentially linear interpolation according to duty cycle). The unit of distance used to express the deviations is the norm of the vector of the midpoint of the morphline (i.e. the distance from the origin to the midpoint of the morphline). Brackets contain standard errors for the estimates. $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$. $*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-values are adjusted for multiple comparisons using the Bonferroni correction (number of independent comparisons = 6). Column 2 (radial deviation) confirms what can be appreciated visually from, for example, figures 3.4 and 3.5, namely a positive bias for conditions in which the duty cycle is close to 0% or 100%, which reverses to a pronounced negative bias as the duty cycle approaches 50%. Column 3 demonstrates was is less obvious, but still appreciable from the figures, namely that there is an apparent bias towards the extremities of the morphline. This could be thought of as a repulsive effect away from the midpoint of the morphline or conversely an attractive effect towards the extremes.

An appreciation of the divergence from linearity is aided by visualising the data. When we do so, as in figure 3.4, it is apparent that for intermediate duty cycles there is a very significant deviation of judgements away from the morph line and towards the norm. This is most pronounced when the duty cycle is 50%, that is when the two faces are present for the same length of time on any one cycle of period 100ms or 200ms. As can be seen from figures 3.5 and 3.10 this effect is pronounced for periods of both 100ms and 200ms. In contrast, the state space model, developed in chapter 2, predicts that the subject's percept should interpolate along the morph line on average, and therefore we would expect the judgements to do so too. These deviations from the morph line are highly statistically significant in nearly every duty cycle, as shown in figure 3.8.

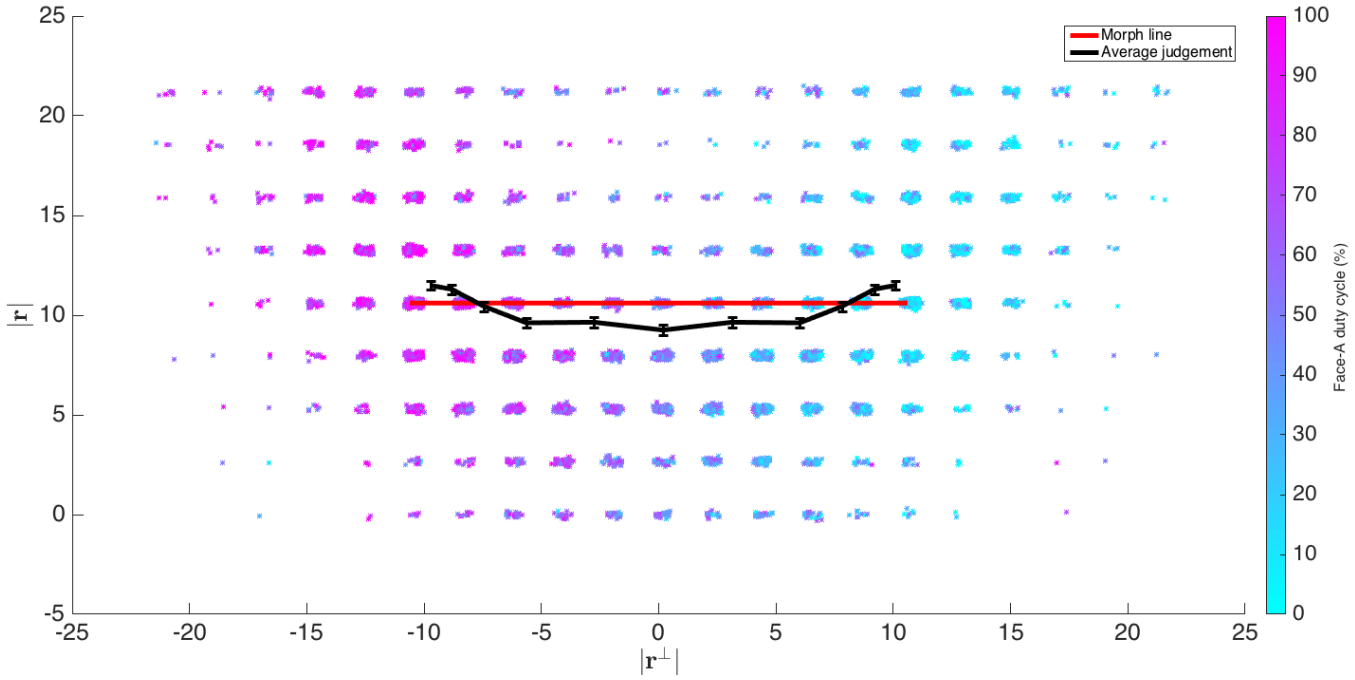


Figure 3.4: Data from the DE (dynamic experiment) for all 20 subjects. Each point represents a match to the target stimulus that the subject made on a single trial. The colour of the point represents the proportional duration of face A (purple) versus face B (blue) on any trial. Data from both the 100ms and 200ms conditions are included. The state space model described in chapter 2 mandates these judgements to, on average, fall along the morphline connecting A and B (plotted in thick red). The actual means from the experimental data are plotted in thick black. It is clear, even visually, that there is very significant deviation from the model predictions, especially where face A and B have similar proportional durations.

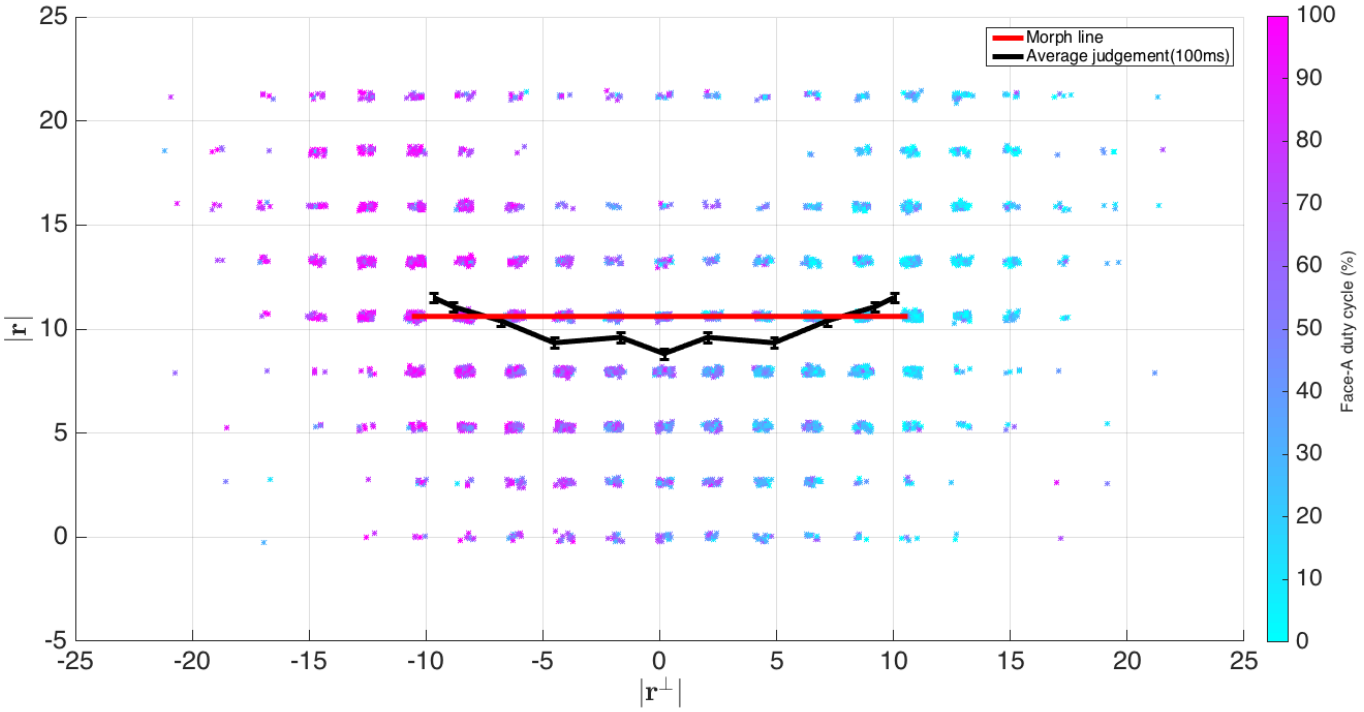


Figure 3.5: Data from the DE (dynamic experiment) for all 20 subjects for a period of 100ms (cf. figures **??** and **??**).

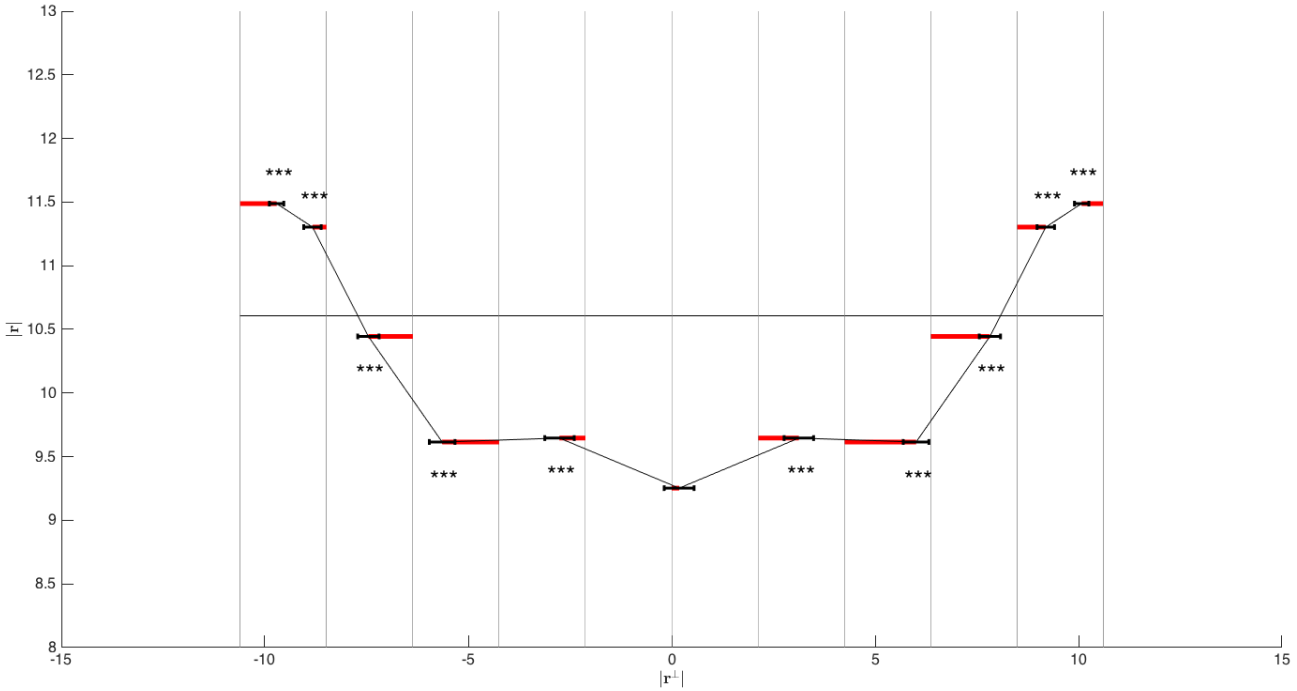


Figure 3.6: Tangential or lateral deviation from prediction of linear interpolation in the DE. Data shown for all 20 subjects in the DE. Two sided t-test were performed wherein the null hypothesis states that the judgements are sampled from a distribution with a mean that lies along the morph line in linear proportion to the duty cycle. Grey vertical lines display the null hypotheses (i.e. the predictions of linear interpolation), and thick red lines represent the deviation from those linear predictions in the tangential direction. Black error bars represent standard errors. $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$. $*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-values are adjusted for multiple comparisons using a Bonferroni correction (number of independent comparisons = 6). See section 3.4 for further discussion.

Furthermore, not only did the means of the response distributions differ considerably from those predicted by the state space model, but the variability of responses also changed as a function of the duty cycle. The state space model in fact makes no explicit prediction about the form of the response distributions, and it was assumed that they would essentially be well approximated by isotropic Gaussians. That this is not the case can be seen from figure 3.9 which shows the response distributions plotted for each of 11 proportional durations $0, 0.1, \ldots, 1.0$. There is a marked trend from narrow (low entropy) anisotropic distributions at extremes of the duty cycle (i.e. $\approx$ 0% or 100%) to broad (high entropy) more isotropic distributions at intermediate duty cycles (i.e. $\approx$ 50%).

Although a prominent bowing effect can be seen in both the 100ms and 200ms conditions of the DE a natural question is whether the judgement lines in these two conditions are statistically indistinguishable. That there is indeed no significant difference between the subjects' judgments in the two conditions is demonstrated in subsection 3.3.3.

The question arises as to whether some or all of this deviation form the morphline can be explained by the design of the experiment. Firstly, as noted previously, the initialisation point on each trial was chosen randomly and uniformly from the 9-by-21 2D grid of faces. Thus the deviation towards the norm cannot be explained by, say, always starting at the norm, moving randomly and stopping after a short time (so that there would be little opportunity to "diffuse" away from the origin). What about edge effects? Since subjects could only move freely up, down, left or right when not at the edge of the 2D face grid perhaps this somehow distorts even a

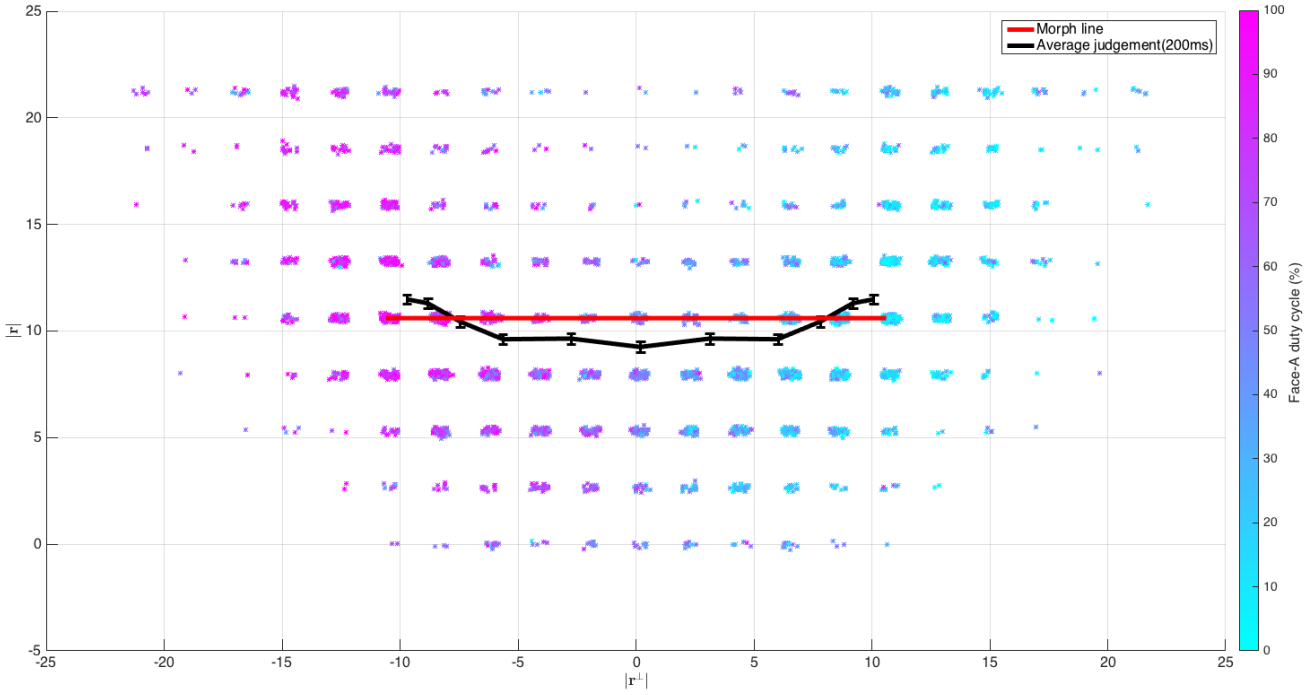


Figure 3.7: Data from the DE (dynamic experiment) for all 20 subjects for a period of 200ms.

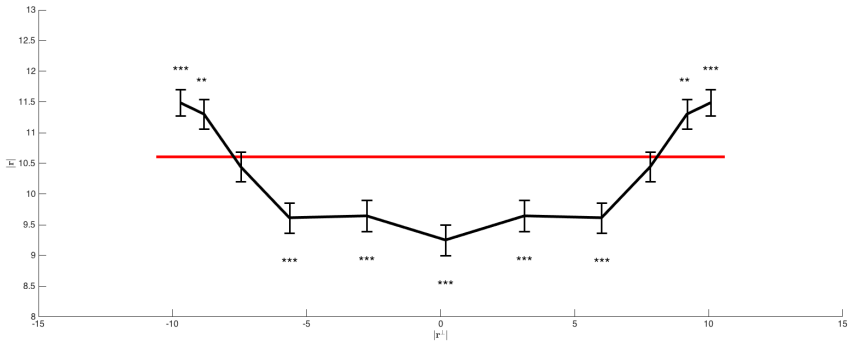


Figure 3.8: Radial deviation from prediction of linear interpolation. Data shown for all 20 subjects in the DE. Two sided t-test were performed wherein the null hypothesis states that the judgements are sampled from a distribution with a mean that lies on the morph line. $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$. $*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-values are adjusted for multiple comparisons using the Bonferroni correction (number of independent comparisons = 6). See section 3.4 for further discussion

random walk with random initialisation and stopping point irrespective of what target stimulus is present. In other words, what is the null distribution assuming that the target face has no effect upon stopping position for the matcher? This null hypothesis can be expressed, for the 2D slice, as a Markov chain with 189 states, or nodes corresponding to the faces within the 2D grid, and an bidirectional edges between two states/nodes if the two faces are adjacent on the grid. To express this graphically would require a graph with 189 nodes and a corresponding transition matrix of dimensions 189 x 189. We can however illustrate the situation for a somewhat smaller 3 x 4 grid, which is nevertheless possesses the same properties as the 9 x 21 grid in terms of analysis. A Markov chain for the analogous 3 x 4 case is shown in figure 3.11 and the corresponding (time-

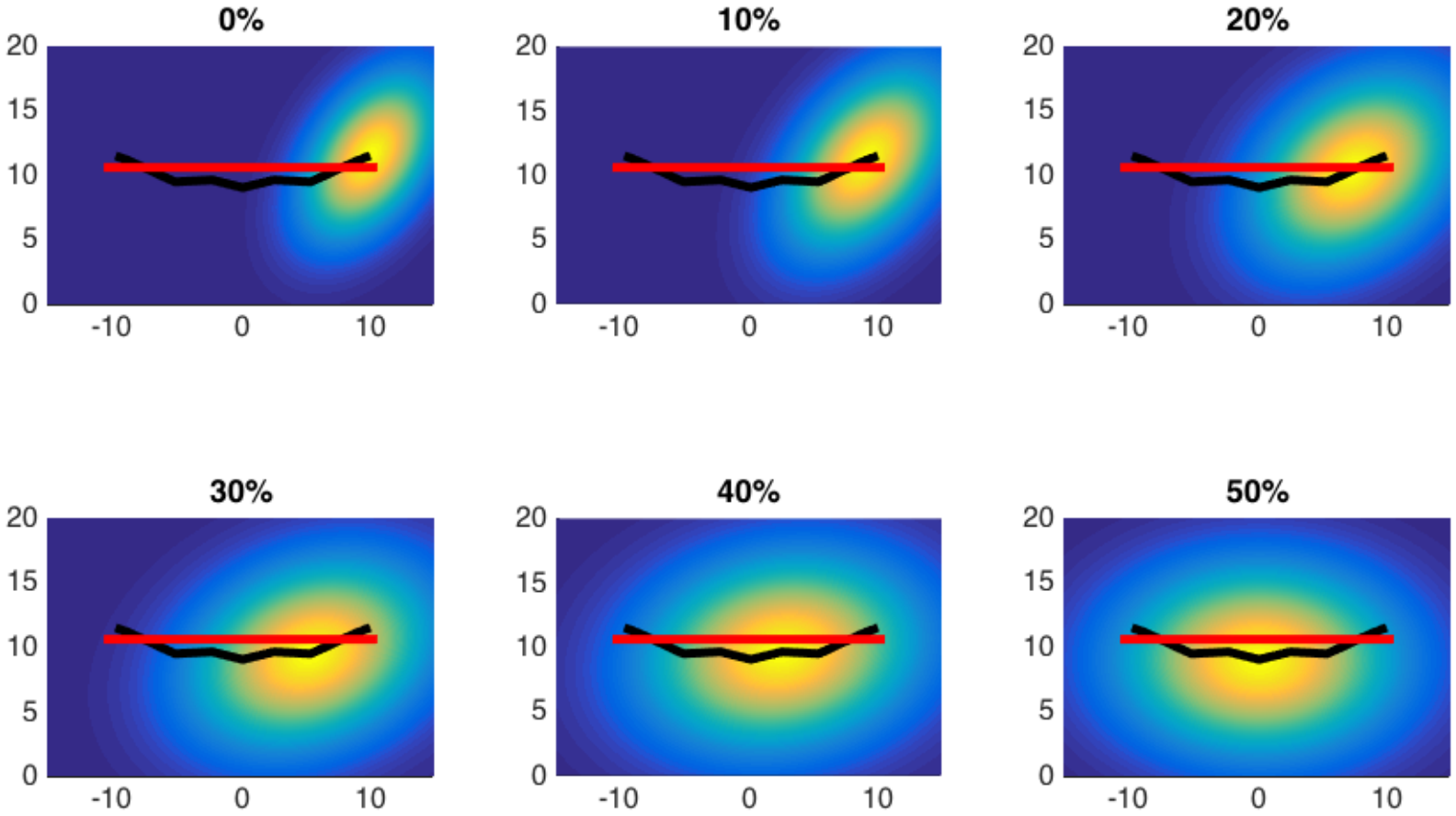


Figure 3.9: Fitted Gaussian distributions to the responses of all 20 subjects for duty cycles 0-50%, duty cycles 60-100% being equivalent to 40-0%. Not only does the mean of the response distribution follow an approximately arc-shaped (black), rather than linear (red), path between duty cycles 0-50%, but clearly the form of the distribution varies too, becoming increasingly broad as duty cycles approach 50%.

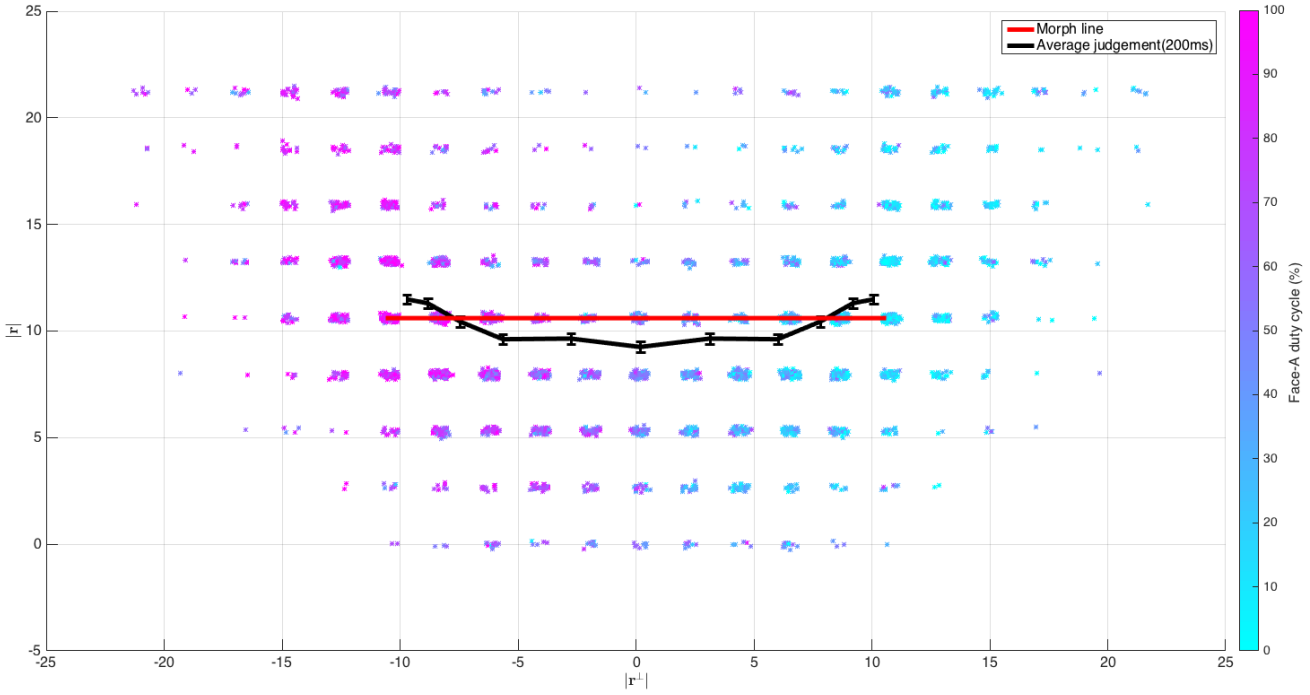


Figure 3.10: Data from the DE (dynamic experiment) for all 20 subjects for a period of 200ms (cf. figures 3.9 and 3.5).

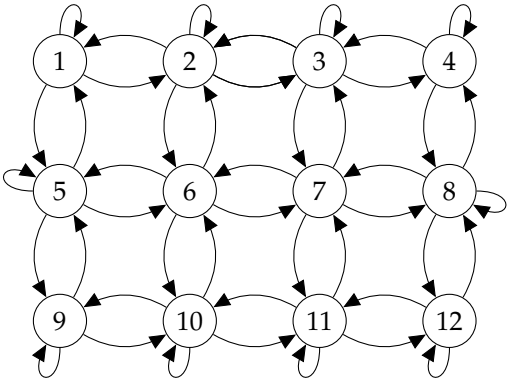


Figure 3.11: A 3 by 4 markov model, analogous to the 9 by 21 plane of faces subjects navigate around with the matcher.

homogeneous) transition matrix, $\mathbf{P}$, in equation 3.4.

The stationary distribution for a positive recurrent, time-homogenous Markov chain of n vertices is given by solving the equation

$$\pi = \pi \mathbf{P} \tag{3.1}$$

Given the constraint that

$$\sum_{i=1}^{n} \pi_i = 1 \tag{3.2}$$

Leaving the details of the algebra to one side, for the sake of brevity, it turns out that any such Markov chain we care to construct will in general have a stationary distribution, π such that

$$\pi_i = 1/n \tag{3.3}$$

The material point we extract from this analysis is thus that a subject moving randomly in the 2D plane, despite the apparent possibility of edge

effect, will nevertheless achieve a uniform distribution across the grid. And the upshot is that the non-uniformity seen in the subject data cannot therefore plausibly be an artefact of the experimental design, at least as regards the possibility of edge effects on a random walk.

