UNIVERSITY OF CAMBRIDGE

(Girton College)

Shadows of the Mind:

Using Discrete Decision Tasks to Extract Mental Representations.

This dissertation is submitted for the degree of Master of Philosophy in Engineering.

Author:
Mohammad M. Ghassemi

Advisor:
Dr. Máté Lengyel

Supervisor:
Professor Daniel Wolpert

Table of Contents

Preface

I am submitting this dissertation to fulfil the requirements for an M.Phil in Engineering from the University of Cambridge. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. My central contribution was on the experimental procedure, collection of data, and analysis of results. The statistical methods employed in this thesis were developed by others, as mentioned in the main body of the text.

1. Introduction

It is a remarkable feat of nature that organisms have an ability to translate a mass of noisy, often conflicting sensory data, into abstract concepts, ideas and categories that impact the way they think and behave. Historically, there have been multiple theories that explain how this phenomenon takes place. If we choose to measure the worth of a theory by the discoveries it inspires, however, then the computational theory of mind is surely the champion among them. In contrast to the behaviourist view, which opposed "speculation" on the inner contents of the mind that influence behaviour, the computational theory of mind embraces the idea that humans learn, experience and interact with the world using internal mental representations of reality (Craik, 1967). These mental representations reflect all an individual subjectively knows about the world. Stephen Pinker provided an intuitive way of understanding the concept of mental representations in his book, *How The Mind Works*:

> *"Plato said that we are trapped inside a cave and know the world only through the shadows it casts on the wall. The skull is our cave, and mental representations are the shadows." (Pinker, 1999)*

If these "shadows of the mind", or mental representations, are indeed responsible for our behaviour, then understanding a mind is synonymous with understanding its corresponding mental representations and the dynamics between them. Cognitive psychologists made the first steps in this direction by designing experiments that require subjects to use a specific facet of a mental representation to perform a task. Then, using metrics such as reaction time, they inferred something about the underlying representation of the subject (Laming, 1968; Rhodes et al., 1987). Bayesian models of cognition take this primitive approach one step further by defining a mental representation as a probability distribution over a set of stimuli. This makes the process of extracting a mental representation

synonymous with describing its corresponding probability distribution (Dayan and Abbott, 2001).

While language is undoubtedly a window into the mind, using language to describe such a distribution is far from intuitive (Ericsson and Simon, 1980). To illustrate this point, imagine you are given the task of describing your mental representation of attractive faces. The distribution spans a high dimensional "face-space" and is probably multi-modal. You could attempt to describe the mean, but how easily could you describe other aspects of the distribution like modality or covariance?

The Bayesian approach to this problem is to collect a data set that is dependent on the underlying distribution of interest (attractive faces). Then, with a sensible model of the human cognitive process during the task and a reasonable *a priori* distribution describing beliefs about attractive faces, we can infer the underlying representation given the observed data using Bayes' Rule. We emphasize here that even an infinite amount of data is meaningless, however, without a statistical model that appropriately describes the subject's behaviour during the task. This fact provides motivation to utilize experimental tasks which maximize information content per trial, while allowing for accurate modelling of subject behaviour during the task.

Discrete decision tasks fulfill this requirement. A discrete decision task consists of several experimental trials in which subjects are presented with sets of stimuli, S, and make a decision about the stimuli based on the instructions of the task. The first question of our study is to identify if it is possible to model complex mental representations, using a trinary, Odd-One-Out (OOO), decision task paradigm in which subjects are presented with a set of three stimuli, for N trials, and asked to choose the "odd-ball" stimuli.

As we hope our previous example with attractive faces alluded to, humans have visual mental representations of faces which allow them to distinguish between minute differences

and develop individualized representations over abstract concepts such as attractiveness or familiarity (Beale and Keil, 1995). Hence we modelled the mental representation of faces in this study in order to gauge our ability to extract fundamentally subjective, complex, multi-modal mental representations in a more general sense. Our ability to perform Bayesian inference in face-space however, will also require us to compute the probability of observed data given every possible subjective distribution in the mind of our subject (from apples to zebras and beyond). It should come as no surprise that such a quantity is analytically intractable.

It is possible, however, to determine the properties of the probability distribution within face-space by drawing sample distributions from it. Generating such samples is not simple, especially in cases like ours when *a priori* information on the distribution is limited or unknown. Markov chain Monte Carlo (MCMC) algorithms are particularly useful for this kind of problem as they can begin from any location in a parameter space and (given enough steps) converge on the underlying target distribution. As we are estimating a face-space distribution, each sample produced by the MCMC describes such a distribution fully. There are many flavours of MCMC sampling methods that are available, of which we chose the Hybrid Monte Carlo (HMC) approach (Duane, 1987). HMC is attractive because it avoids the random walk behaviour associated with the canonical Metropolis Monte Carlo. This increases the independence of samples and the speed by which the chain can identify the target distribution.