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0.5 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0.25 & 0 & 0.25 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0 & 0.25 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0 & 0 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0.5 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.5 \end{pmatrix} \tag{3.4}$$

3.3.3 A comparison of the 100ms condition and the 200ms condition for the Dynamic Experiment

As can be seen from the data in both the 100ms and 200ms condition, there is a pronounced bowing in both conditions. However, an important question is whether the pattern of radial and lateral deviation from the predic-

tions of linearity are identical in the two cases. If it is really the case that, below a certain threshold duration at which one perceives a single face distinctly, it is the relative duration of one stimulus with respect to the other that chiefly explains the location of subjects judgements, then the absolute duration of each stimulus should not matter. To be explicit, we should see no difference between the 100ms and 200ms conditions if this account is at least approximately true. Moreover, were there a significant difference between the judgements obtained in the 100ms and 200ms conditions, then this could invalidate the pooled analysis of the bowing effect presented.

Figure 3.12 confirms visually that the judgements from the 100ms and 200ms conditions are indeed extremely similar. This visual impression is confirmed statistically in that two sampled t-tests performed across conditions shown no significant difference, even at the lowest conventional threshold for signioficance (i.e. $p < 0.05$). The results of these tests are shown in table 3.4

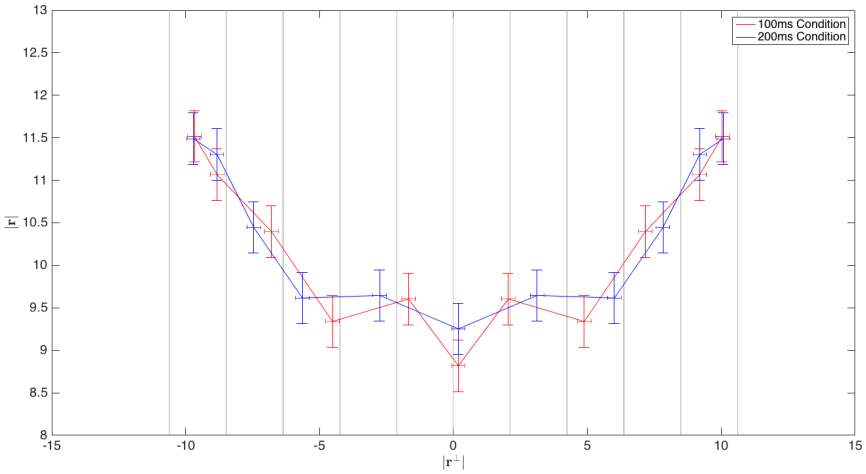


Figure 3.12: Overlaying the judgement curves made from subject data in the 100ms (red) and 200ms (blue) conditions shows there is little difference between the two, neither in lateral nor the radial directions. This is confirmed by two sample t-tests across all duty cycles with Bonferroni corrections, the results of which are shown in table 3.4. Error bars represent standard errors.

Duty cycle of A (%)	$H_0(\|\mathbf{r}^\perp\|)$	p	$H_0(\|\mathbf{r}\|)$	p
0	0	0.87401	0	0.92497
10	0	0.9932	0	0.48461
20	0	0.075039	0	0.88938
30	0	0.012027	0	0.42858
40	0	0.034563	0	0.90583
50	0	1	0	0.21612
60	0	0.034563	0	0.90583
70	0	0.012027	0	0.42858
80	0	0.075039	0	0.88938
90	0	0.9932	0	0.48461
100	0	0.87401	0	0.92497

Table 3.4: Results of two sample t-tests carried out on subjects' jusdgements for the 100ms and 200ms DE. The columns headed by H_0, i.e. columns 2 and 4, denote the result of the hypothesis test for the lateral ($|\mathbf{r}^\perp|$) and radial ($|\mathbf{r}|$) directions respectively, with 0 indicating insufficient evidence to reject the null hypothesis that the data are drawn from the same distribution in both the 100ms and 200ms conditions. p denotes the p-value of each test. The level of significance was $p < 0.05$, with a Bonferroni correction for (6) multiple comparisons (i.e. $p < 0.05/6 = 0.0083$).

3.3.4 Contrast Experiment (CE) Results

The main motivation for performing the CE (Contrast Experiment) was a concern that the bowing effect found in the DE might be an artefact of the experimental design or somehow unrelated to the dynamic features of the stimuli. In this sense the CE is really a control experiment and the results, shown in figure 3.13 bear out the contention that the bowing effect seen in the DE is indeed a consequence of the dynamic nature of the stimuli, since in this static context it is entirely absent, and indeed the data exhibit a small bias in the opposite direction. Figure 3.15 displays the results of a t-test performed for each contrast weighting with the null hypothesis being that the judgements are sampled from a distribution with a mean on the morph line. The null hypothesis is rejected for all but one of the conditions. In short there appears to be a fairly consistent centrifugal radial deviation from the morph line, akin to that seen in the DE for more extreme duty cycles (see table 3.3). Given that this effect is not significant for all conditions an important question is whether there is any systematic change as a function of contrast weighting, analogous to the centrifugal to centripetal trend seen in the DE as duty cycles varied from 0% or 100% to 50%. Table 3.6 shows the results of a paired t-test between the extreme conditions (contrast weightings of 0% or 100%) and intermediate conditions. As can be seen from the second column, every comparison is insignificant at a significance level of 0.05, supporting the hypothesis that there is no systematic change as a function of contrast weighting, and in particular no bowing of the kind seen in the DE. In summary then, the data from the

CE and DE experiments appear to deviate from the predictions of linearity in the same way, with the singular exception of the radial centripetal effect, the bowing, which is only seen in the DE.

The bowing effect, present in the DE and absent in the CE (see table 3.6), relates to the deviation from linear interpolation in the radial direction, and is as discussed prominent in the data. However, there are equally significant deviations in the tangential or lateral direction. These deviations are tabulated in figure 3.5 and displayed graphically in figure 3.16.

Contrast of A (%)	Deviation in $\|\mathbf{r}\|$ (%) (s.e.)	Deviation in $\|\mathbf{r}^{\perp}\|$ (%) (s.e.)
0	+8.5 (± 2.8)***	-6.5 (± 2.5)***
10	+7.5 (± 2.8)**	+11.4 (± 2.3)***
20	+8.9 (± 2.8)***	+23.5 (± 2.4)***
30	+10.6 (± 2.9)***	+36.1 (± 3.0)***
40	+3.1 (± 3.2)	+29.6 (± 3.6)***
50	+10.6 (± 3.0)**	+0.3 (± 4.6)
60	+3.1 (± 3.2)	-29.1 (± 3.6)***
70	+10.6 (± 2.9)***	-35.6 (± 3.0)***
80	+8.9 (± 2.8)***	-22.9 (± 2.4)***
90	+7.5 (± 2.8)**	-10.8 (± 2.3)***
100	+8.5 (± 2.8)***	+7.1 (± 2.5)***

Table 3.5: Radial ($|\mathbf{r}|$) and tangential ($|\mathbf{r}^{\perp}|$) deviations from the predictions made from a "contrast model" (i.e. linear interpolation along the morphline according to the % contrast). The unit of distance used to express the deviations is the norm of the vector of the midpoint of the morphline (i.e. the distance from the origin to the midpoint of the morphline). Brackets contain standard errors for the estimates. $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$. $*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-value thresholds are adjusted for multiple comparisons using the Bonferroni correction (number of independent comparisons = 6).

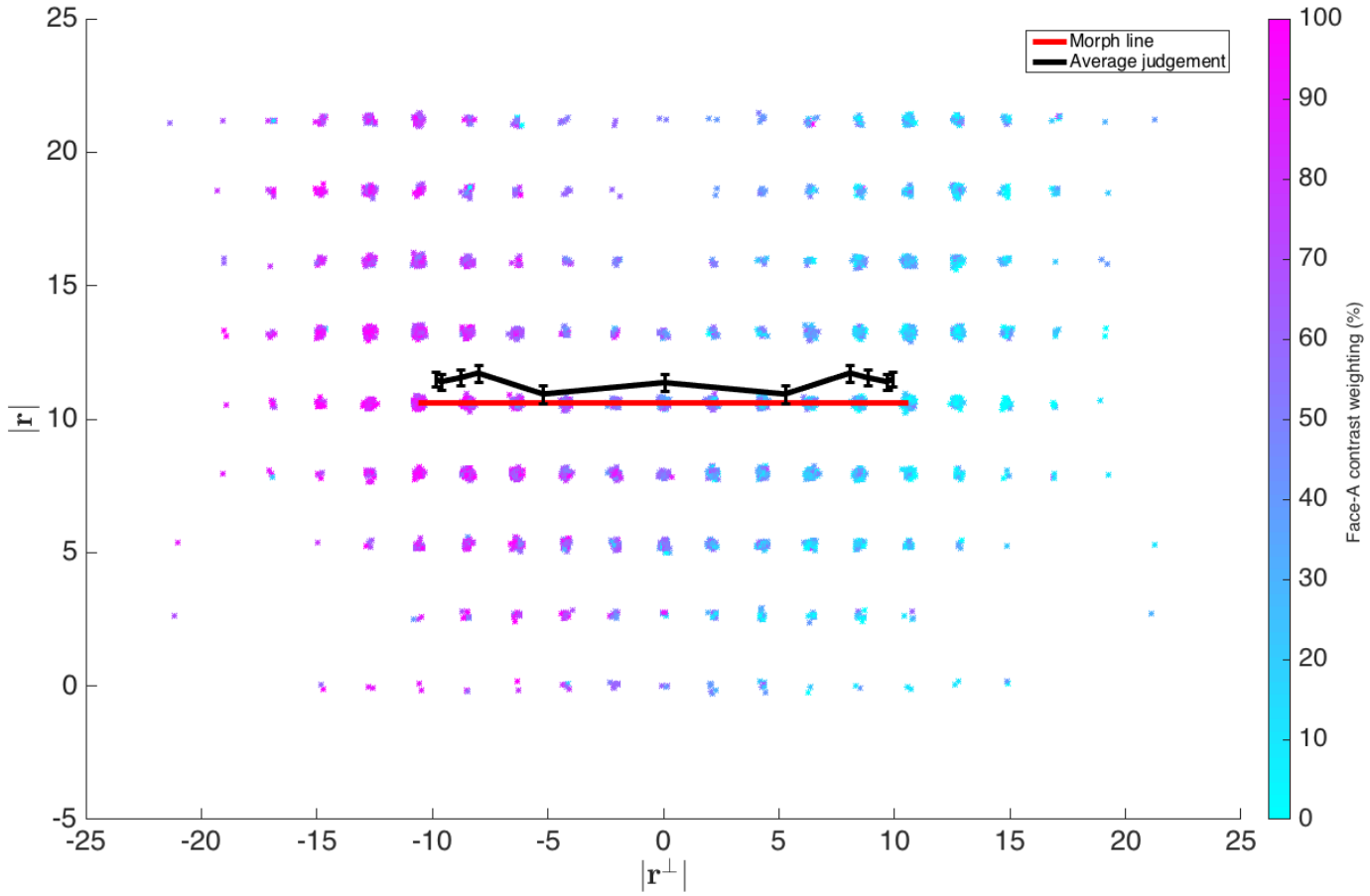


Figure 3.13: Data from the CE (contrast experiment) for all twenty subjects. Each datapoint represents a match to the target stimulus that the subject made on a single trial. The the colour of the data-points represents the contrast weighting of A (*A*methyst) versus face B (*B*lue) on any trial. As the contrast weighting between the two faces varies between 0% and 100% judgements might be expected to fall at corresponding points along the morph line connecting A and B, plotted in thick red. The actual means from the experimental data are plotted in thick black. In contrast to the DE (cf. figure 3.10) here there is no bowing towards the origin, suggesting that the effect seen in the DE is an inherently dynamic effect. A further difference compared to the DE is that when projected onto it judgements are not linearly interpolated along the morph line. Instead there appears to be a moderate winner-takes-all effect whereby judgements are biased towards the stimulus of higher contrast.

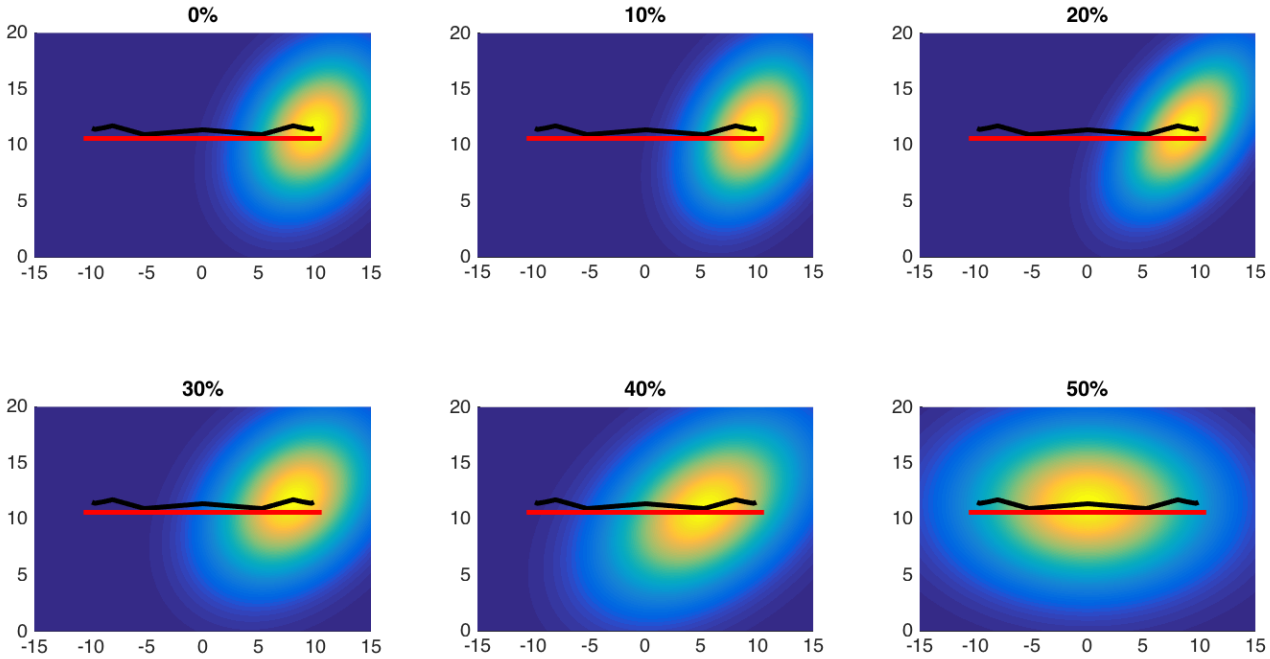


Figure 3.14: Fitted Gaussian distributions to the responses of all 20 subjects for contrast weightings 0-50%, weightings 60-100% being equivalent to 40-0% respectively. The means of the response distributions follow an approximately parallel (black) course to the linear morph line (in red). Statistical support for this observation (i.e. the absence of bowing inward of outward as a function of contrast weighting) is provided in table 3.6. As in the DE the form of the distribution varies too, becoming increasingly broad as duty cycles approach 50%.

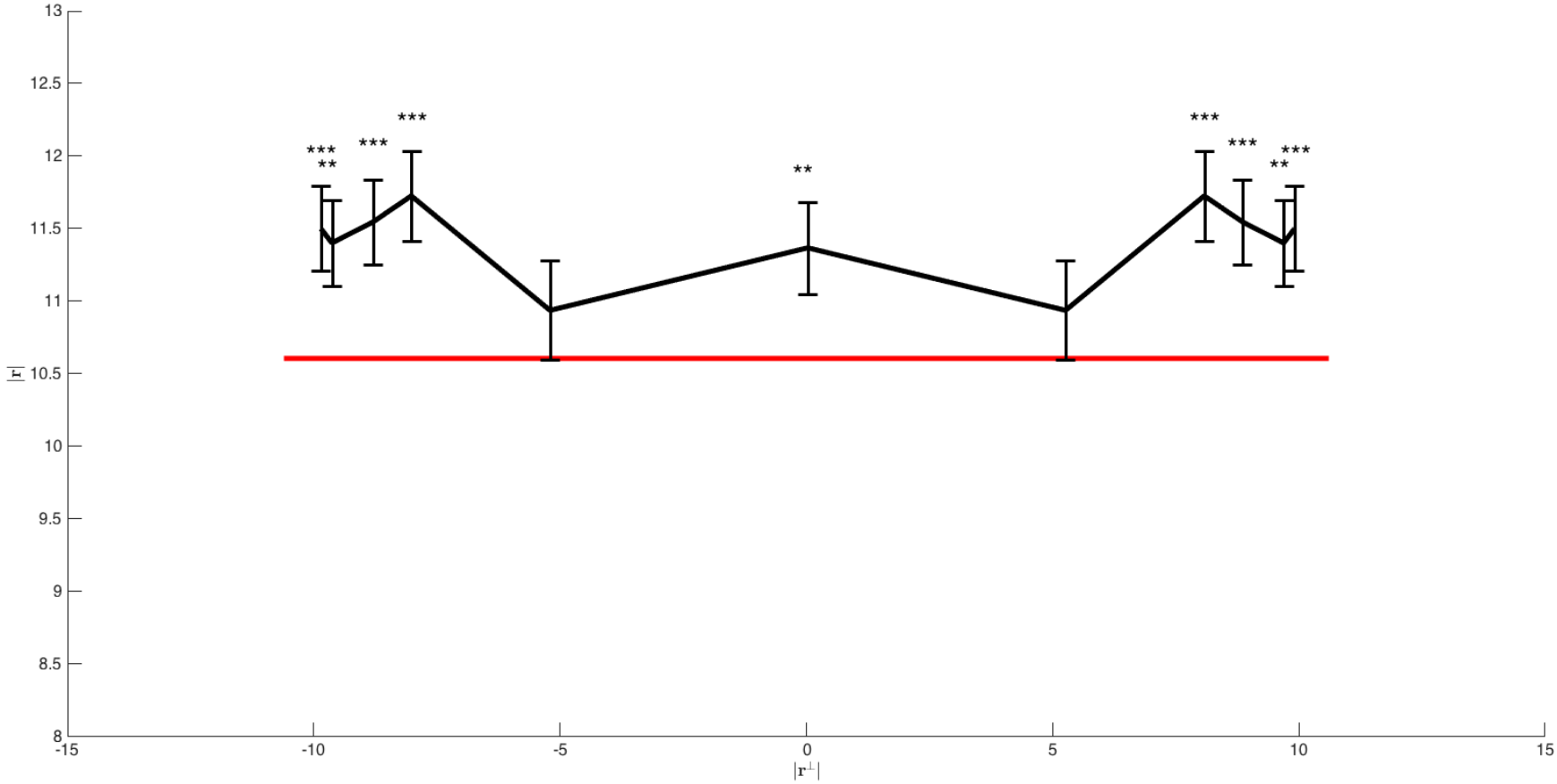


Figure 3.15: This analysis relates to data for all 20 subjects in the CE. Two sided t-tests were performed where the null hypothesis states that the judgements are sampled from a distribution with a mean that lies on the morph line (in red). $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$.$*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-numbers are adjusted for multiple comparisons using a Bonferroni correction (number of independent comparisons = 6).

Contrast of A (%)	H_0	p
0 vs. 0	0	1
0 vs.10	0	0.7358
0 vs. 20	0	0.8738
0 vs.30	0	0.4554
0 vs.40	0	0.0767
0 vs.50	0	0.6636
0 vs.60	0	0.0767
0 vs.70	0	0.4554
0 vs.80	0	0.8738
0 vs.90	0	0.7358
0 vs.100	0	1

Table 3.6: Results of two sample t-tests carried out on subjects' judgements to test for evidence of bowing (or other changes in radial deviation) in the CE. The column headed H_0, i.e. column 2, denotes the result of the hypothesis test, with 0 indicating insufficient evidence to reject the null hypothesis of no difference in radial deviation compared to the extreme conditions (i.e. 0% and 100%). The level of significance was $p < 0.05$, with a Bonferroni correction for (6) multiple comparisons (i.e. $p < 0.05/6 = 0.0083$). Note that a comparison between 0% and 0%, and 0% and 100% is included for clarity of exposition, but returns a p-value of 1, and therefore a -ve result at all levels of significance, by definition. The implication of this is that there is no evidence of systematic variation as a function of contrast weighting, and in particular of bowing.

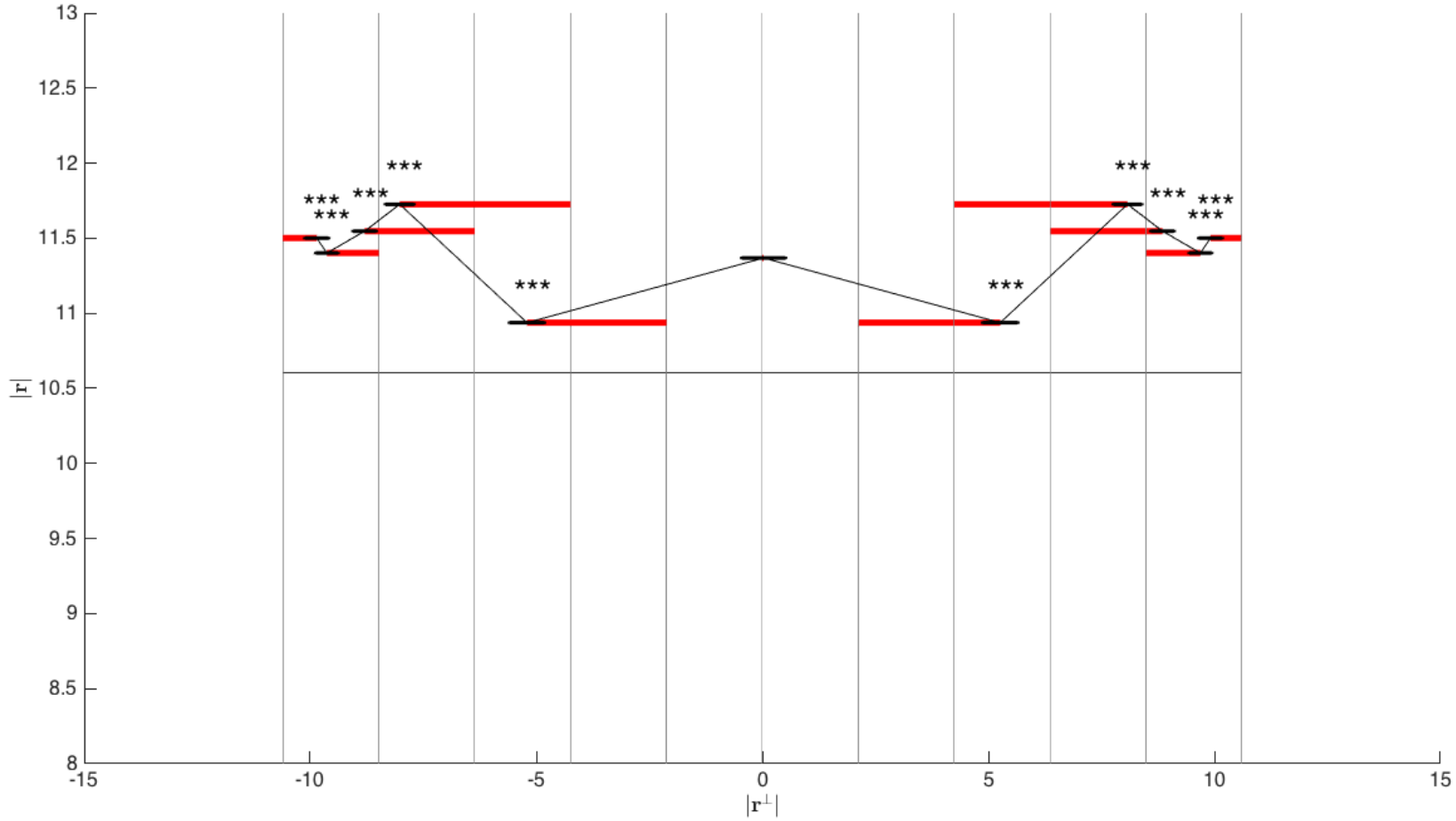


Figure 3.16: Tangential or lateral deviation from prediction of linear interpolation. Data shown for all 20 subjects in the CE. Two sided t-test were performed wherein the null hypothesis states that the judgements are sampled from a distribution with a mean that lies along the morph line in linear proportion to the contrast weighting. Grey vertical lines display the null hypotheses (i.e. the predictions of linear interpolation), and thick red lines represent the deviation from those linear predictions in the tangential direction. Black error bars represent standard errors. $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$. $*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-numbers are adjusted for multiple comparisons using the Bonferroni correction (number of independent comparisons = 6). See section 3.4 for further discussion.

3.4 Discussion

The upshot of the experimental results presented in this chapter can be summarised in the following way: when two faces are rapidly alternated, the percept is not a linear interpolation of the two stimuli. Instead, there appears to be a radial centripetal deviation towards the origin of face space at intermediate duty cycles (30-70 %) and a radial centrifugal deviation at duty cycles close to 0% or 100% (0-10% and 90-100%). In terms of the lateral or tangential deviation there are also deviations from the predictions of linear interpolation, namely a tangential centrifugal bias at duty cycles of 10-90% and a tangential centripetal bias at duty cycles of 0% and 100%. Moreover, the results of the CE reproduce the pattern of radial centrifugal and tangential deviations seen in the DE, but do not reproduce the radial centripetal effect, suggesting that the radial centripetal effect is not a general feature of merging two stimuli, but somehow arises from the rapid alternation of stimuli particular to the DE. This is shown in figures 3.15 and table 3.6.

However, because the CE was designed and performed after the DE, and specifically in light of the curious bowing effect evident in the results of the DE, one cannot exclude the possibility that these differences were due to adaptive effects. For example, one possibility might be that subjects were initially (i.e. during the DE) unfamiliar with the stimuli, and therefore biased towards the norm face (i.e. the origin). However, as they became more familiar with the stimuli (i.e. during the CE) they may have adjusted their estimates to the locations of the actual stimuli (i.e. the ex-

tremes of the morph line). This might therefore explain how the prominent centrifugal radial effect, seen in the DE, was apparently abolished in the CE, without having anything to do in fact with the design of the experiments.

In addition to the radial deviations from linearity, a prominent feature of the data concerns the lateral, or tangential, deviations, displayed numerically in tables 3.3 and 3.5 and graphically in figures 3.6 and 3.16. A comparison of the lateral deviations in the CE and DE reveals that the pattern of lateral deviation is replicated across experimental conditions, however the magnitudes of the deviations are considerably greater in the CE than in the DE. Indeed, comparison of tangential deviations in tables 3.6 (DE) and 3.16 (CE) reveals that the magnitude of the lateral deviation for duty cycles or contrasts other than 0%, 50% and 100% was greater in the (CE) by a factor of around 2 (10%) to 30 (60%). However, the fact that the CE was performed in every case after the DE again makes it impossible to establish conclusively whether this is a difference due to the design of the experiment or an adaptive effect dependent on familiarity with the stimuli.

Were experiments to be repeated, one could examine the possibility of temporal or adaptive effects by having half the subjects perform the CE and then the DE and the other half perform the DE and then the CE. Still better this possibility of subjects becoming familiarised with specific stimuli could have been prevented completely by using a sufficiently large number stimuli such that the subject never saw the same stimuli in more than one trial. In actual fact, as small set of only ten pairs of faces was

used, meaning that this scenario is rendered quite possible. It would be an important aspect of future work to eliminate this possible explanation of the data.

A prominent feature of the data in both the CE and the DE, in addition to the radial centripetal and lateral deviations already discussed, is a radial *centrifugal* deviation. This is seen at duty cycles of 0%, 10%, 90%, and 100% in the DE and at contrast weightings of all except 40% and 60% in the CE. Even in the case of the 40% and 60% contrast weightings, while the effect does not reach statistical significance, the effect is in the direction of a radial centrifugal deviation as can be seen from figure 3.13. A natural thought is that this represents a rediscovery of the controversial caricature effect. Indeed, the experimental design, with the matching and target stimuli side by side, is such that the subject is not in fact matching two stimuli directly, but matching one stimulus to a mental representation of the other stimulus. This would accord with accounts of the caricature effect whereby a caricature is recognised ”better” (e.g. faster and/or with more confidence) than a veridical image [Lee et al., 2000].

Notwithstanding the preceding caveats it is interesting to consider a probabilistic account of the most curious feature of the data in the DE, conspicuously absent in the CE, namely the strong radial centripetal deviation at intermediate duty cycles. In the following chapter I will develop a model showing how such as ”bowing” of the judgements could arise as a consequence of a Bayesian inferential process underpinning face-perception. The motivating intuition is that a rapidly alternating stimulus is inherently variable and therefore possessed of a degree of uncertainty

much greater that a static stimulus. If we consider an analogy with a Bernoulli random variable, $X \sim \mathrm{Bern}\,(p)$, (such as describes a coin which may or may not be biased), the variance and therefore uncertainly about the outcome is greatest when the parameter p is equal to 0.5. In accordance with canonical Bayesian theory, as evidence (represented by the likelihood function) becomes increasingly weak the posterior distribution approximates more and more closely to the prior distribution. In the context of our experimental results weak evidence in the form of an alternating stimulus, in the presence of a strong, central isotropic Gaussian prior, translates into an estimate of the location of the stimulus in face space closer to the norm or origin of face space.

4

A Bayesian Account of a Novel Dynamic Effect in Face Space

This chapter develops a model of probabilistic inference in face space. The basic assumption is that probability distributions are somehow represented in the brain. A Bayesian model is presented in which a prior over face space is combined with data (through a likelihood function) in the process of performing inference. Once the parame-

ters of this model have been fitted, using Markov Chain Monte Carlo, this result in a "bowing" effect that is apparently very similar to the data presented from the DE in chapter 3. This arises as a direct consequence of the Bayesian nature of the model. The chapter concludes with a discussion of some recent work in entomology that is suggestive of a possible class of neural mechanisms underpinning this effect.

4.1 Probabilistic Perceptual Inference

As discussed at length in chapter 1 section 1.4 there is now a broad and firm literature supporting the notion that humans can, and typically do, use prior knowledge of the statistical structure of the world to perform perceptual inference and guide behaviour [Körding and Wolpert, 2004]. Given a prior distribution over some variate of interest and an appropriate likelihood function the proper way to combine these sources of information to obtain a posterior distribution is given by Bayes' rule [Bishop, 2006].

$$P(\theta|\mathbf{I}) \propto P(\mathbf{I}|\theta)P(\theta) \tag{4.1}$$

It is very likely that the brain implements computations approximating Bayesian inference for a broad range of low level inference problems, as discussed at length in section 1.4 of chapter 1. These include, for example, estimating the orientation of line segments in the visual field [Girshick

et al., 2011]. However, it is less obvious that the brain adopts a Bayesian approach to high level problems such as facial identity. This chapter describes how a Bayesian framework can account for some of the curious features of the experimental findings described in chapter 3.

It should be made quite clear at the outset that this model is in no way intended to represent a mechanistic account of inference within face perception. On the contrary its purpose if purely to demonstrate that some of the idiosyncrasies of the data, and in particular the prominent ”bowing” effect seen in the DE (Dynamic Experiment) presented in subsection 3.3.2 can be explained by supposing a Bayesian computation at the heart of the process. It is therefore simply hypothesised this inferential process underpins the visual perceptual machinery relating to face perception, without making further claims concerning its nature.

4.2 Considerations Regarding Modelling Dimensionality

If something like face space does approximate the biological representation present in healthy humans, then it seems the representation is likely to be extremely high dimensional, even considering only the perspective of shape and texture. Indeed, the full Basel Face Model (BFM), which appears to provide a reasonable, albeit imperfect, predictor for similarity judgements in human subjects, comprises a 398 dimensional feature space (199 dimension in the texture model, and another 199 dimensions in the

shape model). Consider for a moment a multivariate normal distribution over this space, already a fairly restrictive assumption. Such a distribution, in its most general form, requires a forbidding 79,401 parameters to be specified [1]. Following Valentine and others [Valentine, 1991b] I suppose that the prior in face space is a isotropic multivariate normal distribution with a **0**-vector mean. This, in contrast, requires a single parameter, essentially specifying the rate at which the density falls off with eccentricity. Alternatively one could consider only a d-dimensional subspace of the full feature space, in which case one need only specify $d+d(d-1)/2$ parameters in the general case. As described in the methods section (3.2) of chapter 3, subjects were required to match the target stimulus with a face selected from a 2D grid within a high dimensional face space. Because judgements in any trial were constrained to a 2D plane we can therefore conduct the modelling in a 2D space, making fitting the parameters of the model to data tractable.

4.3 A Probabilistic Model of Face Perception

The basic motivation behind the Bayesian model outlined in this chapter is the desire to account for some curious features of the results presented in chapter 3. In that chapter we saw that where the stimulus was dynamic there emerged an unusual bias in the responses of subjects, which can be describe as a *bowing towards the origin* in the judgements of subjects. A salient feature of this bowing appears to be that the degree of deviation

[1] 398 to specify the mean vector, and a further 79,003 to specify the covariance matrix

from the morph line is approximately monotonically related to the degree of variability in the stimulus, reaching a maximum where the duty cycle is 50%. This is depicted in figure 4.1 and is given formally by the equation 4.2, where X_d represents the position on the morph line for a given duty cycle, d ($0 \leq d \leq 1$), and M denotes the total length of the morph line. Of note is the formal similarity to the variance of a Bernoulli random variable, $p(1-p)$, i.e. equal up to a constant factor $M^2/100$, also illustrated in figure 4.1

$$\mathrm{Var}[X_d] = \frac{M^2}{100} d(1-d) \tag{4.2}$$

The model proposed in this chapter is a relatively simple one based on some elementary facts about the nervous system, such as that its computations are subject to intrinsic neuronal as well as perceptual noise, which tends to reduce the certainty with which an organism can estimate the value of a variable. Furthermore, as a relatively high level attempt to understand the brain as a Bayesian computational device there is little to say about the implementation of this model at a neuronal level, although in section 4.7 an interesting possibility from the insect swarm intelligence literature is discussed [Seeley et al., 2012]. In Marr's terms this account operates at the representational/algorithmic level [Marr, 1983]. The graphical representation for this declarative model of perception is depicted in figure 4.2, which illustrates the dependency structure of the model [Bishop, 2006].

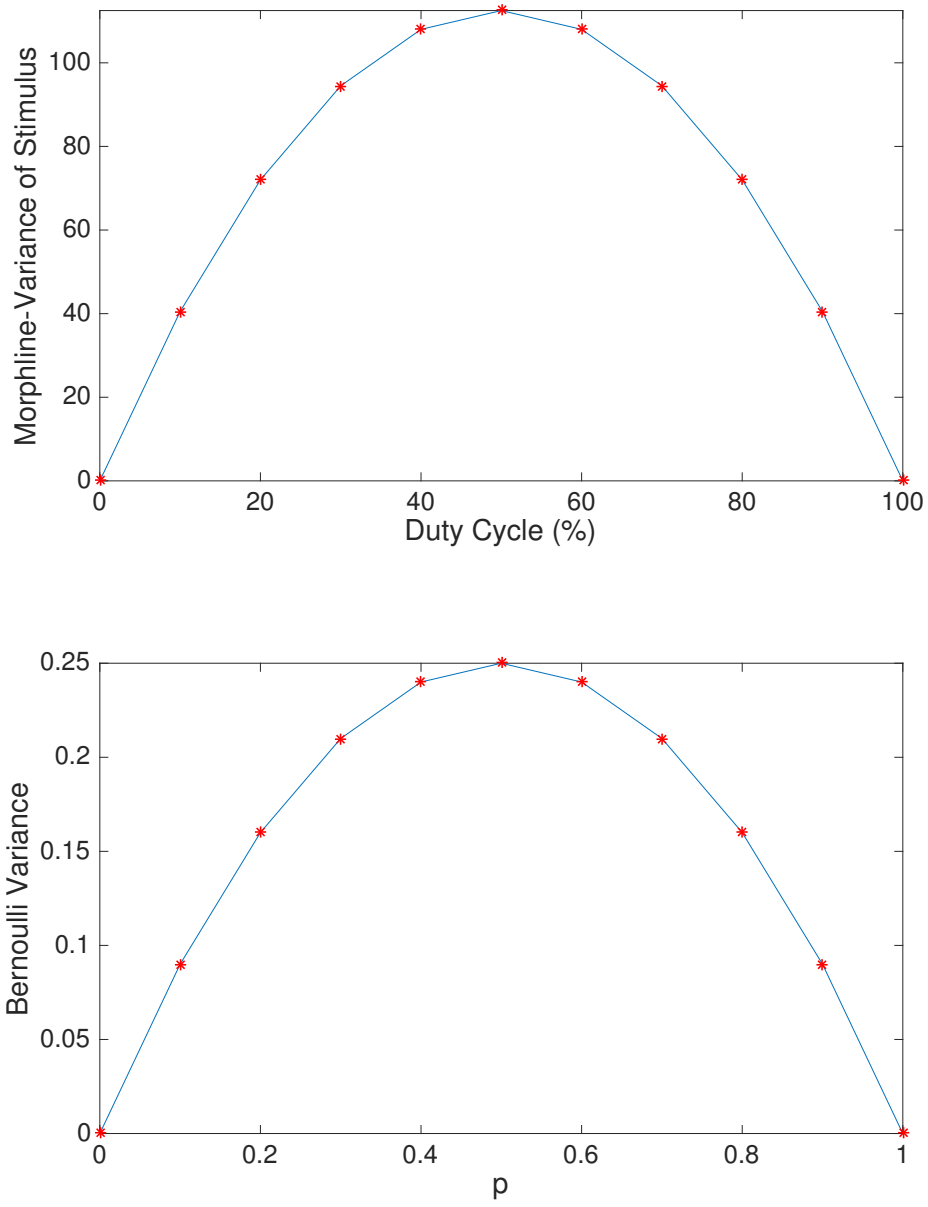
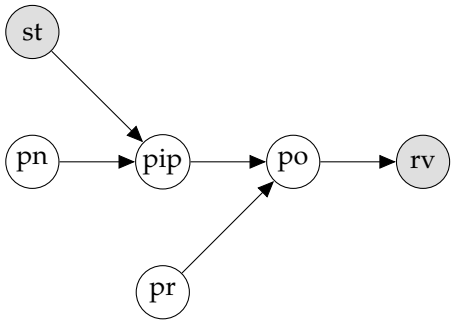


Figure 4.1: Upper panel: the variance of the stimulus varies as a function of the duty cycle of A, reaching a maximum where the duty cycle is 50%. Lower panel: the variance of a Bernoulli random variable as a function of p. As can be appreciated from the curves in the upper and lower panels, they differ only by a constant $M^2/100$, where M is the total length of the morph line.



$$\text{st} \sim \mathcal{N}(\mu_{st}, \Sigma_{st}) \text{ (stimulus)}$$

$$\text{pn} \sim \mathcal{N}(\mu_{pn} = 0, \Sigma_{pn}) \text{ (perceptual noise)}$$

$$\text{pip} \sim \mathcal{N}(\mu_{pip}, \Sigma_{pip}) \text{ (perceptual input)}$$

$$\text{pr} \sim \mathcal{N}(\mu_{pr}, \Sigma_{pr}) \text{ (prior distribution in facespace)}$$

$$\text{po} \sim \mathcal{N}(\mu_{po}, \Sigma_{po}) \text{ (posterior distribution)}$$

$$\text{rv} \sim \mathcal{N}(\mu_{rv}, \Sigma_{rv}) \text{ (response distribution/variability)}$$

Figure 4.2: A graphical model of perceptual inference in face space and specification of the nodes. Shading denotes an "observed" node, in this case the stimulus (st) and the response (rv) of the subject. The remaining unobserved, or latent, nodes represent the random variables representing perceptual noise (pn), perceptual input (pip), the prior distribution over face space (pip), and the posterior distribution over face space (po). See figure 4.4 for a list and description of the free parameters within the model. A property of any random variable is that it is associated with a probability distribution. See sections 4.4 and 4.5 for details of how these distributions are parameterised.

4.4 Parameter Descriptions and Interpretations

There follows an enumeration and description of the seven parameters with the Bayesian model of perception which are fitted using MCMC sampling. See figure 4.2 for a description of the dependency structure of the model in the form of a graphical model of the perceptual process. In section 4.5 these parameters are related mathematically to the graphical model depicted in figure 4.2.

p_1: perceptual noise factor (a nonnegative real value). Determines the magnitude of intrinsic perceptual, and presumably chiefly neuronal, noise. Relates to the random variable 'pn' in figure 4.2.

p_2: prior distribution in face space factor (a nonnegative real value). Determines the width or "strength" of the isotropic Gaussian prior in face space. Relates to the random variable 'pr' in figure 4.2.

p_3: window of integration (a positive integer). Given some Gaussian distribution for the prior and for the perceptual input this parameter actually determines the number of samples that are taken from the perceptual input distribution, pip $\sim \mathcal{N}(\mu_{pip}, \Sigma_{pip})$, in computing the posterior distribution over face space, and thereby represents a discretised window of time, sometimes also referred to as a window of integration. Relates to the random variables 'po', 'pr' and 'pip' in figure 4.2.

p_4: isotropic stimulus variance weight (a nonnegative real value be-

tween 0 and 1). This parameter defines a continuum between the model in which the brain is able to appreciate that the variance in the stimulus occurs within a single axis and the model in which the brain assumes that total variance in the stimulus is attributable to all axes equally and is therefore, so to speak, shared equally among them. Relates to the random variable 'st' in figure 4.2.

p_5: perceptual noise form ratio (a nonnegative real value). This parameter determines the form of the perceptual noise, and represents the ratio of the eigenvalues of the covariance matrix. The effect is to allow the perceptual noise to adopt either an isotropic ($p_5 = 1$) or an anisotropic form ($0 \leq p_5 < 1$). Relates to the random variable 'pn' in figure 4.2. The basis of this feature of the model is the observation that at a duty cycle of 0% or 100% the subject is matching a static, unchanging stimulus and yet the distribution of judgements suggests that there is an increased variance along the radial axis as compared to the tangential vector (i.e. a fitted Gaussian forms a cigar shape orientated towards the origin appreciable in figure 3.9). This phenomenon, a relative insensitivity to radial versus tangential change, has been previously observed [Ross et al., 2010].

p_6: response bias magnitude (a nonnegative real value). This parameter represents a multiplicative bias in the distribution of judgements, which is evident both in the dynamic (figures 3.5 and 3.10) and the contrast (figure 3.13)experiments. Relates to the random variables 'po' and 'rv' in figure 4.2.

p_7: response variance factor (a nonnegative real value). This parameter scales the covariance matrix of the response distribution. It is based on the idea that there exists within the brain a posterior distribution on which the motor response is based. This could result in a response distribution with either a greater or a lesser total variance than the posterior itself. For example, if the posterior were inefficiently utilised then the distribution might be very broad, whereas if it were computed several times, i.e. repeatedly sampled, then it might result in a response distribution somewhat narrower than the posterior (i.e. with a lower total variance).

4.5 Model Definitions and Specifications

It will now be made explicit how exactly the relevant quantities (the parameters of the distributions) for the model are computed. Regarding the stimulus, as described in chapter 3, this is on any given trial in the dynamic experiment a deterministic stimulus with a fixed and regular duty cycle. All the same we can compute the mean and variance for any condition and treat it henceforth *as if* it were a truly random variable.

$$\mu_{st} = [\bar{x}, \bar{y}]^\top = \frac{1}{T} \sum_{t=1}^{T} \mathbf{f}_t \tag{4.3}$$

Where the x and y variables correspond to the radial and tangential coordinates used to determine a position on the 2D-slice through face space described in chapter 3. T represents the number of milliseconds in a duty cycle, either 100 or 200. Similarly, the covariance matrix of the stimulus,

Σ_{st}, is determined by the following expressions

$$\mathrm{Var}[x] = \frac{\sum_{i=1}^{T}(x_i - \bar{x})^2}{T} \tag{4.4}$$

$$\mathrm{Var}[y] = \frac{\sum_{i=1}^{T}(y_i - \bar{y})^2}{T} \tag{4.5}$$

$$\mathrm{Cov}[x, y] = \frac{\sum_{i=1}^{T}(x_i - \bar{x})(y_i - \bar{y})}{T} \tag{4.6}$$

Explicitly then, the covariance matrix is constructed as follows

$$\Sigma_{st} = \begin{bmatrix} \mathrm{Var}[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[x, y] & \mathrm{Var}[y] \end{bmatrix} \tag{4.7}$$

Moving on to the random variable 'pn' we suppose that perceptual noise is unbiased in terms of its mean, so

$$\mu_{pn} = \mathbf{0} \tag{4.8}$$

However, we incorporate into the model the possibility that the form of the noise may be anisotropic, with the long axis of the ellipse parallel or orthogonal to the radial axis. In other words, we suppose that the perceptual noise may be such that the level of variability towards and away from the origin or norm in face space need not be equal to the variability tangential to a sphere in face space. In chapter 5 subsection 5.2.3 I will present evidence, from a separate experiment, that noise in the radial direction is indeed significantly greater than in the tangential direction, a finding that

is consonant with previous research [Ross et al., 2010]. To return to the immediate subject however, three parameters are required: one to control the ratio of the eigenvectors of the anisotropic covariance matrix, i.e. the form of the perceptual noise, one to control the overall spread (entropy) of the distribution and one to control the balance between an isotropic and an anisotropic estimate of the form of the noise in the stimulus. The material question motivating this aspect of the model is whether the brain is able to appreciate that the variance in the stimulus occurs in a single axis, tangential to a (hyper-) sphere in face space, and use this information in producing its estimate of stimulus location in face space? Or, is the brain only able to appreciate the overall variance in the stimulus, $\mathrm{tr}(\Sigma_{st})$, apportioning it isotropically, as it were, to all axes equally. In mathematical terms the stimulus variance contribute to the perceptual input according to equation 4.9 or 4.10.

$$\Sigma_{pip_{iso}} = \mathrm{tr}(\Sigma_{st}) \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + p_1 \Sigma_{pn} \tag{4.9}$$

$$\Sigma_{pip_{aniso}} = \frac{1}{2} \mathrm{tr}(\Sigma_{st}) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + p_1 \Sigma_{pn} \tag{4.10}$$

Rather than deciding this arbitrarily opting for one model or another we can parameterise a continuum between the two as per equation 4.11.

$$\Sigma_{pip} = p_4 \Sigma_{pip_{iso}} + (1 - p_4) \Sigma_{pip_{aniso}}) \tag{4.11}$$

The prior over face space is assumed to be an isotropic Gaussian, fol-

lowing Valentine and others [Valentine et al., 2016] and is a consequence of the assumption of isotropy underlying the whole experimental setup. In chapter 5 this assumption is tested directly and found to be false, strictly speaking. However the anisotropy is not severe enough to completely undermine the merit of our approach, since clearly consistent judgements are made despite the error of this assumption. Finally a single parameter serves to parameterise the "strength" of the Gaussian prior in face space, which is an isotropic Gaussian.

$$\Sigma_{pr} = p_4 \cdot \mathbf{I} \tag{4.12}$$

With the prior and the perceptual input distributions defined we are now in a position to combine these so as to obtain the posterior distribution, through the process known as Bayesian inference. Accordingly [Murphy, 2012] the mean of the posterior is give by

$$\boldsymbol{\mu}_{po} = \left(\boldsymbol{\Sigma}_{pr}^{-1} + p_3\boldsymbol{\Sigma}_{pip}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_{pr}^{-1}\boldsymbol{\mu}_{pr} + p_3\boldsymbol{\Sigma}_{pip}^{-1}\mu_{\mathbf{pip}}\right) \tag{4.13}$$

And the posterior covariance is given by

$$\boldsymbol{\Sigma}_{po} = \left(\boldsymbol{\Sigma}_{pr}^{-1} + p_3\boldsymbol{\Sigma}_{pip}^{-1}\right)^{-1} \tag{4.14}$$

Naturally, one supposes that this posterior distribution is the representation used by the nervous system to orchestrate a motor response, manipulating the matcher to resemble the dynamic stimulus in this case. However while the distribution of responses should utilise this represen-

tation it does not follow that it need be the same, and this seems to be for at least two reasons. Firstly, for reasons which are not completely clear, there appears to be a small but consistent bias in subjects' judgements, such that they consistently place the matcher at a position in face space slightly more centripetal than the target. This is evident in both the DE and CE (Dynamic Experiment and Contrast Experiment, see chapter 3). Despite the apparent oddness of this feature of the data it is nevertheless a prominent and statistically significant one, and the model therefore incorporates a "response bias" parameter, p_6. Thus the mean of the response distribution within the model is given by

$$\boldsymbol{\mu}_{rv} = p_5 \boldsymbol{\mu}_{po} \tag{4.15}$$

Furthermore we suppose that the total variance of the response distribution could either be greater than, less than or indeed equal to, that of the posterior. This could be due to corruption by noise within the motor system, inattention on the part of the subject, which would tend to increase the total variance. Alternatively a resampling process whereby the posterior was computed repeatedly could result in a response distribution with a lower total variance than the posterior distribution. In any case as we are not in a position to adopt anything but an agnostic position on this question currently so the covariance matrix for the response distribution is parameterised so as to allow for these possibilities.

$$\boldsymbol{\Sigma}_{rv} = p_7 \boldsymbol{\Sigma}_{po} \tag{4.16}$$

4.6 Model Parameter Estimation

MCMC (Markov Chain Monte Carlo) [Andrieu et al., 2003] was used to fit the parameters of the model to the data, assuming uniform distributions over the allowable range of the parameters. Of the many flavours of MCMC that could be adopted I opted for an approach based on the Metropolis algorithm [Bolstad, 2010], mainly for reasons of simplicity and transparency in implementation and presentation. This approach allows one to samples from a probability distribution $P(\mathbf{x})$ in the absence of an explicit representation, so long as one can evaluate a function $f(\mathbf{x})$ which is proportional to it. In this case $P(\mathbf{x})$ is the joint distribution with support over 7 dimensional parameter space. Here $\mathbf{x}$ therefore corresponds to a 1 by 7 vector. The function $f(\mathbf{x})$ was defined as the squared symmetric KL-divergence between the empirical response distributions seen in experiments and the predicted response distributions from the perceptual graphical model depicted in figure 4.2. Thus, any choice of values for the parameter vector, $\mathbf{x}$, implies a set of 11 predicted response distributions (one for each duty cycle percentage { 0%, 10%, ..., 100% }), which we denote R_m as opposed to the set of 11 response distributions established empirically R_e. [2] So,

$$R_m = \{\mathcal{N}_{m1}, \mathcal{N}_{m2}, \ldots, \mathcal{N}_{m11}\} \tag{4.17}$$

[2] Due to the inherent symmetry of the experimental condition (e.g. duty cycle percentages of 10% and 90% are equivalent) this in fact reduces to 6.

and,

$$R_e = \{\mathcal{N}_{e1}, \mathcal{N}_{e2}, \ldots, \mathcal{N}_{e11}\} \tag{4.18}$$

From this we define the fucntion $f(\mathbf{x})$ to be the the squared sum of the symmetric KL-divergence between the empirically derived distrbutions,R_e, and the model derived distributions R_m for some $\mathbf{x}$. i.e.

$$f(\mathbf{x}) = \sum_{i=1}^{11} \left(D_{\mathsf{KL}}(\mathcal{N}_{mi}||\mathcal{N}_{ei}) + D_{\mathsf{KL}}(\mathcal{N}_{ei}||\mathcal{N}_{mi})\right)^2 \tag{4.19}$$

Fortunately the KL-divergence can be computed efficiently for two multivariate normal distributions, say $\mathcal{N}_a = \mathcal{N}(\mu_a, \Sigma_a)$ and $\mathcal{N}_b = \mathcal{N}(\mu_b, \Sigma_b)$, and, where k is the dimensionality of the distributions, is as follows

$$D_{\mathrm{KL}}(\mathcal{N}_a \| \mathcal{N}_b) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_b^{-1}\Sigma_a\right) + (\mu_b - \mu_a)^\top \Sigma_b^{-1}(\mu_b - \mu_a) - k + \ln\left(\frac{\det \Sigma_b}{\det \Sigma_a}\right)\right). \tag{4.20}$$

Using the KL-divergence in this way to estimate the target distribution represented a significant computational benefit and can be evaluated in constant time with respect to the size of the dataset, n, and therefore linear in the number of MCMC iterations, m, i.e. $\mathcal{O}(m)$. In contrast using the likelihood of the data requires evaluating each datum under a multivariate normal and is therefore linear in the size of the data, n, and in the number of MCMC iterations, m, i.e. $\mathcal{O}(mn)$. The practical implication for MCMC was that using the likelihood was approximately an order of magnitude slower and produce comparable results in terms of the quality of the MCMC. This accounts for why a somewhat atypical function (the

sqaured KL-divergence) was used to estimate the target distribution over the parameters of the model.

4.6.1 MCMC for Parameter Estimation

This section details the particular incarnation of the MCMC used for fitting the model to the data. Where a random vector was sampled from parameter space each element was sampled independently from a uniform distribution over a fixed interval corresponding to a plausible range. The suitability of the intervals chosen was confirmed by the marginal distributions obtained, which strongly suggest that the support in each case is a subset of the parameter intervals chosen for sampling. It remains of course possible that the support only partially intersects with the chosen intervals, however this is a general issue with MCMC and not something peculiar to this application [Andrieu et al., 2003].