While MCMC based inference allows us to fit the parameters of a statistical model that describes a subject's behaviour to the observed data, verification of the model's accuracy is another issue altogether. The simplest way to gauge the accuracy is to expose the trained model to novel task stimuli and compare responses to those of the actual subject. Gauging the accuracy of our method based solely on within-task predictions can be problematic as we

have no way of knowing if the extracted distribution depicts the actual mental representation we are looking for (faces), some other mental representation, or no distribution at all (structured noise). Thus, it is necessary to perform a control task, which is dependent on the same underlying representation, but assumes a distinct generative model. The control paradigm we opted for was the Two Alternative Forced Choice task (2AFC), an established and well tested psychometric experiment in which subjects are asked to choose the more familiar of two stimuli (LaBerge, 1962; Laming, 1968; Gold and Shadlen, 2002). We can extract the underlying distribution according to data from the 2AFC task and perform cross-validation to gauge the effectiveness of the OOO task and, more importantly, the integrity of the extracted distributions.

Our approach builds on the work of others that explain human behaviour using Bayesian models (Chater and Manning, 2006; Körding and Wolpert, 2006; Steyvers et al., 2006; Tenenbaum et al., 2006). We posit that a subjective distribution can be inferred based on task performance and then gauged using a likelihood value from Bayesian analysis. The ability to gauge the effectiveness of our model, in conjunction with algorithms that adjust our model's parameters (Monte Carlo methods) should allow us to optimize our model, given the underlying assumptions.

2. Methods

Our aim in this section is to outline the method used for inferring the mental representations of subjects based on their responses to our discrete decision tasks. To begin, we need an accurate statistical model that explains how subjects generate responses to a task given an underlying mental representation of interest, $\mathcal{P}$. We construct our statistical model building on the classical "ideal observer" model which states that an ideal subject should behave in a way that minimizes error given behavioural constraints (decision noise for instance) and the information available to them in each trial of the task (Geisler, 2003).

Although it has been used with some success in the past, the ideal observer model is flawed in that it does not account for the variety of sub-optimal behaviours that are typically exhibited by subjects during experimental studies. Hence, we follow the lead of Huszár et al. (2010) and modulate this classical definition to include several sources of non-ideal behaviour and a dependence of the responses on the mental representation. We believe this approach provides a more realistic model of subject behaviour during the task and will allow us to more accurately parameterize and predict our subject's behaviour. Once developed, the statistical model of subject behaviour can be used to infer a mental representation of interest. Figure 1(A) graphically illustrates such a model in the most general sense.

Before continuing we must acknowledge Ferenc Huszár, Neil Houlsby and Dr. Máté Lengyel for their foundational work in this area. The models, approach, and procedures used in this section were derived from their prior and current work in the area.