Algorithm 1 MCMC with squared symmetric D_{KL}

1: **procedure**
2: *Initialisation*
3: $\mathbf{x} \leftarrow$ random vector
4: *loop*:
5: $\mathbf{x}' \leftarrow \mathbf{x} +$ random vector
6: $\alpha = \frac{f(\mathbf{x}')}{f(\mathbf{x})}$
7: $u \sim \mathsf{U}(0, 1)$
8: **if** $\alpha > u$ **then** $\mathbf{x} \leftarrow \mathbf{x}'$

This algorithm was implemented in the Matlab [3] programming language. The algorithm was run for $200,000$ iterations of the MCMC loop shown in pseudocode. Of the parameter values sampled, those which re-

[3] R2015b

sulted in the minimum (squared symmetric) KL-divergence between the model's response distributions, R_m, and those established by experiment, R_e, were used as an estimate of the true maximum. The values thereby obtained are in themselves of limited interest, since, as I have mentioned previously, it seems highly unlikely that the numerical value of any parameter within the model relates directly to a biological parameter of significance. This is to be expected as the model was conceived with a proof-of-principle objective: can a Bayesian mode of inference explain the prominent bowing effect observed in the experimental data.

4.6.2 MCMC Validation

As described previously (section 4.6), the MCMC procedure allows one to obtain an estimate of the posterior distribution over parameter space. For my purposes I wish to obtain point estimates for my probabilistic model of inference, which I can do by taking the (estimated) maximum of the joint distribution over parameter space. MCMC is a widely used algorithm in part due to its very broad applicability across multiple model types, however this is also due to the empirical finding that it often works extremely well. The word "works" must here be taken with a pinch of salt since although MCMC does come with asymptotic guarantees this is almost never the case in practical use setting where sampling must of necessity be finite. As a consequence there are several standard techniques for assessing the quality of the MCMC procedure.

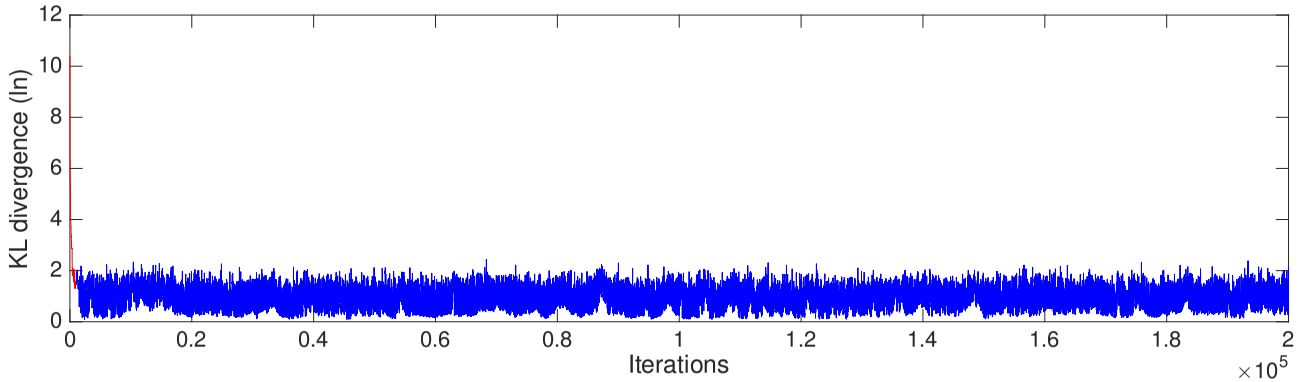


Figure 4.3: Convergence of MCMC. Logarithmic KL-divergence (ordinate) is plotted against iterations (abscissa) The burn-in period (red) and convergent period (blue) for a single chain in the MCMC procedure is shown. Evidently the convergence is very rapid. The burn-in period as a steep red line on the far left of the figure (iterations 1-1000)

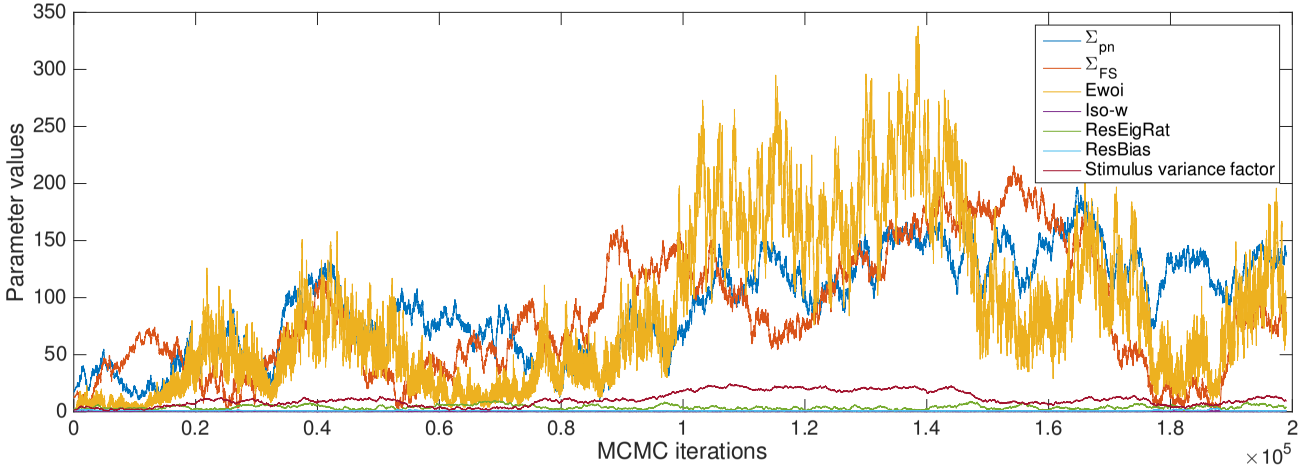


Figure 4.4: Samples obtained during 200,000 iterations of the MCMC algorithm. Some of the parameters, such as the effective window of integration (Ewoi), were sampled from across a wide interval, while others tended to remain confined to a relatively small interval.

4.6.3 MCMC Results

Figures 4.8 4.9 and 4.10 illustrate the Bayesian step in the probabilistic model of face perception presented here. For each duty cycle between 0 and 50 % the prior, likelihood and posterior are shown, figures 4.8 4.9 and 4.10 respectively. In each panel iso-probability ellipsoids [4] are shown for the complementary functions. In each case the prior is represented by cyan ellipsoids (actually circles since the prior is isotropic), the posterior by black ellipsoids, and the likelihood by red ellipsoids. Several features are of note. Firstly one can observe that the form of the likelihood (representing the distribution of the data plus perceptual noise) varies systematically from a duty cycle of 0% to 50%, becoming broader and more isotropic. This corresponds to the variability in the data increasing from a minimum at a duty cycle of 0% to a maximum at 50%, as discussed in section 4.3. In contrast, but as we would expect, the prior distribution remains constant. The interaction then of the prior and likelihood is seen in the variation of the posterior distribution across duty cycles, where we again see a systematic change from narrow elongated distributions at 0% to broad and increasingly isotropic distributions at 50%.

Thus, the model is successful in reproducing the so-called bowing seen in the data, at least in the sense of reproducing a centrifugal bias at more extreme duty cycles and a centripetal bias at intermediate duty cycles. This is the pattern seen in the DE, though of course not in the CE, an issue discussed in section 4.7.

[4] Defined as the ellipsoid which circumscribes 39% of the probability mass

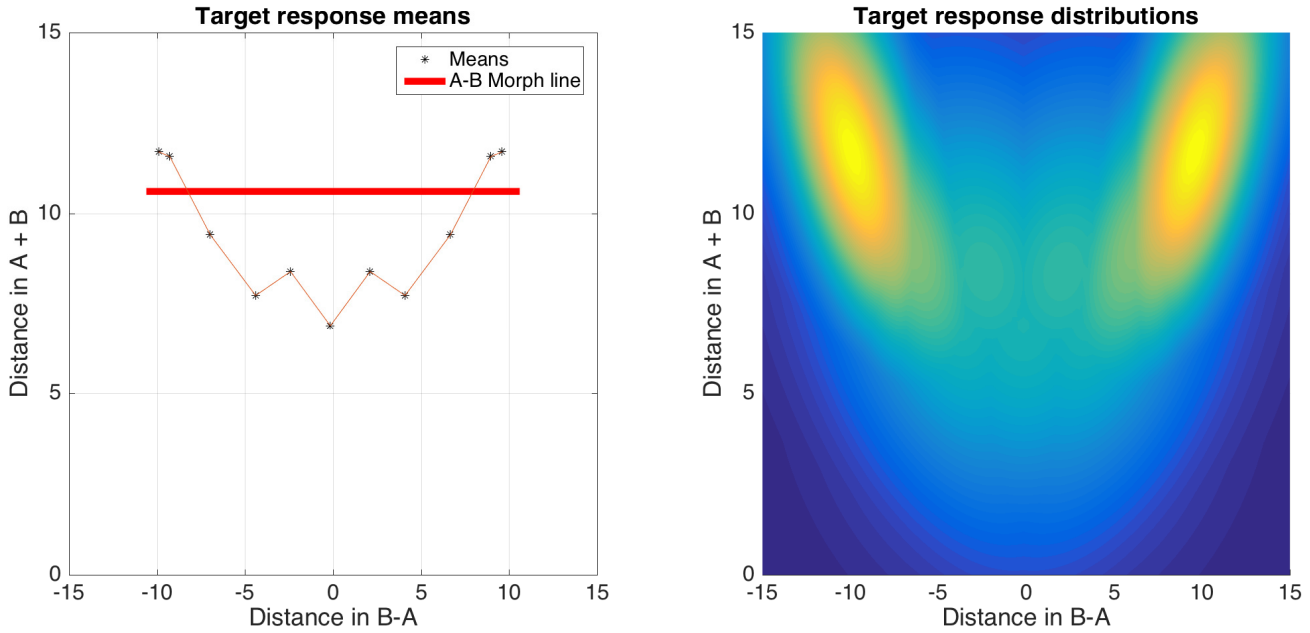


Figure 4.5: cf. figure 4.6The left panel shows the means for all duty cycles (black asterisks), with data from both the 100ms and 200ms conditions. The blue line connecting the points represents a simple linear interpolation. The thick red line shows the morph line connecting faces A and B. The right panel shows the fitted Gaussian distributions for all duty cycles. The mean of each fitted Gaussian corresponds to one of the dark asterisks in the left panel. Note that Gaussians with the lowest total variance are most prominent as their density is greater at their peak (i.e. at their means))

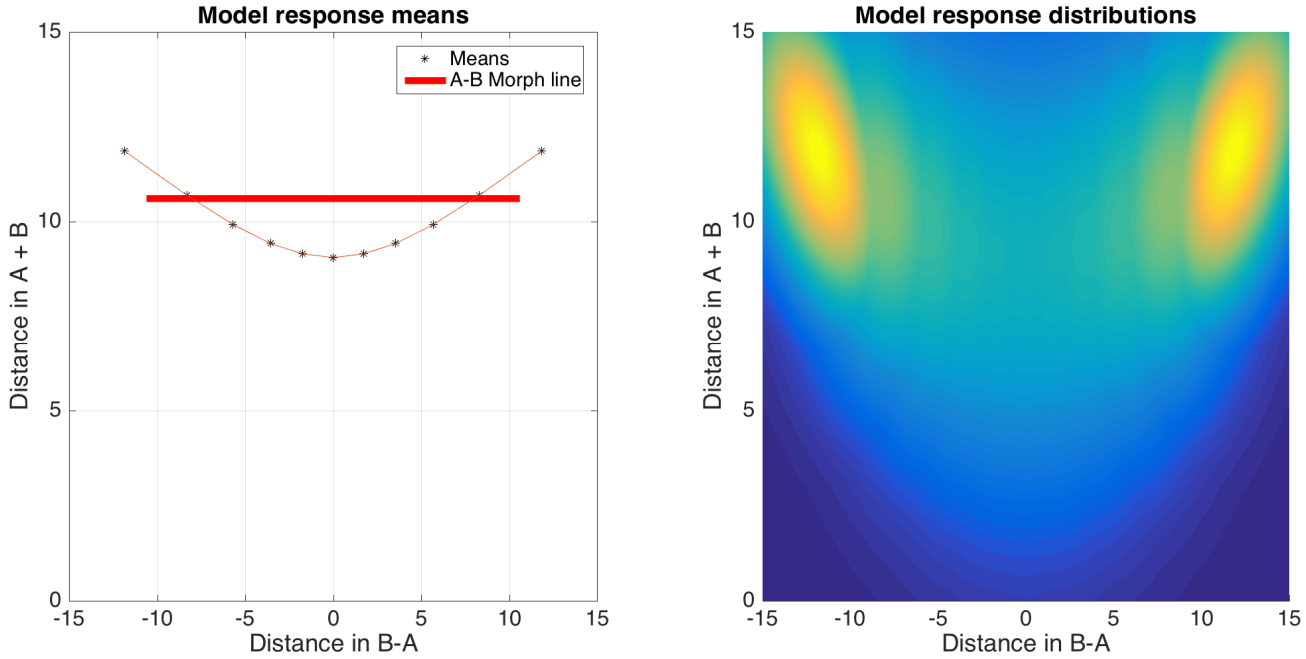


Figure 4.6: cf. figure 4.5.The left panel shows the means for all duty cycles (black asterisks), taken from the Gaussian response distributions predicted by the fitted model. The blue line connecting the points represents a simple linear interpolation. The thick red line shows the morph line connecting faces A and B. The right panel shows the predicted Gaussian response distributions for the fitted model. The mean of each fitted Guassian corresponds to one of the dark asterisks in the left panel. Note that Gaussians with the lowest total variance are most prominent as their density is greater at their peaks (i.e. the means))

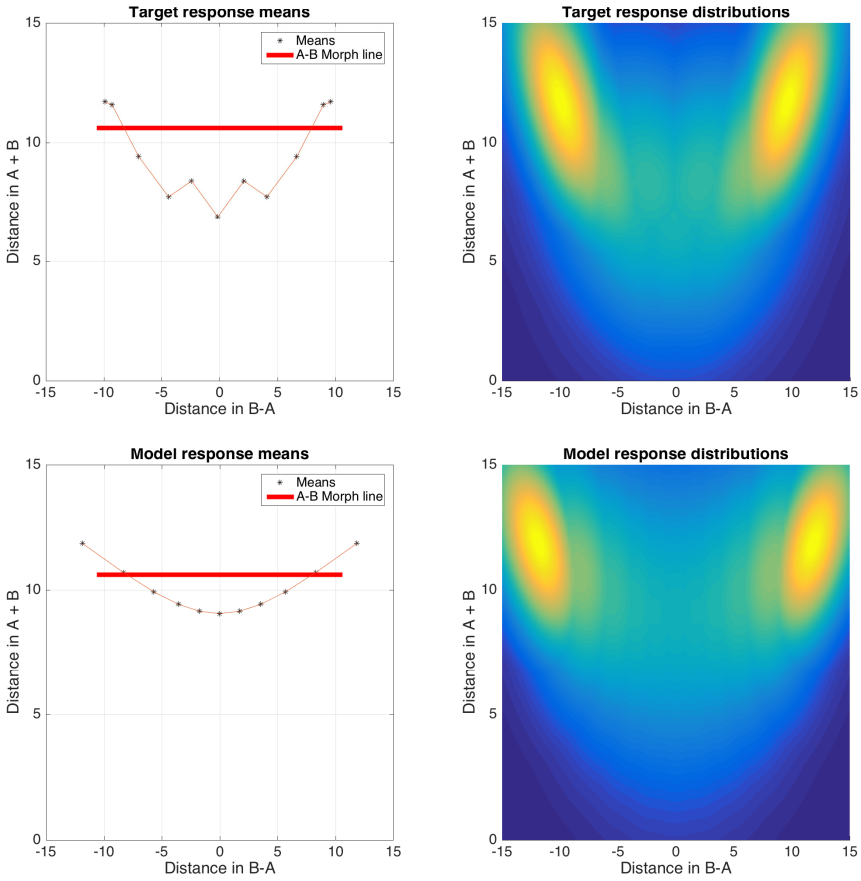


Figure 4.7: cf. figures 4.5 and 4.6. The upper and lower two panels are exact reproductions of those in figures 4.5 and 4.6 respectively, presented again here for ease of comparison. While there is obviously some discrepancy between the two some basic features of the two sets of response distributions are the similar. That is, a small centripetal bias most evident at duty cycles of 0% and 100% and an increasing ”bowing”-effect towards the origin at increasingly intermediate duty cycles, reaching a maximum magnitude at 50%.

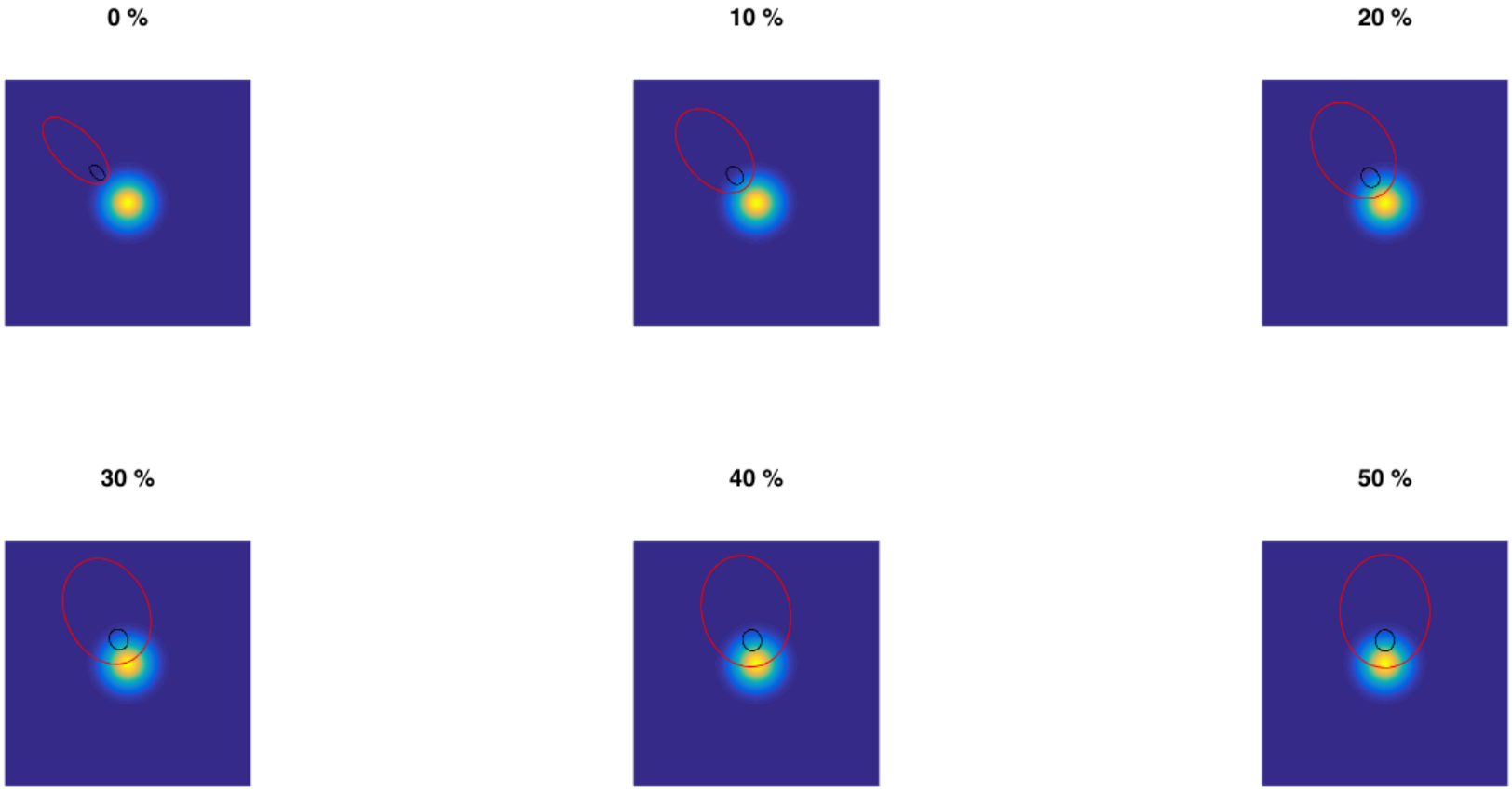


Figure 4.8: Prior distribution over face space, with the (normalised) likelihood (red) and posterior (black) represented by iso-probability ellipsoids. Each panel relates to a specific duty cycle between 0% and 50%. Note that the prior is identical for each duty cycle. cf. figure 4.9 and 4.10

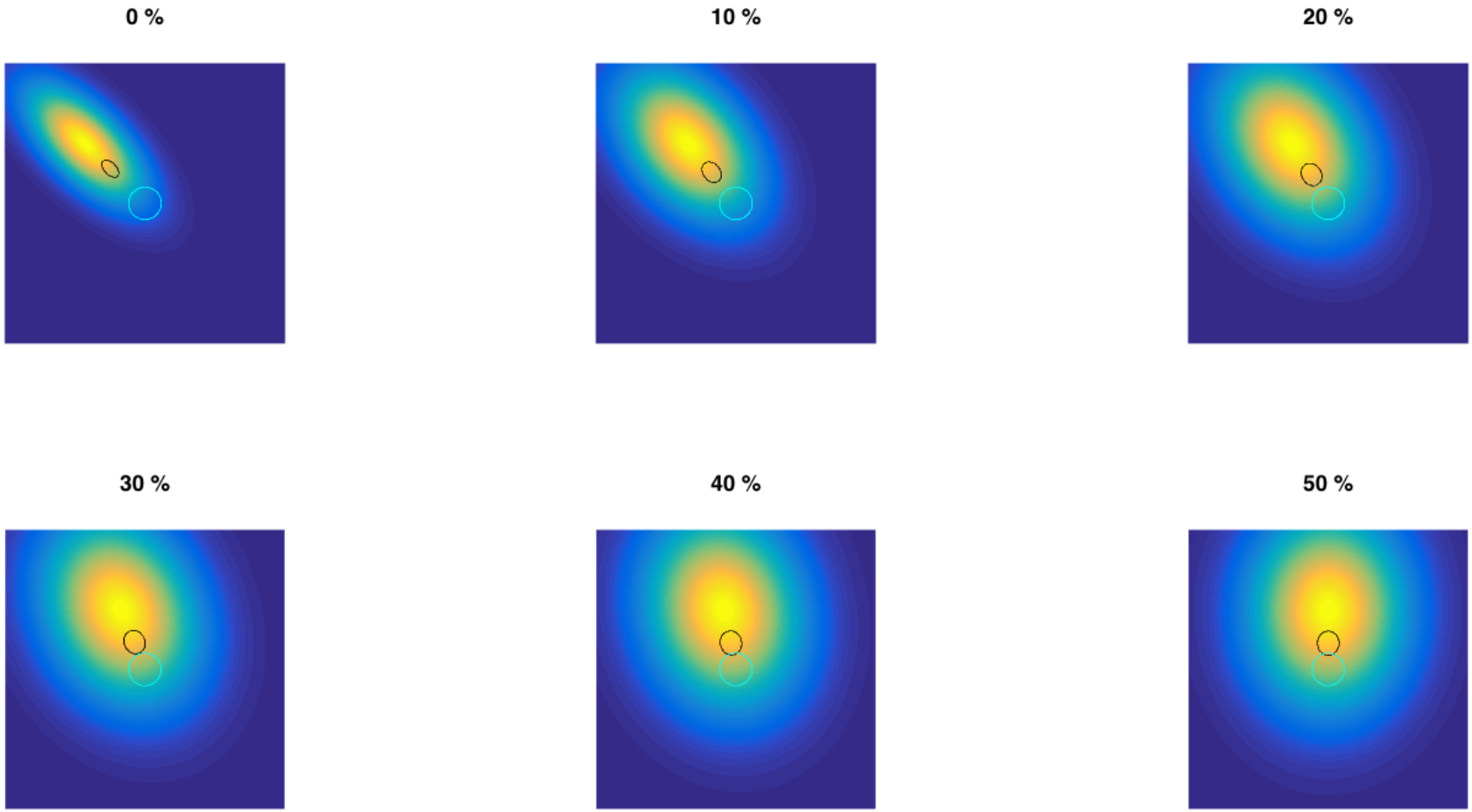


Figure 4.9: Normalised likelihoods within face space representing the distribution of the data, with the prior (cyan) and posterior (black) represented by iso-probability ellipsoids. Each panel relates to a specific duty cycle between 0% and 50%. Note that the prior is identical for each duty cycle. cf. figure 4.8 and 4.10

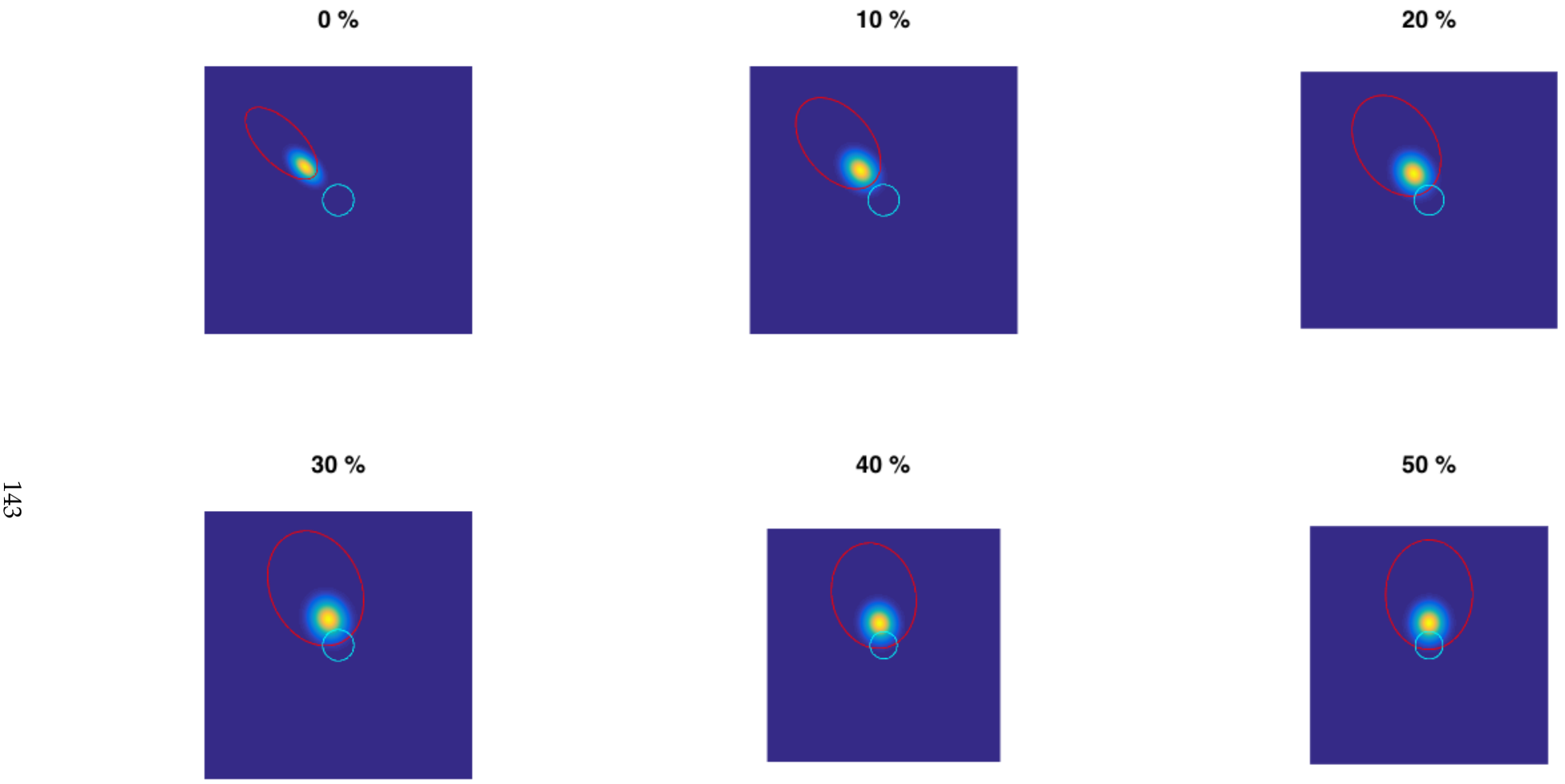


Figure 4.10: Posteriors within face space resulting from the combination of the prior and likelihood according to Bayes' rule. The prior (cyan) and (normalised) likelihood (red) are represented by iso-probability ellipsoids. Each panel relates to a specific duty cycle between 0% and 50%. Note that the prior is identical for each duty cycle. cf. figure 4.9 and 4.8

4.7 Discussion

How convincing is the preceding model for the curious results presented in chapter 3? In particular how convincing an explanation is this for the bowing, the bias in judgements towards the norm at intermediate duty cycles and away from the norm at extreme duty cycles, which appears moreover to be an inherently dynamic effect (section 3.2.2)? Our faith in the proposed explanation might be shored up were we able to posit a plausible neuronal mechanism by which this effect might arise. One possibility that will be explored at some length is that lateral inhibition may provide a mechanism which could account for at least some of the psychophysical results presented in chapter 3.

Lateral inhibition has a long history in the study of mind, with the seeds of the concept discernible in the writings of Descartes and contemporaries [Jacobson, 1993]. However, it began to be studied physiologically and psychophysically only in the 20th Century, through the seminal work of Georg von Békésy [Békésy, 1967] and others. Physiological and anatomical studies demonstrate that lateral inhibition permeates the mammalian visual system from the retina through to cortical and deep brain structures, indeed it appears to be an essentially universal feature of complex nervous systems [Isaacson and Scanziani, 2011]. Despite these extensive phenomenological investigations there remains uncertainty about the precise functional role of inhibition within nervous systems. Is it to inhibit excitatory neurons that would otherwise destabilise into epileptic activity? Is it, as has been plausibly argued [Rolls and Treves, 1998], that lateral

inhibition is crucial for the implementation of winner-takes-all competition in biological neural networks? Doubtless, as commonly found with biological structures, multiple functions are supported. However, the recent ascendancy of Bayesian approaches to neural computation, reviewed at some length in chapter 1 section 1.4, provides a framework in which local lateral inhibition can be understood as a crucial functional component in probabilistic neural computation [Bill et al., 2015].

The concept of lateral inhibition is not new in the face recognition literature. For instance, the well known Interactive Activation and Competition (IAC) model [Burton et al., 1990] posits a connectionist architecture comprising pools of interactive simple processing units. Examples include Face Recognition Unit (FRU), dealing with visual recognition, Person Recognition Unit (PRU), involved in recognition from voice or other modalities, and Person Identity Unit (PIU), key for the representation of complex semantic information about an individual, beyond the modality of recognition (e.g. whether they are a family member). Between pools of these units there are excitatory links, allowing, say, the image of a person to trigger the relevant FRU and subsequently PIUs. However, within pools all units inhibit one another (i.e. laterally inhibit one another). In particular, regarding the model of face space developed in chapter 2, the IAC model supposes that all FRUs mutually inhibit one another. In terms then of a functional role, inhibition can be seen as enforcing the assumption that faces only possess a single identity, implementing a kind of winner-takes-all competition. Crucially, the experimental design of the DE violates this basic assumption, since by design it involves two rapidly alternating but

distinct entities.

In keeping with the IAC and other models allocating a central computational role to inhibition in face perception, an interesting mechanistic possibility for the centripetal radial deviation presented in chapter 2, is suggested by rather beautiful work done in insect decision-making by Seeley and colleagues [Seeley et al., 2012]. They studied populations of honey bees, *apis mellifera*, making binary decisions about which of two prospective nest sites to occupy. It had been well known for some time that bees use the waggle dance not only to direct to food sources but also to describe and thereby advocate a potential nesting site to other spectating bees. However, these researchers were directly inspired by the analogy with neurons [Passino et al., 2008] to look for inhibitory cross signals between bees advocating, by waggle-dancing for, a particular site and those advocating a different potential site. They found these cross-inhibitory signals in the form of high frequency head butts delivered bidirectionally between bees from competing parties, christened *stop signals* by the group. The effect of receiving such a stop signal was to reduce the probability that the recipient would continue repeating his waggle dance i.e. of his continuing his advocacy for a particular site.

As part of their investigations Seeley and colleagues considered a number of candidate dynamical system models of the decision making process, derived from their observations of bee behaviour. They found that the best explanation of their data was provided by a so-called *discriminate stop signal* model of decision making. Essentially this means that the stop signals are not issued at random but to those bees committed to a compet-

ing alternative, making the recipients more likely to become uncommitted. Symbolically, where A and B represent two species (bees in this case) committed to alternatives 'A' and 'B' respectively,

$$\mathrm{A} + \mathrm{B} \xrightarrow{\alpha_\mathrm{A}} \mathrm{A} + \mathrm{U} \tag{4.21}$$

and,

$$\mathrm{A} + \mathrm{B} \xrightarrow{\alpha_\mathrm{B}} \mathrm{B} + \mathrm{U} \tag{4.22}$$

where α_A and α_B are rate constants.

Using a van Kampen expansion [Van Kampen, 2007] they derive the following mean-field population-level differential equations.

$$\frac{\mathrm{d}\Psi_\mathrm{A}}{\mathrm{d}t} = \gamma_\mathrm{A}(1 - \Psi_\mathrm{A} - \Psi_\mathrm{B}) - \Psi_\mathrm{A}[\alpha_\mathrm{A} - \rho_\mathrm{A}(1 - \Psi_\mathrm{A} - \Psi_\mathrm{B}) + \sigma_\mathrm{B}\Psi_\mathrm{B}] \tag{4.23}$$

$$\frac{\mathrm{d}\Psi_\mathrm{B}}{\mathrm{d}t} = \gamma_\mathrm{B}(1 - \Psi_\mathrm{A} - \Psi_\mathrm{B}) - \Psi_\mathrm{B}[\alpha_\mathrm{B} - \rho_\mathrm{B}(1 - \Psi_\mathrm{A} - \Psi_\mathrm{B}) + \sigma_\mathrm{A}\Psi_\mathrm{A}] \tag{4.24}$$

Where Ψ_A and Ψ_B represent proportions of the population committed to option A and B respectively. And where represents γ_i, α_i, ρ_i and σ_i represent rate constants for spontaneous commitment, spontaneous abandonment, recruitment and stop-signal induced abandonment respectively (see supplementary material for reference [Seeley et al., 2012] for further

details).

The dynamics implicit in these differential equations can be visualised as a vector field, shown in figure 4.11. It can be appreciated that in a situation in which the ”best” option alternates, each transition will cause the state of the system to converge to the opposite fixed point, which will furthermore be reach via a curved, or bowed, trajectory. The suggestion might be then that the alternation of distinct faces pushes the perceptual state back and forth between the corresponding fixed points. However, when this alternation is rapid, as in the case of the DE, the dynamics of the system are such that instead of being suspended according to a linear interpolation between the stable states, the percept is suspended at a position biased towards the origin. Thus, by analogy, it seems possible that the underlying mechanism for the effect we see is some form of cross-inhibition between populations of neurons differentially sensitive to face A or face B.

The apparent similarity of decision making by populations of insects and neurons is one emphasised by Seeley and colleagues [Seeley et al., 2012] in their original paper, so much so that they describe the phenomenon of (discriminate) stop signals as *cross-inhibition*, a term lifted unapologetically from the neuronal literature and a synonym for lateral inhibition. The functional importance of this cross-inhibition in the case of bees appears to be, *inter alia*, to rapidly break deadlocks. Without cross-inhibition two very similar alternatives can result in the state of the system hovering in equilibrium between the two options. In some models, such as drift diffusion models, of decision making a deadlock of this kind will even-

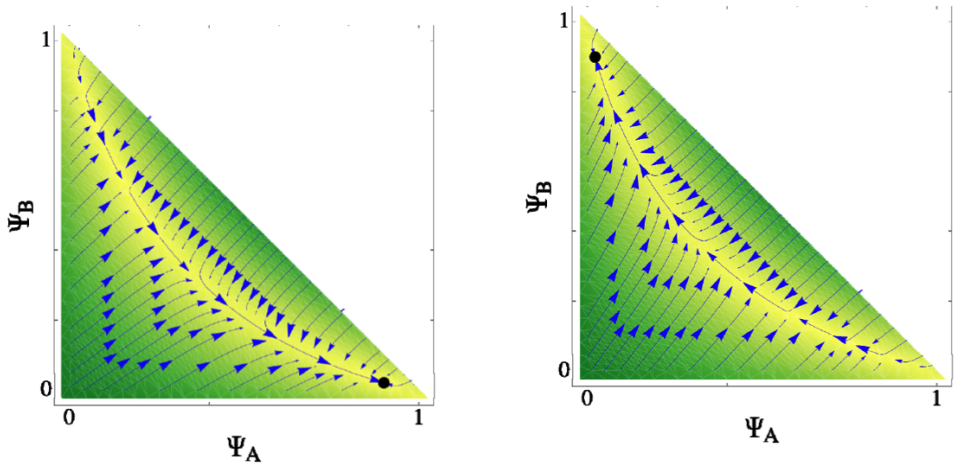


Figure 4.11: Illustration of the dynamics of a population of bees making a decision about two candidate nests, A and B, assuming a *discriminate stop signal model* of decision making (see reference [Seeley et al., 2012]). Bees can be in one of three states, committed to A, committed to B or uncommited. The horizontal and vertical axes correspond to the proportion of bees committed to option A and B respectively. The left panel illustrates the situation in which A is the preferable option. So bees are more likely to become committed to A and, when committed, waggle-dance more and issue more stop signals to bees committed to B. The right panel shows the reverse situation in which B is the preferable option. By analogy the situation in which A is dominant corresponds to stimulus/face A being presented, and likewise for B. Stable fixed points are represented by solid black dots, on which surrounding vectors converge. If one considers the situation in which A and B alternate then it is clear that the state of the system will shuttle back and forth along a curved, or bowing, trajectory between A and B's respective fixed points. If the alternation is sufficiently rapid, as in the DE, then the state will not have time to evolve fully to either stable state and will be suspended between the two. Moreover, the suspension will not correspond to a linear interpolation between states A and B, but will be biased towards the origin thanks to the effects of lateral inhibition. Adapted from [Seeley et al., 2012]

tually be broken by the effect of noise [Ratcliff and McKoon, 2008], but unless the magnitude of the noise is relatively large this will take a prohibitively long time. It may be argued that in most biologically relevant circumstances simply making a decision, even randomly, is much more adaptive that remaining paralysed by indecision between two similar alternatives. A clearer perceptual analogue of this effect may be the many well know bistable percepts, such as the Necker cube. Each interpretation is equally plausible, yet only one is "opted for" by our perceptual systems. Inhibitory neuronal activity is a crucial component of most modelling efforts to understand these phenomena, and recent work has provided evidence that this is indeed the case *in vivo* [van Loon et al., 2013].

A question therefore is whether a network of neurons, driven by specific stimuli, and connected by reciprocal inhibitory connections can effect Bayesian inference in a principled fashion. That this can indeed be done is supported by work over the past decade, largely in the form of modelling studies [Ma et al., 2006, Beck et al., 2008, Bill et al., 2015].

Even supposing that neuronal populations are capable in principal of implementing Bayesian computations it remains to be shown that this is in fact the path evolution has chosen. There is of course no *a priori* guarantee that the somewhat rarefied principles of probabilistic inference predominated in the evolution, for all the current vogue in touting this as a unifying framework for neuroscience. As Anderson observes, somewhat archly, in *The Adaptive Character of Thought*: "the gambler's fallacy may lead someone to lose money in Las Vegas, but if it leads him to try for a third child after two boys (because a girl is due), then it is quite adaptive" [Anderson,

1990, p. 32].

Finally, in this chapter and the modelling therein we have addressed only the bowing seen in the DE, and made no attempt to model the results of the CE. As demonstrated chapter 3 (see table 3.5 in section 3.3.4) there is no statistically significant evidence of a bowing effect in the CE, although there was evidence of a radial centrifugal effect as well as significant tangential deviations. The ability of the Bayesian model presented in this chapter to replicate, to some extent at least, the bowing seen in the DE stems chiefly from the combination of a variable likelihood with a strong prior. In order then to replicate results in which this bowing is not seen it is only necessary to eliminate the prior, or rather "flatten" the prior. Given that the model as currently described utilises a Gaussian prior it is unable to converge to a truly flat prior, but only approximates it better and better as the determinant of the covariance matrix increases. Thus an attempt to fit the model directly to the CE will not result in convergence. Having said that, however, Bayesian inference with a flat prior is equivalent to maximum likelihood, which is exactly how the CE data are modelled in figure 3.14 of chapter 3. Thus, a worthy avenue of further investigation would be to develop a single model that could account for both the DE and the CE data without special manipulation of core model features, such as the nature of the prior distribution in face space. The prior being, presumably, relatively static on the timescale of these experiments (i.e. seconds/minutes). It is likely that developing such a model would require modelling the dynamic nature of the stimuli itself, since it would be necessary to represent the DE's 200ms condition, the DE's 100ms condi-

tion and the CE within a single input representation. The CE can after all be thought of as a version of the DE in which the period has been reduced to 0 (i.e. infinitely fast alternations).

4.8 Conclusion

This chapter has demonstrated that a simple Bayesian framework can account for certain otherwise rather perplexing results arising from the experiments described in chapter 3. It has been convincingly demonstrated at lower perceptual levels that Bayesian inference does appear to be a widespread, if not ubiquitous, feature of perceptual inference [Knill and Pouget, 2004]. However, if Bayesian inference is to act as a unifying principle across domains of cognition it is crucial to demonstrate that it is a feature of high level object recognition. This model, and the data it appears to account for, therefore represents a step in that direction. I have previously noted the distance that exists between this model and the neuronal mechanisms which must underlie all computation within the brain. A natural direction for further research would thus be to explore how a neuronal mechanism could implement such a Bayesian computation in high dimensional face space. I have however drawn attention to some recent work in swarm intelligence, which suggests a possible class of neuronal mechanisms for this effect (i.e. cross-inhibition).

5

A Comparison of Human Face Space and Basel Face Space

The Basel Face Space (BFS) model has been utilised to perform the experiments reported in chapter 3 as well as in multiple other studies. In this chapter results are presented from an experiment which systematically probed the relationship between pairs of faces within BFS and the judgements human subjects made regarding their per-

ceived similarity. In assessing this relationship three central analyses are performed. Firstly, a number of possible functions relating the geometry of BFS to human subject similarity judgements are fitted, and it is found that, of those considered, a logistic function appears to produce the best correspondence. Secondly, the general question of isotropy is addressed, and it is shown that a significant proportion of the variance in responses is indeed explained by the absolute, as opposed to the relative, geometry of pairs of faces. Together these results constitute a quantification of the degree to which BFS is a good, albeit imperfect, model of human face space. Thirdly, a more specific issue around isotropy, which for clarity we might call *directionality*, is addressed, wherein it is shown that the gradient of dissimilarity between pairs of faces, as a function of BFS distance, is significantly greater in the tangential than the radial direction. In the discussion some tentative conclusions are drawn.

5.1 Introduction

The concept of face space has been extensively discussed in previous chapters and efforts have principally centred on the question of how information is integrated over time (chapters 2 and 3), perhaps conforming to Bayesian principles at some functional level (chapter 4). Throughout the theoretical and experimental investigations described in those chapters there has existed an implicit set of assumptions concerning face space, in particular that BSF (Basel Face Space) provides a reasonable approxi-

mation to the face space instantiated in the brain.

Recently, a body of work has arisen addressing the question of whether modern computer vision approaches to object recognition share essential characteristics with biological object recognition [Khaligh-Razavi and Kriegeskorte, 2014]. In this chapter this is explored by asking if one can predict human perceptual judgements from their *relative* geometry and, subsequently, whether *absolute* geometry in BFS is of additional significance. This concerns the question of isotropy, and if in general it matters from which angle a pair of faces are sampled, given some relative geometry. Persisting in this theme, a further issue relating to isotropy, for clarity referred to with the antonym *directionality*, in face space is then addressed, specifically whether there is a systematic difference between tangential versus radial change in terms of perceptual similarity.

Previous work has addressed the issue of isotropy in face space, for example [Ross et al., 2010]. Building on previous work (e.g. [Leopold et al., 2001]), they argued that the anisotropy suggested by their results supported a norm based coding model. Since then, however, further work has shown that exemplar encoding can yield effects that were previously thought to be characteristic only of norm based encoding schemes [Ross et al., 2014]. Moreover, as discussed in chapter 1, the norm vs exemplar debate, is not well motivated from a theoretical standpoint and indeed proponents of the norm-based coding scheme have yet to propose a coherent "norm" itself, in the sense of a metric. In this thesis an inference from the experimental results contained in this chapter and the underlying coding scheme is not attempted. It remains valid to note, however, that

while results such as these may not obviously imply a particular coding scheme (i.e. numerous models are consistent with these results) it does constrain the set of possibles models, and is not therefore a completely hopeless endeavour.

5.1.1 Experimental Design and Methods

This experiment addressed the question of whether reliable relationships could be found between the geometry of faces within the BFS model and the percept of similarity-versus-dissimilarity in human subjects. Pairs of faces from BFS were generated and present on a large 43" touchscreen display (Panasonic TH-43LFE8-IR), which allowed subjects to drag the pairs of faces, each face occupying approximately 1 degree of arc horizontally, and each face pair separated by the same, from a siding onto a large central arena for arrangement (see figure 5.1). Also present in the siding was an "identity" marker which the subject was instructed to place according to where he/she felt the demarcation At the beginning of each trial 8 randomly selected face pairs, along with the identity marker, were displayed in the siding. Once all 9 of these items had been placed in the arena the user terminated the trial by pressing a button marked "Done" and a new trial would begin (see figure 5.2). In a single session there were a total of 232 face pairs presented, 29 trials of 8 face pairs per trial. All 26 subjects completed 2 trials in which the same faces were presented but in a shuffled order, so that a given face pair was usually ($\approx 80\%$ of the time [1])

[1] Although on any trial in session 2 it was most probable that at least one of the $\binom{8}{2} = 28$ possible pairings of face pairs occurred together in a trial in session 1

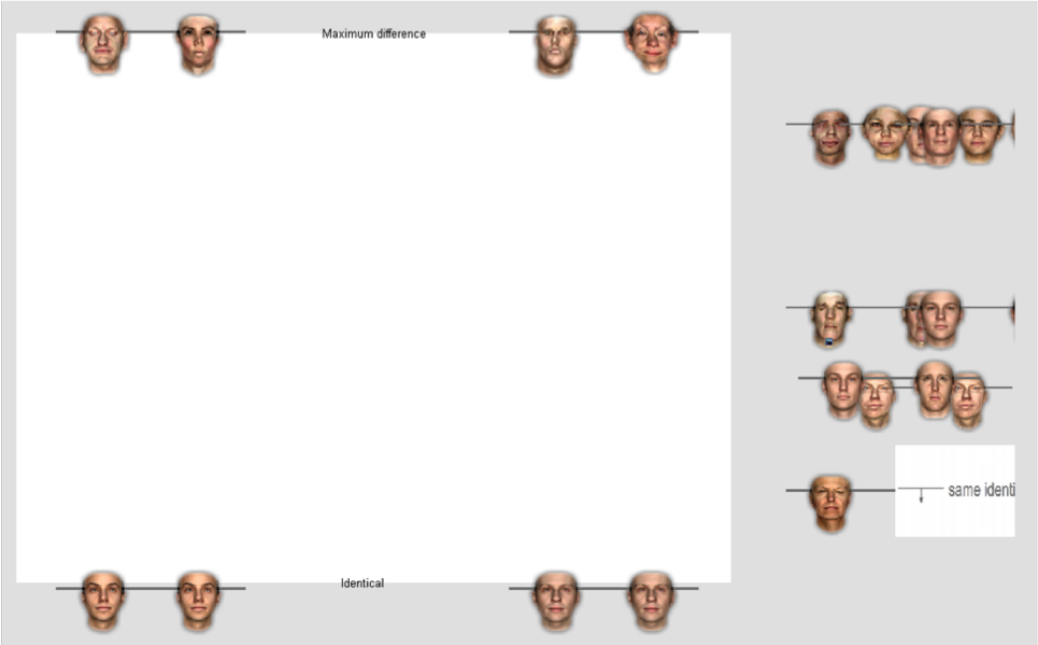


Figure 5.1: Example screen shot of experimental display at the beginning of each trial. The faces to be arranged are initially placed in a siding to the right of the main arena in which arrangement takes place. Note the anchor faces, which are pairs of identical faces (separation in BFS = 0) at the bottom of the screen and pairs of anti-faces (separation in BFS = 80) at the top of the screen. These face pairs cannot be moved but reinforce for the subject the perceptual continuum between extremes of similarity and dissimilarity.

accompanied by 7 different face pairs in session 2 compared to session 1.

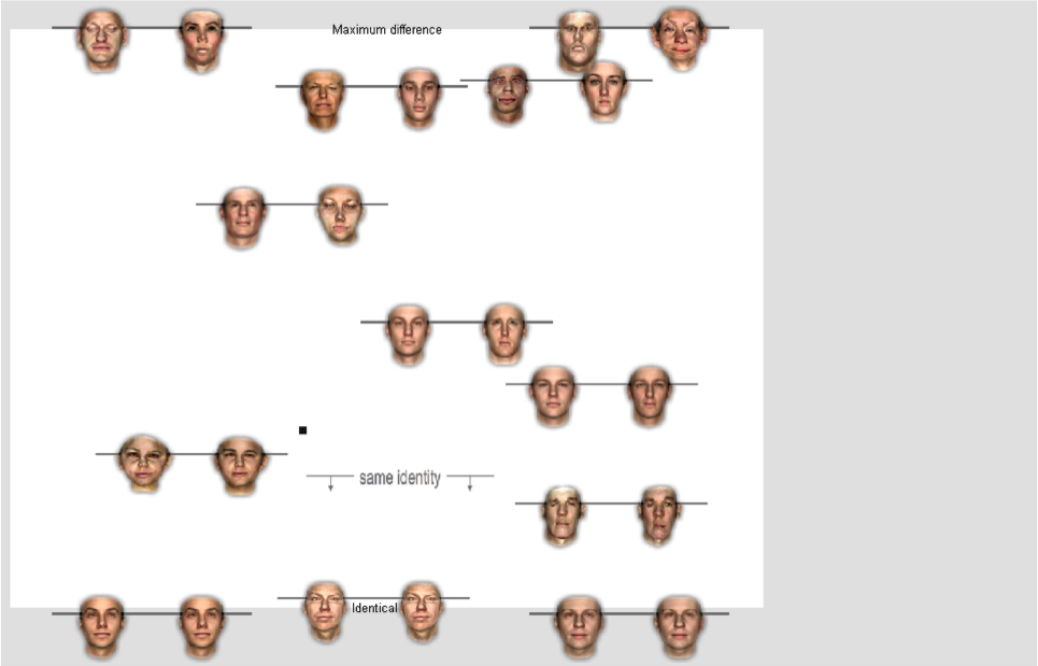


Figure 5.2: Example screen shot of experimental display at the end of each trial. Subjects can move the pairs from the siding across a large touch screen to any point in the arena using their fingers. Faces are deposited at a location on the arena corresponding to the internal similarity of each pair, and a marker denoting the point beneath which all pairs are considered to be of the same identity (i.e. the same person) is also positioned.

For a subset of 15/26 of the subjects a third session, session 3, was conducted in which 232 face pairs were presented in the same fashion as in sessions 1 and 2. However in this case although the relative geometry of the pairs remained the same the vectors were resampled from high dimensional face space so that on average they tended to be orthogonal compared to those presented in sessions 1 and 2. Indeed, because the space from which vectors are sampled is a 398 dimensional one can be extremely confident that in every case the inner product of the mean vector from sessions 1 and 2 with that in session 3 was essentially 0 (i.e. orthogonal). This manipulation allowed the question of just how much the intrinsic geometry of BFS determined the perceptual similarity to be addressed. If the relative geometry were of little importance, and absolute geometry was instead the chief determinant of perceptual similarity, then there should be a much better correlation between judgements in sessions 1 and 2 (which share both relative and absolute geometry) than between 1 and 3 or 2 and 3 (which share only relative geometry). On the other hand if relative geometry were the chief determinant of perceptual similarity then there should be comparable correlations between all sessions, since they share the same relative geometry.

Within the full Basel Face Model a particular face corresponds to a single 398 dimensional vector or, to put the same thing another way, a point in 398 dimensional real space $\mathbb{R}^{398}$. If we consider any two vectors then we can describe their absolute location with 2 such 398 element vectors. However their *relative* geometry can also be expressed by the lengths of the vectors r_1 and r_2 and the angle between them θ. Thus we can charac-

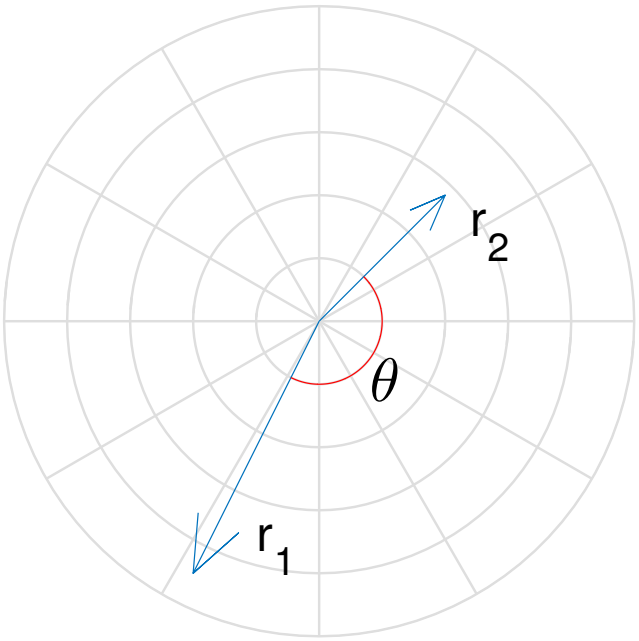


Figure 5.3: Any pair of vectors emanating from the origin lie in a 2D plane within high dimensional space. We can fully characterise their relative geometry in terms of three numbers, corresponding to the length of the two vectors, their radii r_1, r_2, and the angle between them θ.

terise the relative geometry of two faces within BFS compactly in terms of a tuple (r_1, r_2, θ) of three real numbers [2]. This characterisation of relative geometry is illustrated in figure 5.3.

Relative geometry in these terms can be visualised as a 3D parameter space wherein the 3 axes correspond to the three parameters r_1, r_2 and

[2]A tuple is a finite list of ordered elements. Notice that in fact r_1, r_2 need not be ordered since in terms of relative geometry (r_1, r_2, θ) is equivalent to (r_2, r_1, θ).

θ. Each point in this space corresponds to a definite angle and pair of vector lengths i.e. a particular relative configuration or geometry within BFS. Because of the relative nature of this characterisation there are then an infinite number of pairs of faces which satisfy any given configuration, excepting the situation in which both vectors r_1 and r_2 are the null vector.

Because the parameter space for relative geometry is continuous it would require impracticably large quantities of data to characterise it fully by empirical methods, so it is necessary to select some sampling scheme whereby a full characterisation can be approximated by making certain assumptions, like that the function changes only relatively slowly w.r.t. parameter space. Secondly, one is confronted with the issue of whether the sampling should take place in BFS or in the parameter space itself. There are many considerations to reflect upon in this regard, but an illustrative difficulty is to consider how well our parameter space would be characterised were we to sample i.i.d. pairs of faces randomly from a 398 dimensional Gaussian within BFS. A moment's reflection reveals that this would result in a dense sampling of pairs of approximately orthogonal vectors (at $\theta \approx \pi/2$) and extremely sparse sampling of pairs forming more acute or obtuse angles due to the very high dimensionality of BFS. For reasons such as this a deterministic grid-sampling procedure was adopted, conducted within *parameter* space rather than BFS itself. Figures 5.4 and 5.5 illustrate the scheme settled upon, an 8-by-8-by-8 regular grid of parameter space.The reason why a denser grid than 8 per axis was not used is that for a density, d, the number of unique points within such a grid is proportional to d^3. The reason why the number of unique points (i.e. unique relative

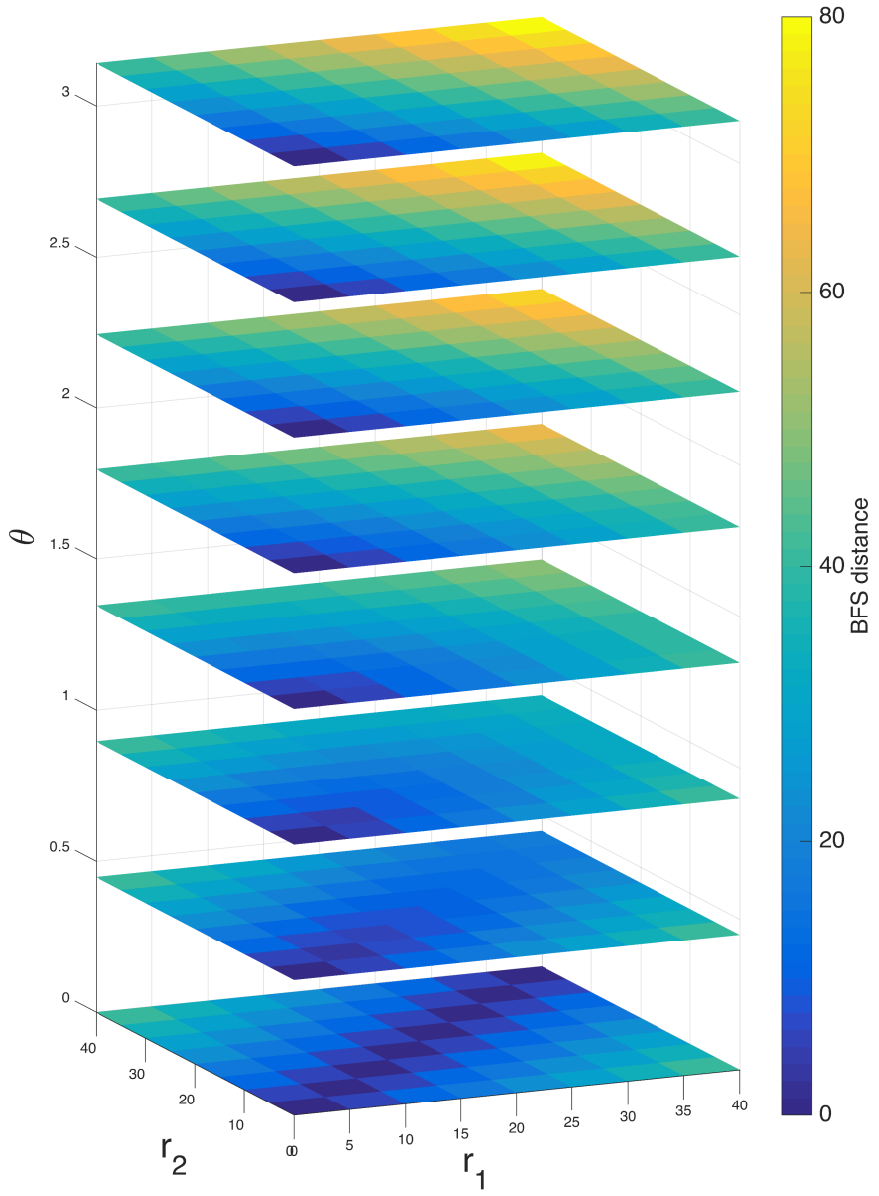


Figure 5.4: The relative geometry of two vectors can be represented as a point in three dimensional parameter space. This figures shows parameter space with axes, r_1 r_2 and θ. r_1-by-r_2 slices are stacked along θ (in radians), representing an 8-by-8-by-8 grid sampling of three dimensional parameter space. Colours from deep blue to bright yellow represent Euclidean distance in BFS. See also figure 5.5 which shows these slices laid out individually from the bottom ($\theta = 0$) up.

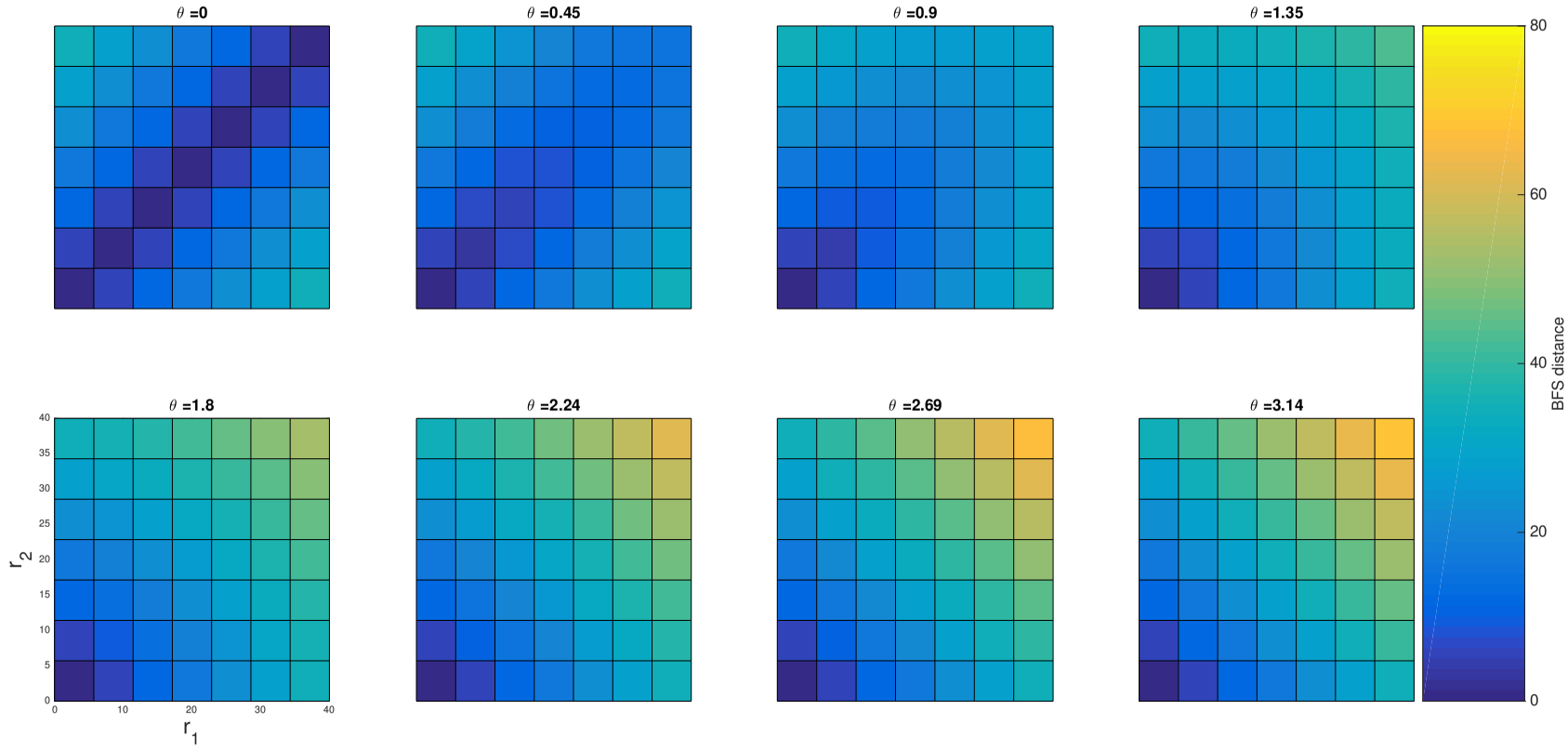


Figure 5.5: This figures shows the same r_1-by-r_2 slices as shown in figure 5.4 with the corresponding value of θ printed above each slice.

geometries) is not exactly d^3 is that there is redundancy within the parameter space, which is consequently non-uniform in the sense that if we were to simply sample each point from the grid some n number of times then there would again be some relative geometries which would be sampled more often than others. For example there is only one point in parameter space corresponding to the situation in which the separation between two points (a pair of faces) in BFS is equal to 80, when the tuple (r_1, r_2, θ) = $(40, 40, \pi)$. In figure 5.4 this is the furthest (and most yellow) square in the topmost slice ($\theta = \pi$); in figure 5.5 the top right point in the bottom right panel. In contrast for any situation in which $r_1 \neq r_2$ the tuples (r_1, r_2, θ) and (r_2, r_1, θ) are geometrically equivalent in BFS but correspond to different points in parameter space. In fact they are reflections of one another in the plane $r_1 = r_2$. These redundancies correspond to the axis of symmetry (r_1=r_2) seen in all of the slices in figure 5.5 and to the fact that the most leftward columns and bottom rows are identical within and between the slices. There is another axis of symmetry ($r_1 + r_2 = 40$) in the upper left panel of figure 5.5, but this does not in fact correspond to a redundancy in relative geometry. Instead, this represents faces separated only in terms of eccentricity (i.e. the angle between the vectors is 0). Once these redundancies have been eliminated it turns out that a total of 232 unique points characterise parameter space.

5.2 Results

5.2.1 Predicting Human Judgements from Basel Face Space

Were BFS (Basel Face Space) a perfect representation of human face space then we would expect to find a linear relationship connecting the distance between two faces in BFS and the subjective dissimilarity to a human subject. In order to probe this issue we fitted a number of functions to the data acquired in our experiments. We fitted both linear and non-linear models.

These were of the following form, where d_h denotes human dissimilarity judgements and d_b denotes Euclidean distance in Basel Face Space.

Firstly,

$$d_h = m \cdot d_b + c \tag{5.1}$$

a simple linear model, which corresponds to the situation in which Basel face space is an essentially perfect model of human face space.

Next we have a sigmoid function

$$d_h = \frac{1}{1 + \exp\{-m \cdot d_b + c\}} \tag{5.2}$$

which is essentially the linear regression model passed through a "squashing function".

Next we have a logarithmic function

$$d_h = \ln(m \cdot d_b + c) \tag{5.3}$$

and then an exponential function

$$d_h = \exp(m \cdot d_b + c) \tag{5.4}$$

These functions were all fitted using Matlab's inbuilt non-linear fitting tool, and the results can be seen in figure 5.6.

Finally I fitted a further function, but which took as inputs three variables: the length of two vectors and the angle between them, denoted r_1, r_2, and θ. A linear combination of these three variables (plus a bias term a_3), passed through a sigmoid function yielded the output of the function. I call this the "polar model", since the inputs reflect the relative polar coordinates. This had the following form

$$f(\mathrm{r}_1, \mathrm{r}_2, \theta) = \frac{1}{1 + \exp\{a_0 \cdot \mathrm{r}_1 + a_1 \cdot \mathrm{r}_2 + a_2 \cdot \theta + a_3\}} \tag{5.5}$$

This function cannot be represented in the same was as those preceding it and displayed in figure 5.6, since it takes not one but three arguments.

In order to assess the goodness-of-fit the residuals were for each model, along with the MSE (Mean Squared Error), computed using 4-fold cross validation. Both the residuals and the MSEs are displayed in figure 5.7. Regarding the MSE the sigmoid model surpasses the other models, closely followed by the linear model. This is supported by the proportion of variance explained by each model (i.e. R^2), printed above each corresponding bar. With regards to the residuals only the logistic, the linear and the polar models could be construed visually, at first blush, as approximating a Gaussian. The other two models (logarithmic and exponential) show

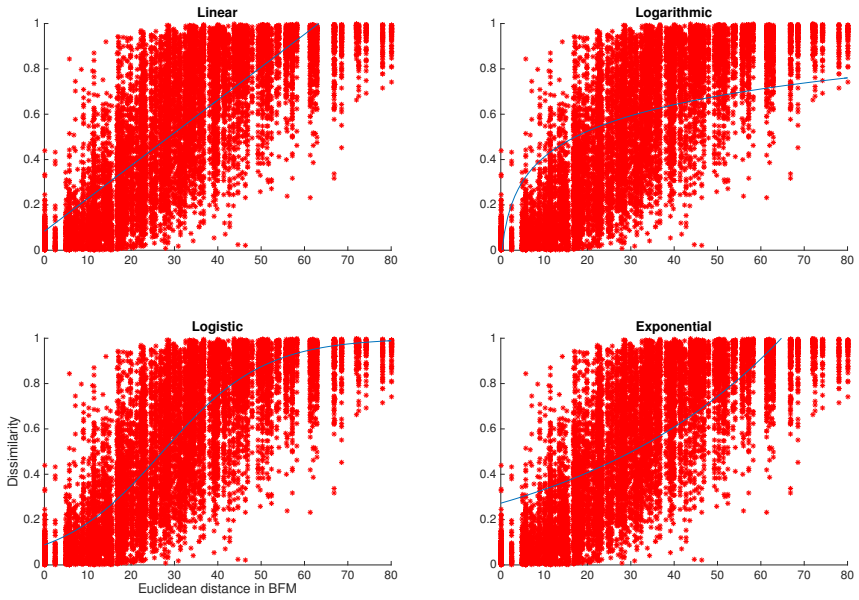


Figure 5.6: All judgements for all subjects. Euclidean distance is plotted on the abscissa against perceptual dissimilarity on the ordinal. In each panel a different function has been fitted to the data. From top left to bottom right these correspond to the linear model eq. 5.1, the logarithmic model eq. 5.3, the sigmoid/logistic eq. model 5.2, and the exponential model eq. 5.4.

clearly non-Gaussian signs, such as being bimodal and /or skewed. However, an Anderson-Darling test for normality (see table 5.1) does not support the assumption of normality for any of the models, so strictly none of them passes muster in terms of producing Gaussian residuals. That said, the relative magnitudes of the test statistic in each case suggests that the polar (2.7), logistic (18.7) and linear (30.6) models are much closer to normality that the exponential (115.2) or logarithmic (221.7), so in this sense the models could be meaningfully ranked.

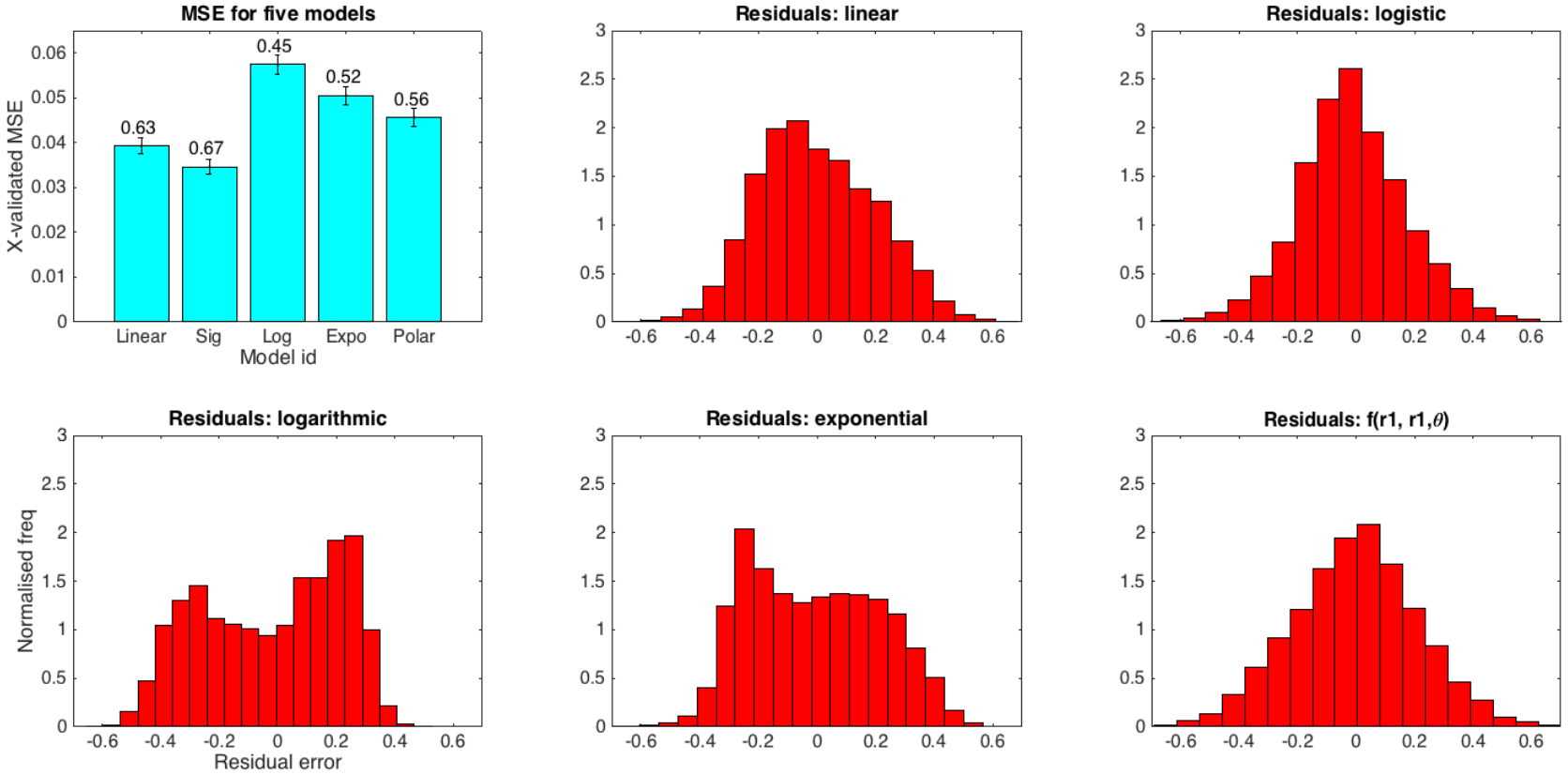


Figure 5.7: Upper left panel: MSE (Mean Squared Error) for five models relating Basel Face Space representations to perceptual dissimilarity judgements. Error bars are standard errors. From left to right within the top left figure the Mean Squared Error for each of 5 functions/models fitted to the data and evaluated using 4-fold cross-validation. From left to right these correspond to equations 5.1, 5.2, 5.3, 5.4 and 5.5 respectively. Above each error bar the proportion of variance explained by the model is printed in black, i.e. R^2. The red histograms from upper left (the upper middle panel being first) to bottom right correspond to the panels from top left to bottom right in figure 5.6. See table **??** for results of an Anderson-Darling test for normality of the residuals.

Model	H_0	AD test statistic	p-value
Linear	1	30.6	<0.001
Logistic	1	18.7	<0.001
Logarithmic	1	221.7	<0.001
Exponential	1	115.2	<0.001
Polar	1	2.7	<0.001

Table 5.1: Results of an Anderson-Darling test for normality of residuals. The null hypothesis of normally distributed residuals is rejected for all of the models at a significance level of 0.05. The test statistic, given in column three, provides some insight into the strength of the evidence against normality in each case. The polar model yields the smallest, while the logarithmic model yields a test statistic nearly two orders of magnitude greater.

In summary, of the models assessed none yields convincingly normal residuals, as assessed by the Anderson-Darling test for normality (table 5.1). However, the linear, the logistic and the polar models, assessed by the magnitude of the Anderson-Darling test statistic, are much closer to normality than the exponential and logarithmic models. The MSE analysis reinforces this point quantitatively, in that the models displaying the *most* non-Gaussian residuals also display the highest error, but suggests that overall the logistic model is slightly superior to the linear and polar models. However, a complication is that superior model as assessed by MSE (logistic) is different from the superior model as assessed by the Anderson-Darling test statistic (polar). Suffice to say that this is not the first time a psychometric curve has approximated a sigmoid (assessed by MSE), however it is interesting that a model based on polar coordinates provides a good fit in terms of the normality of residuals. Further work could seek to elucidate the apparent tension between model performance assessed by MSE and by the Anderson-Darling test. Overall, this analysis

supports the contention that there really is shared structure between biological face space and BFS. Subsequent sections will further explore and quantify this relationship, in particular by addressing questions around isotropy or directionality in face space.

5.2.2 Relative and Absolute Geometry in Biological and Basel Face Space

The experimental design, whereby a subset of 15 of the subjects repeated the experiment with identical geometry but a resampled set of particular faces, allowed us to test a certain aspect of isotropy. In particular, it allowed us to ask the following question: does the relative geometry of two faces in face space account for the variability we see in judgements, beyond the effects of noise in subjects' judgements? Were relative geometry the sole determinant of perceptual similarity, then the correlations between a pair of sessions in which the same faces are used (i.e. constant relative and absolute geometry across sessions) and a pair of session in which different faces are used, but relative geometry is held constant across sessions, should be the same. In contrast, insofar as absolute geometry is a determinant of perceptual similarity we should see a corresponding reduction in the correlation between two session in which the absolute geometry is changed, compared to sessions in which both the relative and absolute geometries are shared.

Figures 5.8 and 5.9 display the results of such a correlational analysis, using two separate measurements, the Pearson and the Spearman coeffi-

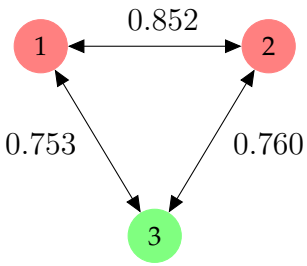


Figure 5.8: Pearson correlation coefficients between sessions 1, 2 (red) and 3 (green) for 15 subjects who completed an additional 3rd session in which the relative geometries of face pairs were preserved but resampled randomly from BFS. Note that the Pearson correlations between sessions 1 and 2 are significantly (see table 5.4) higher, 0.852 (0.8425, 0.8607), than those between sessions 1 and 3, 0.753 (0.7382, 0.767), and sessions 2 and 3, 0.760 (0.7454, 0.7735). cf. figure 5.9

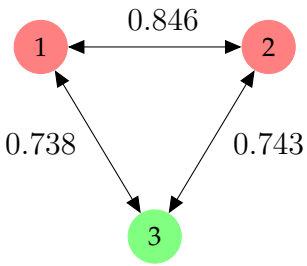


Figure 5.9: Spearman correlation coefficients (and 95% confidence intervals) between sessions 1, 2 (red) and 3 (green). The non-parametric Spearman correlations show the same patterns as the Pearson correlations. Spearman correlations between sessions 1 and 2 are significantly (see table 5.5) higher, 0.846 (CI: 0.8247, 0.8668), than those between sessions 1 and 3, 0.738 (CI: 0.7170, 0.7590), and sessions 2 and 3, 0.743 (CI: 0.7221, 0.7641). cf. figure 5.8

cients respectively. Sessions 1 and 2 (red) consisted of identical geometries and identical faces, while for session 3 (green) the relative geometry was preserved, but the faces were randomly resampled from Basel face space, being typically orthogonal therefore. Figure 5.8 displays the Pearson correlation coefficients between all three sessions, for all subjects.

There was a significantly lower correlation between sessions in which different faces were used, 0.753 (0.7382, 0.767) & 0.760 (0.7454, 0.7735), ver-

sus those in which the same faces were used, 0.852 (0.8425, 0.8607), suggesting the assumption of isotropy is, strictly speaking, false. The results of this analysis are presented in table 5.2 and an analogous analysis, based on the Spearman correlation and with the same conclusions, is presented in table 5.3.

Sessions compared	Pearson (CI)	H_0	p
1-2	0.852 (0.8425, 0.8607)	1	< 0.001
1-3	0.753 (0.7382, 0.767)	1	< 0.001
2-3	0.760 (0.7454, 0.7735)	1	< 0.001

Table 5.2: Pearson correlation coefficients (and 95% confidence intervals) computed, between all permutations of session pairs, for 15 subjects who completed all three sessions (1, 2 and 3). H_0 denotes the null hypothesis of a correlation of 0, and p the associated p-value. cf. figure 5.8

Sessions compared	Spearman (CI)	H_0	p
1-2	0.846 (0.8247, 0.8668)	1	<0.001
1-3	0.738 (0.7170, 0.7590)	1	< 0.001
2-3	0.743 (0.7221, 0.7641)	1	< 0.001

Table 5.3: Spearman correlation coefficients (and 95% confidence intervals) computed, between all permutations of session pairs, for 15 subjects who completed all three sessions (1, 2 and 3). H_0 denotes the null hypothesis of a correlation of 0, and p the associated p-value. cf. figure 5.9.

Although by doing so we, strictly speaking, acquire no new information, we can obtain a slightly different perspective by converting Pearson coefficients into explained variance, R^2. Doing so we can say that the absolute geometry accounts for $R_a^2 = 0.852^2 = 0.73 = 73\%$ of variance, whereas relative geometry accounts for $R_r^2 = 0.75^2 = 0.57 = 57\%$. This leaves a residue of 16% of variance which can be attributed to the absolute geometry of pairs of faces. This is confirmed by an ANOVA for both Pearson and

Spearman correlations, the results of which are presented in tables 5.4 and 5.5.

Geometry	R^2	Explained variance (CI)	H_0	p
Absolute (A)	0.73	73 (71, 75)%	-	-
Relative (R)	0.57	57 (55, 59)%	-	-
Δ (A-R)	-	16 (13, 19) %	1	< 0.001

Table 5.4: ANOVA for absolute versus relative geometry based on estimates of Pearson correlation coefficients (presented in table 5.2). The null hypothesis, H_0, is that there is no difference between explained variances (i.e. that the $\Delta = 0$). cf. table 5.5

Geometry	R^2	Explained variance (CI)	H_0	p
Absolute (A)	0.72	72 (68, 75)%	-	-
Relative (R)	0.55	55 (52, 58)%	-	-
Δ (A-R)	-	17 (12, 21)%	1	< 0.001

Table 5.5: ANOVA for absolute versus relative geometry based on estimates of Spearman correlation coefficients (presented in table 5.3). The null hypothesis, H_0, is that there is no difference between explained variances (i.e. that the $\Delta = 0$) cf. table 5.4

16% perhaps provides a more intuitively graspable and memorable figure with which to quantify the size of the effect. The broad conclusion is that the absolute geometry does indeed add significant information over an above the relative geometry, but that this information gain is relatively small compared to the information already present in the relative geometry.

This conclusion, namely that absolute geometry does provide some additional information, is echoed by the analysis in which the Spearman correlation coefficients between sessions were computed, shown in table 5.3 and figure 5.9. The Spearman test makes no requirement of linearity, only

monotonicity, and is therefore less demanding in its assumptions. However, as can be seen the numbers are essentially indistinguishable (i.e. have overlapping confidence intervals) from those obtained in the Pearson analysis. This suggests that whatever the (presumably) monotonic function relating BFS to biological face space (for example those shown in figure 5.6) it is unlikely to affect the correlation analysis presented here.

It may therefore be reasonable to assume isotropy as a *biased* approximation. This is based on the result that the majority of the available information is contained within the relative geometry, but clearly not all. It would of course be important, when designing experiments using BFS to ensure that the experimental paradigm is not sensitive to this, relatively small, violation in the assumption of isotropy. At any rate, the fact that the violation is now quantified means that experimentalists are henceforth in a position to simulate their proposed design and assess whether isotropy is likely to be an issue in advance. The following section, i.e. 5.2.3, addresses a further issue related to isotropy, namely the equivalence, or otherwise, of tangential versus radial distance in BFS and biological or perceptual face space.

5.2.3 Tangential and Radial Distance in Basel and Perceptual Face Space

Figure 5.10 gives a summary of the combined similarity judgements obtained for all 26 subjects. Here each panels shows how similarity judgements vary with reference to a reference face at an eccentricity that in-

creases in steps from the "norm" face at the origin (upper left) to a point (i.e. a face) within the surface of the hyper-sphere bounding the sample space from which all faces were drawn (bottom right). The colours, cool for similar, warm for dissimilar, indicate how subjective dissimilarity varies with Euclidean distance in BFS, with respect to a reference face. An obvious question is: does perceptual similarity vary in the same way at the centre of BFS as at the periphery? We can begin to address this question by fitting a paraboloid bowl to the data for a given eccentricity, denoted by the scalar value l. At each eccentricity we are interested in the form of the paraboloid which has a value of 0 at the reference face so we can constrain the general form of the paraboloid as follows:

$$z = A(x - l)^2 + By^2 \tag{5.6}$$

where z corresponds to dissimilarity, x represents position on the axis of the radial vector ($x = \pm|\mathbf{r}|$), y the position on the orthogonal axis (i.e. parallel to the tangential vector) ($y = \pm|\mathbf{r}^{\perp}|$), and as stated previously l is a scalar representing the eccentricity of the reference face. A and B are the parameters which control the rapidity with which dissimilarity increases with distance in BFS. Higher values of A or B corresponds to steeper paraboloids, which can be appreciated by observing that the partial derivatives of z with respect to x and y are

$$\frac{\partial z}{\partial x} = 2Ax - 2Al \tag{5.7}$$

and

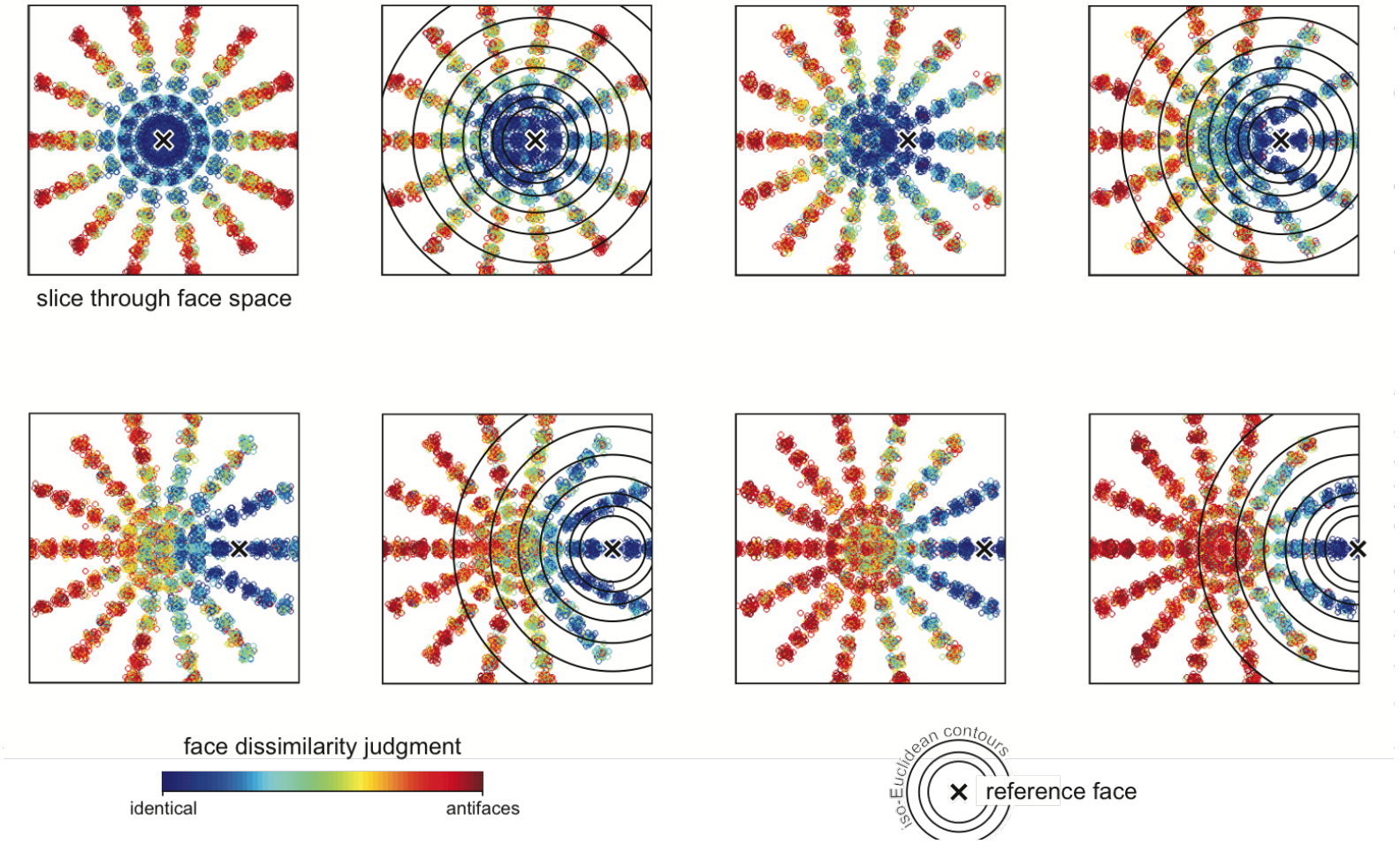


Figure 5.10: For a reference face (marked by an X) at a given distance from the average face (center), each panel shows how dissimilarity judgments between the reference and any other face vary across a slice through face space. Cool colours indicate similarly appearing faces, while warm colours dissimilar ones. The points reflect our regular face space sampling grid in polar coordinates. Black circles (shown for every other slice) indicate iso-Euclidean-distance contours with respect to the reference face. See also figure 5.11

$$\frac{\partial z}{\partial y} = 2By \tag{5.8}$$

So the second partial derivatives are

$$\frac{\partial^2 z}{\partial x^2} = 2A \tag{5.9}$$

$$\frac{\partial^2 z}{\partial y^2} = 2B \tag{5.10}$$

So we can see that the curvature in x is proportional to A, while the curvature in y is proportional to B. In order to obtain this figure however, we first need to fit the model to our data.

Because formula 5.6 is linear in the parameters A and B we can fit directly using closed form linear regression. i.e. where

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} \tag{5.11}$$

$$\mathbf{X} = \begin{pmatrix} x_1^2 & y_1^2 \\ x_2^2 & y_2^2 \\ \vdots & \vdots \\ x_n^2 & y_n^2 \end{pmatrix} \tag{5.12}$$

and

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \tag{5.13}$$

Then the (least mean squares) estimate of the parameters $\hat{\beta}$ is given by the closed form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{z} \tag{5.14}$$

Using this formula to fit a series of appropriately constrained paraboloids to the data we obtain a pair of parameter values for each eccentricity ($|r|$). The fitted paraboloids are plotted in three dimensions at each eccentricity $|\mathbf{r}|$ in figure 5.11 and the corresponding parameter values are plotted as a function of $|\mathbf{r}|$ in figure 5.12. In figure 5.12 we can thus appreciate how the curvature of the fitted paraboloid decreases with increasing eccentricity. In keeping with previous findings, which utilised a different experimental approach, [Ross et al., 2010] the curvature is, at all positive eccentricities, i.e. $|\mathbf{r}| > 0$, greater in the tangential direction than in the radial. Another way of putting this might be to say that subjects are less sensitive to changes in the radial direction than to those in the tangential direction. This observation is verified and quantified by the analysis presented in figure 5.12.

A further nuance, that can be appreciated visually from figure 5.12, is that not only are subjects less sensitive to change in the radial than the tan-

gential direction, but the magnitude of the difference becomes greater and greater with increasing eccentricity. This can be appreciated by expressing the corresponding curvatures as ratios. The phenomenon is quantified in figure 5.13, where a highly significant linear increase is seen in the ratios of curvatures expressed as a function of eccentricity (Pearson CC = 0.95; $p < 0.001$). This was something that was not made explicit in previous results [Ross et al., 2010], and indeed it is only possible to see this effect as a consequence of the novel methodology here adopted. It is an open question, and a worthy avenue of further study, as to what coding scheme's could account for this feature of the data.

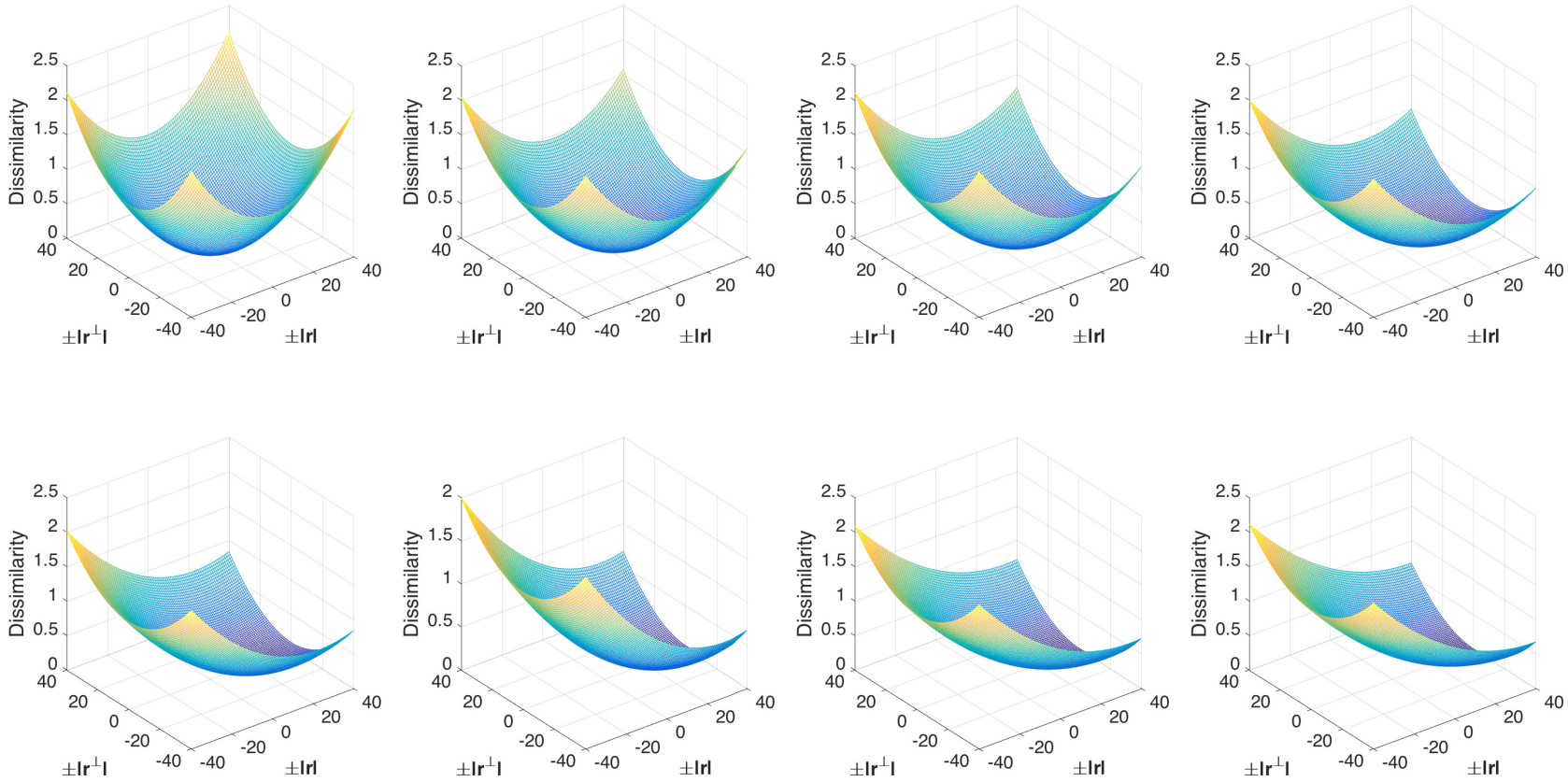


Figure 5.11: Fitted paraboloids, constrained to have a value of 0 at the position of the reference face. From top left to bottom right the reference face is located at regular steps between the norm (at the origin) and a face on the surface of the sampled hypersphere on the axis $\pm|\mathbf{r}|$. See also figure 5.12

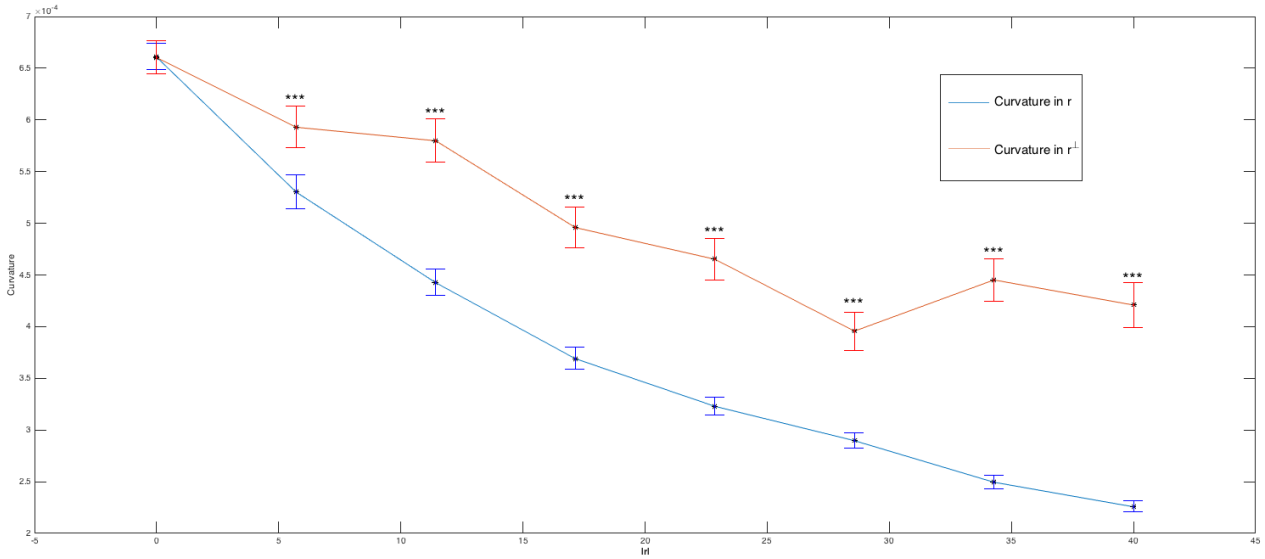


Figure 5.12: The fitted values of parameters A and B, see equation 5.6,plotted against the eccentricity of the reference face (l). Where $l = 0$ the reference face is at the origin and so all directions are equivalent, by our assumption of isotropy. The paraboloid here is therefore a so-called a paraboloid of revolution. With increasing eccentricity however the values of A and B decrease, corresponding to broader, less curved paraboloids. Of note the broadening is greater in radial direction $\mathbf{r}$ than in the tangential direction, $\mathbf{r}^{\perp}$, which can be appreciated visually by a close examination of figure 5.11. This is in keeping with a previous finding that humans are more sensitive to change in the tangential than the radial direction [Ross et al., 2010]. Error bars are 95% confidence intervals obtained by bootstrapping. 2 sample t-tests were performed to identify instances where the parameter estimates deviated at statistical significance. $***$ denotes significance at $p < 0.001$. $**$ denotes significance at $0.001 < p < 0.01$.$*$ denotes significance at $0.01 < p < 0.05$. The absence of an asterisk implies insufficient evidence against the null hypothesis. All p-numbers are adjusted for multiple comparisons using the Bonferroni correction (number of independent comparisons = 6). See also table 5.6

Eccentricty	Curvature in r (CI)	Curvature in $r^{\perp}$ (CI)	H_0	p
0	0.6607 (0.6509, 0.6770)	0.6602 (0.6476, 0.6806)	0	0.478
5.7	0.5301 (0.5127, 0.6074)	0.5929 (0.5074, 0.6123)	1	< 0.001
11.4	0.4425 (0.4329, 0.5946)	0.5798 (0.4247, 0.6026)	1	< 0.001
17.1	0.3690 (0.3586, 0.5072)	0.4959 (0.3501, 0.5159)	1	< 0.001
22.9	0.3229 (0.3126, 0.4773)	0.4652 (0.3005, 0.4887)	1	< 0.001
28.6	0.2897 (0.2824, 0.4016)	0.3957 (0.2712, 0.4125)	1	< 0.001
34.3	0.2495 (0.2425, 0.4500)	0.4452 (0.2276, 0.4640)	1	< 0.001
40	0.2257 (0.2197, 0.4284)	0.4209 (0.2032, 0.4449)	1	< 0.001

Table 5.6: Results of t-test comparison of means using 29 independent estimates of parameter values (from 29 subjects). As would be expected, at the origin (eccentricity 0)all directions are radial by definition, and there is no statistically significant difference when (arbitrary) orthogonal directions are compared (allowing one of the pair to be considered tangential for the purpose of analysis). However, at eccentricities > 0 there is a reduction in curvature in both the radial and tangential directions, but a reduction which is much more pronounced in the radial than the tangential directions. (CI) denotes confidence intervals, H_0 the null hypothesis of no difference between means, and 0 or 1 in the fourth column denotes acceptance or rejection of H_0 repsectively. All curvature values are multiplied by 1,000 for clarity of presentation. See also figure 5.12

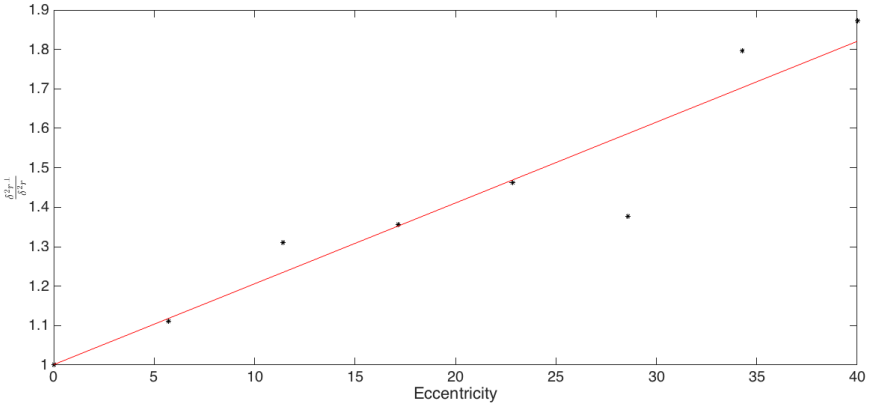


Figure 5.13: The ratios of curvatures (i.e. row 3 divided by row 2 of table 5.6) as a function of eccentricity. There is an approximately monotonic, positive and linear trend, suggesting that not only is the radial curvature less than the tangential curvature, but that this difference is magnified as a function of distance. This linear trend is highly significant (Pearson correlation coefficient = 0.95; $p < 0.001$)

5.3 Discussion

In subsection 5.2.1 the question was addressed of how well human dissimilarity judgements could be predicted from the "dissimilarity" in BFS. A number of different models we adduced, most of which utilised the BFS Euclidean distance between faces as the input, and outputted a prediction between 0 and 1 (0 being identical, 1 being antifaces). Consideration of both the MSE and the residuals would seem to favour the logistic and polar models, respectively, as the best approximations, as can be appreciated from figures 5.6 and 5.7. However, an obvious objection is that this analysis considered only five possible models (linear, logistic, logarithmic, exponential, and polar; see subsection 5.2.1). While it is impossible, in the absence of more information, to derive analytically the "true" function relating BFS to biological face space, one suggestion would be to train a neu-

ral network to perform the mapping between BFS biological face space, which, by the *Universal Approximation Theorem*, can approximate any continuous function [Jordan and Bishop, 2014]. However, quite apart from the absence of any guarantee that this function can in fact be learned, it is quite unclear what the success or failure of such as exercise would establish beyond the conservative conclusion that there is a highly non-random relationship between BFS and biological face space.

Subsection 5.2.2 addressed the question of the relative contribution of relative and absolute geometry to predicting perceptual judgements. It was found that the absolute geometry did indeed account for some of the explained variance, evidenced by a significantly lower correlation coefficient between sessions in which only the relative geometry was shared compared to sessions in which both absolute and relative geometry were shared. This was found to be true regardless of whether the Pearson correlation (see figure5.8) or the Spearman correlation (see figure5.9) was used, arguing that this is a fairly robust result.

An interesting finding from the analysis presented in subsection 5.2.3 is the observation that perceptual dissimilarity increases as a function of distance in BFS less rapidly in the radial direction (i.e. $\mathbf{r}$) than in the tangential direction (i.e. $\mathbf{r}^{\perp}$), depicted graphically in figure 5.12 and visually in figure 5.11. Visually it can be appreciated from figure 5.11 that the fitted paraboloids become shallower with eccentricity, and furthermore that the gradient in the radial axis flattens more slowly that that in the gradient in the tangential axis (see figure 5.13). A related finding, albeit obtained by a different experimental approach, was described in a 2010 paper by

Ross and colleagues [Ross et al., 2010]. They showed that discrimination between two faces was better when they lay on different radial vectors than when they both lay on the same vector within face space, for some given Euclidean distance. They further argued that this implied support for a norm based coding scheme since "vector angle is important in face perception" [Ross et al., 2010]. It seems to be true that angle matters in face perception, an assertion that could perhaps have been made prior to this thesis, but it is less obvious what neural coding scheme this implies. As previously noted, a coherent metric for norm based coding has yet to be adduced, and as Burton and Volkey have argued, intuitions can be treacherous when contemplating high dimensional spaces. Thus, biologically plausible simulations are required to underpin experimental findings [Burton and Vokey, 1998]. Indeed, recent simulation work from the Poggio lab suggests that the width of tuning curves may be a much more important parameter than the distribution of tuning preferences in terms of explaining the central psychophysical properties of face perception [Tan and Poggio, 2016]. In light of these observations, then, a natural direction in which to pursue further work would be to fit a range of physiologically informed neural models to the data obtained. Such a study would be particularly propitious in light of the fact that, while this data is consonant in certain respects to previous research, it has been acquired through a novel experimental paradigm that enables systematic examination of parameter space. The results of this approach, in particular the finding that the ratio of curvatures changes systematically as a function of eccentricity, go some way to providing the type and degree of constraint required for such an

exercise (biologically plausible simulations) to succeed.

In summary, the results presented in this chapter firstly provide an assessment of how well BFS (Basel Face Space) maps onto human face space. As might have been expected *a priori* the correspondence turns out to be good, but far from perfect. Furthermore the demonstration of anisotropy provided by the analysis in section 5.2.2 (i.e. that absolute geometry plays some role in determining perceptual similarity) suggests that there may be systematic differences in the way that faces are represented in BFS, rather than it, say, being a noisy but unbiased estimator of human judgements.

At bottom then, for researchers contemplating using BFS (Basel Face Space), these results counsel caution in assuming models such as BFS to be accurate approximations to the human representation. However they also demonstrate that the correlation may be sufficiently large that many studies can usefully reply on such a correspondence.

6

Conclusion

6.1 Summary of Theoretical and Experimental Results

This chapter summarises the main theoretical and empirical findings of the preceding chapters. The chief strengths and weaknesses of the results are highlighted and assessed as a whole. Likewise, several

improvements in experimental design regarding the work in chapter 3, now evident in hindsight, are enumerated. And finally, fertile avenues for future investigation are identified.

Following the literature review presented in chapter 1, the original work of this thesis begins, in chaper 2, with a speculative model of how information might be integrated over time in face space. The empirical, mainly neurophysiological, support for such as model was discussed, citing circumstantial but supportive evidence, such as the exponential decay seen in the dynamics of neuronal firing throughout much of the nervous system. Additionally, a normative motivation was drawn from results in statistical time series analysis, in particular exponential smoothing, which is a special case of the well-known Kalman filter. It was shown through simulations that this relatively simple state space model could account for some of the core phenomena observed in the psychophysical literature around face perception, including adaptation, priming, and the skewed reaction time distributions seen in two-alternative forced choice experiments. It was also demonstrated, in subsection 2.6.1, that a readout mechanism could be used in such a state space, based on accumulator models of decision making and accommodating phenomena such as identification and misidentification of an individual.

One of the strengths of a well-defined model, which can be simulated in a mathematical programming language such as Matlab, is that it can be probed using essentially arbitrary inputs. Moreover, results can be obtained, and iterated upon, rapidly. This allows hypotheses to be ex-

plored easily without the arduous, though ultimately desirable, process of deriving analytical, closed-form conclusions where possible. Accordingly, although the experimental prediction chosen for empirical verification/falsification was a relatively simple one (i.e. that a rapidly alternating stimulus would give rise to a linearly interpolated percept, dependent on the duty cycle of a pair of stimuli), the model allows arbitrarily complex inputs to be assessed, where confident intuition surely breaks down.

The empirical work presented in chapter 3 rested on the prediction that the percept of facial identity can be titrated to approximate any interpolated point between two alternating stimuli. It transpired that the empirical results differed considerably in both the radial and the tangential directions, as shown figures 3.15 on page 109 and 3.16 on page 111 respectively. Frequentist statistics were used to confirm the significance of these deviations from the model predictions, leaving the question of what the correct account of the findings was an open one. One option, discussed in section 3.4, was that there was a learning effect, possible because a finite set of pair of faces (10) was used in the experimental design so that subjects became familiar with the stimuli as the experiment proceeded. Relatedly, being conceived and performed only subsequently to the results of the DE, the CE was performed after the DE and it is therefore possible that apparent abolition of the so-called bowing effect was dependent on the temporal order of the experiments (i.e. the amount of learning that had occurred prior to the commencement of each experiment would differ systematically between the DE and the CE).

A conspicuous feature of the deviations from model predictions in the

radial direction is a pronounced centripetal deviation at intermediate duty cycles (30 - 70%) and and equally pronounced centrifugal deviation at more extreme duty cycles (0-10% and 90-100%). This pattern of deviation was dubbed "bowing", and is arguably the most curious feature of the data obtained in the DE, since the centripetal component is absent in the CE while, apparently, all other identified characteristics of the data have been preserved, including the centrifugal deviation at the pattern of tangential deviations. The key difference in the experimental design between the DE and the CE, at least insofar as it was intended, was that in the stimulus was not static in the DE. Indeed, every effort was made to keep the experiments identical except for this particular feature, so as to address the question of whether the dynamic component is responsible for the observed centripetal deviation. Given that the CE also consisted of a titrated merging of stimuli it is hard not to conclude that the dynamic element is the crucial one. Nor could the resulting centripetal effect be considered small, reaching a magnitude of ≈15% (i.e. as a proportion of the distance to the origin; see table 3.3 on page 90) at its maximum, achieved at a duty cycle of 50%. This is approximately 50% greater than the largest centrifugal effect seen at around 10% (see table 3.5 on page 106).

At the time this finding (i.e. centripetal deviation dependent on dynamic stimuli) was made it seemed possible, indeed likely, that this was a Bayesian effect, due to the increased uncertainty implicit in a rapidly changing stimulus. While this does remain plausible, and constitutes the motivation for chapter 4, the possibility that a learning effect could account for this data should have been eliminated prior to investing heavily

in a modelling effort. This would be an important component of future experimental work, were it to be undertaken.

The advantages of hindsight to one side, chapter 4 represents an effort to understand the centrifugal deviation found in the DE within a normative, probabilistic framework. The basic idea is that, whatever dynamical system implements inference within perceptual face, space, the degree of uncertainty, or entropy, in the stimulus varies as a function of the duty cycle. The precise sense in which this is true is somewhat subtle, for example because the duty cycle is strictly speaking deterministic rather than stochastic, but we obtained some insight by modelling the stimulus as a Bernoulli random variable. The entropy of a Bernoulli random variable, varies as a function of the parameter p (see figure 4.1 on page 121), and it was supposed that the entropy of the stimulus varied similarly as a function of the duty cycle. On the basis of this simple assumption the chapter develops a Bayesian inferential model of face perception, the parameters of which were subsequently fitted using the DE data and a MCMC (Markov Chain Monte Carlo) based approach. It was found that the model could replicate certain features of the data, namely the radial centrifugal and centripetal effects. The fact that the model could account for the centrifugal deviations cannot be seen as surprising, since it was achieved by virtue of a simple bias parameter (see description of p_6 in section 4.4 beginning on page 123). However, the centripetal effect arises as a consequence of a static, strong prior and a process of Bayesian inference (see figures 4.8, 4.9 and 4.10 on pages 140, 141 and 142). Notwithstanding its success in this regard, a weakness of the model was that it could not be fitted to the CE

experimental data without modification. The required modification is enforcing a flat prior, which is strictly speaking impossible with a Gaussian distribution and finite covariance matrix. Moreover, using a flat prior is equivalent to performing maximum likelihood fitting, which is precisely the fit performed in chapter 3 (see figure 3.14 on page 108). Thus the model *can* be adapted to account for the CE data, but in a manner of speaking the cure is worse that the disease, since the required adaptation involves eliminating the core of the model, namely the combination of a strong (i.e. informative) and static prior with a variable likelihood (i.e. data of a varying entropy) through a process of Bayesian inference. A further weakness of the Bayesian model presented in chapter 4 is that it has almost no capability to account for the very significant tangential deviations, from linear interpolation, seen in the data from both the CE and the DE. Indeed the maximum magnitude of the tangential deviations in these experiments is approximately twice that ($\approx$30%) of the maximum radial deviation seen in either the DE or the CE ($\approx$15%) (see tables 3.5, pg 106, and 3.3, pg 90, and figures 3.6, pg 94, and 3.16, pg 111). It would seem that some kind of nonlinearity is at play here, and certainly, the pattern and magnitude of tangential deviations must be seen as a further nail in the coffin with regard to the prediction of linear interpolation derived from the exponential smoothing model of chapter 2. Accordingly, a worthy course of future investigation might be to look into the possible causes of this tangential deviation and probe it further experimentally.

Chapter 5, addresses in part one of the key assumptions underlying the experimental approach inherent in both the DE and CE as well as dozens

of other articles present in the face space literature, namely the validity of Basel Face Space (BFS) as an approximation of human face space. This is a crucial assumption to test since many experimental paradigms, including the one adopted in chapter 3, depend upon BFS or similarly conceived models for the validity of their inferential process. To put things the other way, if these models do not correspond, at least to some significant degree, to biological face space, then it is nigh on impossible to draw conclusions based on their use. Encouragingly, it was found that the BFS does indeed display a high correlation with human judgements, with correlation coefficients in the neighbourhood of 0.8-0.85 and highly statistically significant (see tables 5.2, pg 171, and 5.3, pg 171). It was further shown, by performing a subsequent experimental manipulation on a subset of 15 subjects, that even where the experimental stimuli were resampled so as to preserve only the relative and not the absolute geometry the correlations remained high and significant, in the range 0.73-0.76 (see tables 5.2, pg 171, and 5.3, pg 171 , and figures 5.8, pg 170, and 5.9, pg 170). That said, the reduction in correlation seen between sessions wherein stimuli preserved absolute and relative geometry and those in which only relative geometry was preserved was itself highly significant and as shown though an ANOVAs (see tables 5.4, pg 172, and 5.5, pg 172). The conclusion, albeit somewhat qualified, was that BFS is a reasonable but imperfect approximation to human face space and therefore, in appropriate circumstances, an acceptable tool for experimental manipulations. Chapter 5 subsequently addressed an additional question relating to isotropy, namely systematic directionality, and in particular whether it matters in terms of perceptual curvature.

It was noted that previous work, and in particular a paper from Michael Lewis' lab [Ross et al., 2014], argued that directionality is of significance, and in particular that perceptual curvature is greater in the tangential direction than in the radial. Using a different experimental paradigm, and the BFS versus the Stirling face database and *Psychopmorph* [B.Tiddeman and Perrett, 2001], the same partial result was found here. Ross et al. comment that particular "participants were more sensitive to changes made to a face in a direction oblique to the caricature [radial] vector" [Ross et al., 2014]. Notwithstanding the parallels between the results of Ross *et al.* and those presented here, the method deployed in chapter 5 offers additional insight, not evident from previous work. In particular, it utilised a paraboloid fit to estimate the "bowl", so to speak, of perceptual change at varying eccentricities, finding that not only was the tangential curvature of perceptual change consistently greater than in the radial direction, but that the ratio of these two curvatures became markedly greater with increased eccentricity. This is demonstrated in figure 5.13 on page 182, and was seen to be highly statistically significant. An important question is what the most natural and plausible interpretation of this finding is. Here again modelling would be of assistance in exploring the implications of the various coding schemes that have been proposed, and were discussed at some length in section 1.3.2 of chapter 1 (page 17). To summarise chapter 5, it provides, first and foremost, a quantification of the accuracy of BFS as a model of biological face space, using correlation coefficients as the representative metric. The question of whether there are systematic anisotropies is subsequently addressed, in particular between radial and

tangential vectors. And it was finally demonstrated that not only is perceptual change greater, per unit distance in BFS, in the tangential than the radial directions, but that this difference becomes exaggerated with eccentricity.

Overall, this thesis has examined two important aspects of perceptual face space. Firstly, the temporal integration of information, and secondly the utility of a computer-vision based model of face space as a model of biological or perceptual face space. The implications of these findings suggest some interesting avenues for future work, which have also been explored and potential improvements on the work already conducted have been discussed.

Acronyms

BFS Basel Face Space. 3, 18, 37, 152, 154, 163, 166, 168, 172, 173, 182, 183, 185, 191

CE Contrast Experiment. 6, 76, 78, 79, 82, 86, 88, 105, 106, 109, 110, 112, 113, 114, 128, 136, 150, 188, 190, 191

DBS Deep Brain Stimulation. 10

DE Dynamic Experiment. 6, 76, 78, 79, 82, 86, 88, 94, 95, 101, 103, 105, 106, 108, 112, 113, 114, 116, 118, 128, 136, 144, 147, 150, 188, 190, 191

FFDE Flashed Faces Distortion Effect. 75

FRU Face Recognition Unit. 144

GCM Generalised Context Model. 23

IAC Interactive Activation and Competition. 144, 145

MCMC Markov Chain Monte Carlo. 116, 129, 131, 132, 133, 190

MSE Mean Squared Error. 165, 166, 168, 182

PCA Principal Component Analysis. 17, 23

PIU Person Identity Unit. 144

PRU Person Recognition Unit. 144

PSSH Physical Symbol System Hypothesis. 11

SED Situated Embodied Dynamic. 13

Bibliography

[Anderson, 1990] Anderson, J. R. (1990). *The Adaptive Character of Thought (Studies in Cognition)*. Psychology Press, 0 edition.

[Andrieu et al., 2003] Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1-2):5–43.

[Barlow, 1961] Barlow, H. B. (1961). *Possible principles underlying the transformations of sensory messages*, pages 217–234. MIT Press, Cambridge, MA.

[Bartlett et al., 2002] Bartlett, M., Movellan, J., and Sejnowski, T. (2002). Face recognition by independent component analysis.

[Beck et al., 2008] Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–52.

[Behrman and Davey, 2001] Behrman, B. W. and Davey, S. L. (2001). Eyewitness Identification in Actual Criminal Cases: An Archival Analysis. *Law and Human Behavior*, 25:475–491.

[Békésy, 1967] Békésy, G. V. (1967). Sensory inhibition. *Princeton University Press*.

[Berkes P, 2011] Berkes P, Orbán G, L. M. F. J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–7.

[Bill et al., 2015] Bill, J., Buesing, L., Habenschuss, S., Nessler, B., Maass, W., and Legenstein, R. (2015). Distributed Bayesian Computation and Self-Organized Learning in Sheets of Spiking Neurons with Local Lateral Inhibition. *PLoS ONE*, 10(8):e0134356+.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Blakemore et al., 1970] Blakemore, C., Nachmias, J., and Sutton, P. (1970). The perceived spatial frequency shift: evidence for frequency-selective neurones in the human brain. *J. Physiol. (Lond.)*, 210(3):727–750.

[Blanz and Vetter, 1999] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Bolstad, 2010] Bolstad, W. (2010). *Understanding Computational Bayesian Statistics*. Wiley Series in Computational Statistics. Wiley.

[Brainard, 1997] Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436.

[Brooks, 1990] Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1):3 – 15.

[Bruce and Young, 1986] Bruce, V. and Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3):305–327.

[Brunton et al., 2013] Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–8. n/a.

[B.Tiddeman and Perrett, 2001] B.Tiddeman and Perrett, D. (2001). Moving facial image transformations based on static 2d prototypes. *Proc. 9th Int. Conf. In Central Europe on Computer Graphics, Visualization and Computer Vision 2001*.

[Buckingham, 2006] Buckingham, G., D. L. M. L. A. C. W. L. L. M. C. C. A. T. B. P. J. B. (2006). Visual adaptation to masculine and feminine faces influences generalized preferences and perceptions of trustworthiness. *Evolution and Human Behavior*, 27:381–389.

[Burton et al., 1990] Burton, A. M., Bruce, V., and Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *Br J Psychol*, 81 (Pt 3):361–380.

[Burton and Vokey, 1998] Burton, A. M. and Vokey, J. R. (1998). The face-space typicality paradox: Understanding the face-space metaphor. *The Quarterly Journal of Experimental Psychology Section A*, 51(3):475–483.

[Buzsaki, 2006] Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

[Callier and Desoer, 1991] Callier, F. and Desoer, C. (1991). *Linear System Theory*. Springer Texts in Electrical Engineering. Springer-Verlag.

[Cardwell, 1971] Cardwell, D. (1971). *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age*. Heinemann educational books. Heinemann Educational Books.

[Chang, 2017] Chang, Le. Tsao, D. (2017). The code for facial identity in the primate brain. *Cell*, 169:1013–1028.

[Chatfield, 2003] Chatfield, C. (2003). *The Analysis of Time Series: An Introduction, Sixth Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 6 edition.

[Clifford et al., 2007] Clifford, C. W., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., and Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, 47(25):3125 – 3131.

[Deutsch, 2012] Deutsch, D. (2012). Creative blocks: The very laws of physics imply that artificialintelligence must be possible.what's holding us up? *Aeon*.

[Dreyfus, 1992] Dreyfus, H. L. (1992). *What Computers Still Can'T Do: A Critique of Artificial Reason*. MIT Press, Cambridge, MA, USA.

[Dyster et al., 2016] Dyster, T. G., Mikell, C. B., and Sheth, S. A. (2016). The co-evolution of neuroimaging and psychiatric neurosurgery. *Frontiers in Neuroanatomy*, 10:68.

[ELLIS, 1975] ELLIS, H. D. (1975). Recognizing faces. *British Journal of Psychology*, 66(4):409–426.

[Fetsch et al., 2014] Fetsch, C. R., Kiani, R., and Shadlen, M. N. (2014). Predicting the Accuracy of a Decision: A Neural Mechanism of Confidence. *Cold Spring Harb. Symp. Quant. Biol.*, 79:185–197.

[Fiser, 2010] Fiser, J, B. P. O. G. L. M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.*, 14(3):119–30.

[Fishel and Loeb, 2012] Fishel, J. and Loeb, G. (2012). Bayesian exploration for intelligent identification of textures. *Frontiers in Neurorobotics*, 6:4.

[Fries, 2015] Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, 88(1):220–235.

[Gallistel and King, 2009] Gallistel, C. R. and King, A. P. (2009). *Memory and the computational brain : why cognitive science will transform neuroscience*. Wiley-Blackwell.

[Girshick et al., 2011] Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci*, 14(7):926–32.

[Gold et al., 2010] Gold, J., Law, C.-T., Connolly, P., and Bennur, S. (2010). Relationships between the threshold and slope of psychometric and neurometric functions during perceptual learning: implications for neuronal pooling. *Journal of neurophysiology*, 103(1):140–154.

[Goldstein and Chance, 1980] Goldstein, A. G. and Chance, J. E. (1980). Memory for faces and schema theory. *The journal of psychology*, 105(1):47–59.

[Gruter et al., 2008] Gruter, T., Gruter, M., and Carbon, C. C. (2008). Neural and genetic foundations of face recognition and prosopagnosia. *J Neuropsychol*, 2(Pt 1):79–97.

[Herzmann et al., 2010] Herzmann, G., Kunina, O., Sommer, W., and Wilhelm, O. (2010). Individual differences in face cognition: brain-behavior relationships. *Journal of cognitive neuroscience*, 22(3):571–589.

[Houlsby et al., 2013] Houlsby, N. M. T., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., and Lengyel, M. (2013). Cognitive Tomography Reveals Complex, Task-Independent Mental Representations. *Curr Biol*, 23(21):2169–2175.

[Isaacson and Scanziani, 2011] Isaacson, J. S. and Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron*, 72(2):231–243.

[Jacobson, 1993] Jacobson, M. (1993). Foundations of neuroscience. *Springer*, (2nd ed.):277.

[Jaynes, 1988] Jaynes, E. T. (1988). How Does the Brain Do Plausible Reasoning? *Maximum-Entropy and Bayesian Methods in Science and Engineering*.

[Jordan and Bishop, 2014] Jordan, M. I. and Bishop, C. M. (2014). Neural networks. In Gonzalez, T. F., Diaz-Herrera, J., and Tucker, A., editors, *Computing Handbook, Third Edition: Computer Science and Software Engineering*, pages 42: 1–24. CRC Press.

[Khaligh-Razavi and Kriegeskorte, 2014] Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 10(11):1–29.

[Knill and Pouget, 2004] Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*, 27(12):712–9.

[Körding and Wolpert, 2004] Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–7.

[Learned-Miller et al., 2006] Learned-Miller, E., Lu, Q., Paisley, A., Trainer, P., Blanz, V., Dedden, K., and Miller, R. (2006). Detecting acromegaly: screening for disease with a morphable model. *Med Image Comput Comput Assist Interv*, 9(Pt 2):495–503.

[Lebedev and Nicolelis, 2006] Lebedev, M. A. and Nicolelis, M. A. L. (2006). Brain-machine interfaces: past, present and future. *Trends Neurosci.*, 29(9):536–46.

[Lee et al., 2000] Lee, K., Byatt, G., and Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: testing the face-space framework. *Psychological Science*, 11(5):379–385.

[Leibo et al., 2017] Leibo, J. Z., Liao, Q., Anselmi, F., Freiwald, W. A., and Poggio, T. (2017). View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation. *Current Biology*, 27(1):62 – 67.

[Leopold et al., 2006] Leopold, D. A., Bondar, I. V., and Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–575.

[Leopold et al., 2001] Leopold, D. A., O'Toole, A. J., Vetter, T., and Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature neuroscience*, 4(1):89–94.

[Lewis, 2004] Lewis, M. (2004). Face-space-r: Towards a unified account of face recognition. *Visual Cognition*, 11(1):29–69.

[Lewis and Ellis, 2000] Lewis, M. B. and Ellis, H. D. (2000). Satiation in name and face recognition. *Memory & Cognition*, 28(5):783–788.

[Light et al., 1979] Light, L. L., Kayra-Stuart, F., and Hollander, S. (1979). Recognition memory for typical and unusual faces. *J Exp Psychol Hum Learn*, 5(3):212–228.

[Ma et al., 2006] Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci*, 9(11):1432–8.

[Marr, 1983] Marr, D. (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt & Company.

[Mitchell, 1995] Mitchell, M. (1995). A complex-systems perspective on the computation vs. dynamics debate in cognitive science. *Proceedings of the 20th Annual Conference of the Cognitive Science Society: Cogsci98*.

[Morton et al., 1979] Morton, J., Smith, N., and Marshall, J. (1979). *Structures and processes*. Psycholinguistics series. Paul Elek.

[Murphy, 2012] Murphy, K. P. (2012). Machine learning a probabilistic perspective.

[Newell and Simon, 1976] Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19(3):113–126.

[Nosofsky, 2011] Nosofsky, R. (2011). The generalized context model: An exemplar model of classification. pages 18–39.

[O'Keeffe, 2011] O'Keeffe, J. (2011). How will neuroscience influence the practice of psychiatry in the next ten years? *Opticon 1826*, (10):1–9.

[Or and Eckstein., 2015] Or, Charles C.-F., M. F. P. and Eckstein., M. P. (2015). "initial eye movements during face identification are optimal and similar across cultures.". *Journal of Vision*, 15(13).

[Palmer, 2005] Palmer, J, H.-A. S. M. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J Vis*, 5(5):376–404.

[Passino et al., 2008] Passino, K., Seeley, T., and Visscher (2008). Swarm cognition in honey bees. 62(3):401–414.

[Paysan et al., 2009] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3D Face Model for Pose and Illumination Invariant Face Recognition. In Tubaro, S. and Dugelay, J.-L., editors, *AVSS*, pages 296–301. IEEE Computer Society.

[Pfeifer and Bongard, 2006] Pfeifer, R. and Bongard, J. C. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence (Bradford Books)*. The MIT Press.

[Popper, 2002] Popper, K. (2002). *The Logic of Scientific Discovery*. Routledge Classics. Taylor & Francis.

[Psalta et al., 2014] Psalta, L., Young, A. W., Thompson, P., and Andrews, T. J. (2014). Orientation-sensitivity to facial features explains the Thatcher illusion. *Journal of vision*, 14(12).

[Ratcliff and McKoon, 2008] Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*, 20(4):873–922.

[Ratcliff and Smith, 2004] Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, 111(2):333–367.

[Rhodes G., 2003] Rhodes G., Jeffery L., W. T. L. C. C. W. G. N. K. (2003). Fitting the mind to the world: Face adaptation and attractiveness after effects. *Psychological Science*, 14:558–566.

[Rice, 1995] Rice, J. A. (1995). *Mathematical statistics and data analysis*. Duxbury Press, 2 edition.

[Richler and Gauthier, 2014] Richler, J. J. and Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychol Bull*, 140(5):1281–1302.

[Rips, 1973] Rips, L . J. Shoben, E. J. S. E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behovior*, 12:1–20.

[Rolls and Treves, 1998] Rolls, E. and Treves, A. (1998). Neural networks and brain function. *Oxford University Press: Oxford*.

[Romo, 2012] Romo, D. R. A. (2012). Brain mechanisms for perceptual and reward-related decision-making. *Progress in Neurobiology*.

[Ross et al., 2014] Ross, D. A., Deroche, M., and Palmeri, T. J. (2014). Not just the norm: exemplar-based models also predict face aftereffects. *Psychon Bull Rev*, 21(1):47–70.

[Ross et al., 2010] Ross, D. A., Hancock, P. J. B., and Lewis, M. B. (2010). Changing faces: Direction is important. *Visual Cognition*, 18(1):67–81.

[S., 1987] S., R. G. B. S. C. (1987). Identifications and ratings of caricatures: Implications for mental representations. *Cognitive Psychology*, 19:473–497.

[Seeley et al., 2012] Seeley, T., Visscher, K., Schlegel, T., Hogan, P., Franks, N., and Marshall, J. (2012). Stop signals provide cross inhibition in collective decision-making by honeybee swarms. *Science (New York, N.Y.)*, 335(6064):108–111.

[Shadlen et al., 2008] Shadlen, M. N., Kiani, R., Hanks, T. D., and Churchland, A. K. (2008). *Neurobiology of decision making: An intentional framework*, pages 71–101. The MIT Press, Cambridge.

[Simo, 2013] Simo, S. (2013). *Bayesian filtering and smoothing*. Cambridge University Press.

[Sirovich and Meytlis, 2009] Sirovich, L. and Meytlis, M. (2009). Symmetry, probability, and recognition in face space. *Proceedings of the National Academy of Sciences*, 106(17):6895–6899.

[Tan and Poggio, 2016] Tan, C. and Poggio, T. (2016). Neural Tuning Size in a Model of Primate Visual Processing Accounts for Three Key Markers of Holistic Face Processing. *PLoS ONE*, 11(3):e0150980.

[Tangen et al., 2011] Tangen, J. M., Murphy, S. C., and Thompson, M. B. (2011). Flashed face distortion effect: Grotesque faces from relative spaces. *Perception*, 40(5):628–630.

[Thompson, 1980] Thompson, P. G. (1980). Margaret thatcher : a new illusion. *Perception*, pages 483–484.

[Torres-Sanchez et al., 2017] Torres-Sanchez, S., Perez-Caballero, L., and Berrocoso, E. (2017). Cellular and molecular mechanisms triggered by Deep Brain Stimulation in depression: A preclinical and clinical approach. *Prog. Neuropsychopharmacol. Biol. Psychiatry*, 73:1–10.

[Turk and Pentland, 1991] Turk, M. A. and Pentland, A. P. (1991). Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition*, pages 586–591. IEEE.

[Usher and McClelland, 2001] Usher, M. and McClelland, J. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550–592.

[Valentine, 1991a] Valentine, T., F. A. (1991a). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology*, 82:87–102.

[Valentine, 1991b] Valentine, T. (1991b). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A*, 43(2):161–204.

[Valentine and Bruce, 1986] Valentine, T. and Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, 15(5):525–535.

[Valentine and Endo, 1992] Valentine, T. and Endo, M. (1992). Towards an exemplar model of face processing: the effects of race and distinctiveness. *Q J Exp Psychol A*, 44(4):671–703.

[Valentine et al., 2016] Valentine, T., Lewis, M. B., and Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, 69(10):1996–2019. PMID: 25427883.

[Van Kampen, 2007] Van Kampen, N. G. (2007). *Stochastic Processes in Physics and Chemistry, Third Edition (North-Holland Personal Library)*. North Holland, 3 edition.

[van Loon et al., 2013] van Loon, A. M., Knapen, T., Scholte, H. S., St John-Saaltink, E., Donner, T. H., and Lamme, V. A. (2013). GABA shapes the dynamics of bistable perception. *Current biology : CB*, 23(9):823–827.

[Walther et al., 2013] Walther, C., Schweinberger, S. R., Kaiser, D., and Kovacs, G. (2013). Neural correlates of priming and adaptation in familiar face perception. *Cortex*, 49(7):1963–1977.

[Wang and Lai, 2011] Wang, S. F. and Lai, S. H. (2011). Reconstructing 3D Face Model with Associated Expression Deformation from a Single Face Image via Constructing a Low-Dimensional Expression Deformation Manifold. *IEEE Trans Pattern Anal Mach Intell*, 33(10):2115–2121.

[Wenger, 2001] Wenger, M. J., . T. J. T. E. . (2001). Computational, geometric, and process issues in facial cognition: Progress and challenges.

[Westheimer, 2008] Westheimer, G. (2008). Was Helmholtz a Bayesian? *Perception*, 37(5):642–650.

[Wilson, 2001] Wilson, H. (2001). Vision, low-level theory of. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social and Behavioral Sciences*, pages 16232 – 16237. Pergamon, Oxford.

[Young et al., 1985] Young, A. W., Hay, D. C., and Ellis, A. W. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology*, 76(4):495–523.

[Yu and Cohen, 2009] Yu, A. J. and Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1873–1880. Curran Associates, Inc.