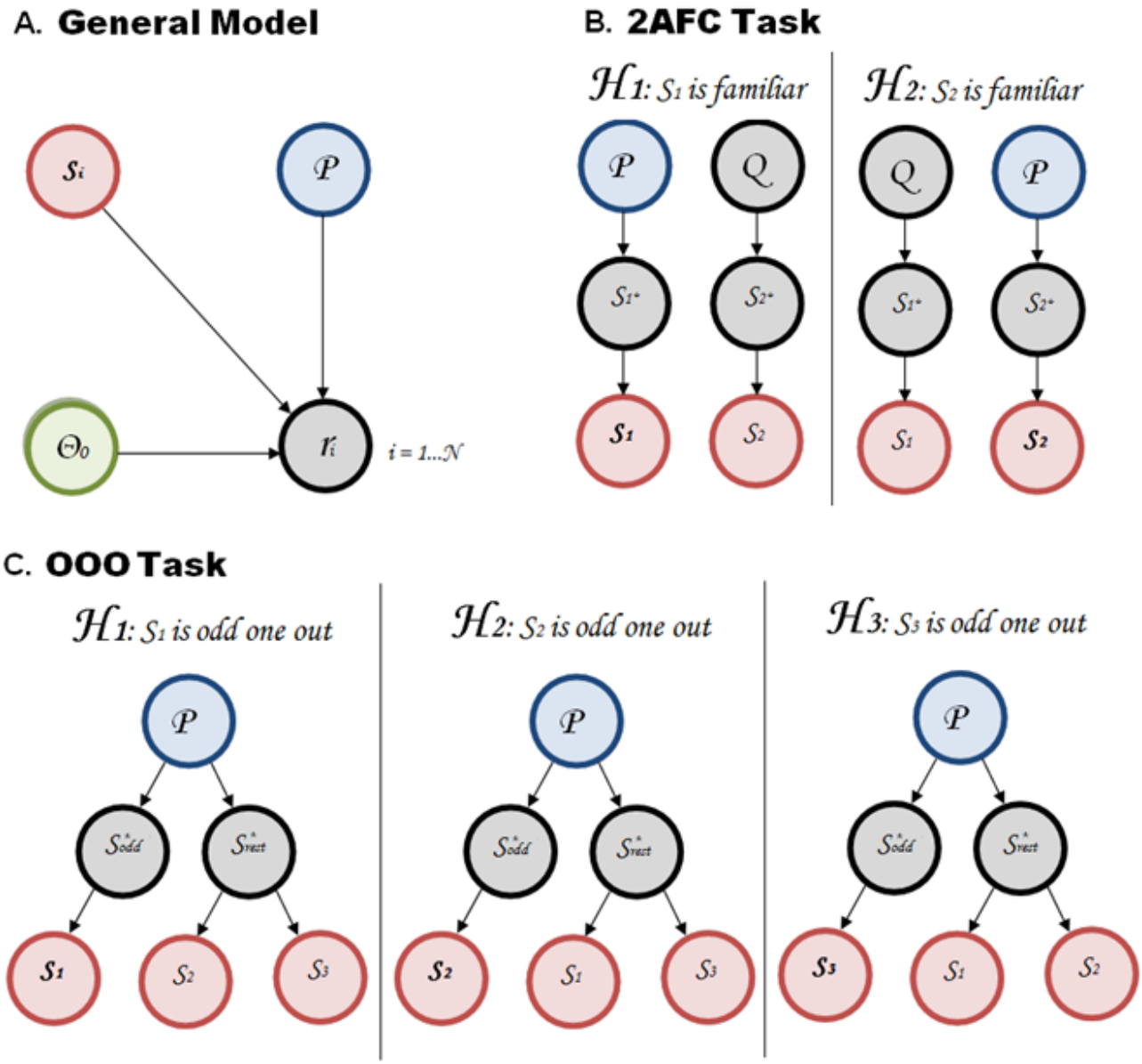


Figure 1. Graphical models illustrating the outcomes of various hypotheses for each task. (A) In general, subjects generate responses r_i given a set of stimuli S_i , an underlying distribution $\mathcal{P}$, and behavioural parameters θ_0. (B): In the 2AFC experiment, subjects were presented with two stimuli. Hypothesis H_i corresponds to the belief that stimulus s_i is more familiar. In the model, s_i is a noisy version of the underlying stimulus, s_i^* drawn from $\mathcal{P}$ while the other observed stimulus, s_j, is noisy version of an underlying stimuli drawn from another distribution Q. (C): In the OOO experiment subjects were presented with three stimuli. Hypothesis H_j corresponds to the belief that stimulus s_j is the "odd-one-out", or noisy version of the underlying stimulus S^*_{odd}, while the other observed stimuli were noisy versions of a separate underlying stimulus, S^*_{rest}.

2.1 General Model

Our general model for a discrete decision task (Figure 1a) assumes that subjects have an underlying distribution $\mathcal{P}$ over all possible stimuli which drives their decision making process. We aim to estimate $\mathcal{P}$ and the set of parameters θ_0 that describe the non-ideal aspects of our subject's behaviour during the task. We accomplish this using a series of N experimental trials in which subjects generate a response r_i, for each trial i, depending on the nature of presented stimuli S_i. During the task, we assume that subjects have hypotheses, H_i, that relate to each of their potential responses r_i with a one-to-one mapping between hypotheses and response types. Subjects gauge the likelihood of each hypothesis, for each trial, i, by computing $p(S_i|H_j, \mathcal{P}, \theta_0)$. Inference can then be performed according to the generative model of the task to extract a posterior probability for each hypothesis.

$$\mathbb{P}[\boldsymbol{H_j}|\boldsymbol{S_i}, \boldsymbol{\mathcal{P}}] = \frac{\boldsymbol{p}(\boldsymbol{S_i}|\boldsymbol{H_j}, \boldsymbol{\mathcal{P}})}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{R}} \boldsymbol{p}(\boldsymbol{S_i}|\boldsymbol{H_k}, \boldsymbol{\mathcal{P}})} \quad (1)$$

The R term in Equation 1 refers to the number of response options in the task (3 for the OOO task and 2 for the 2AFC task). With the ability to gauge each hypothesis, subjects would ideally generate responses by choosing the hypothesis with the highest corresponding posterior probability. We point out here that the likelihood term in Equation 1 is always dependent on the subjective distribution $\mathcal{P}$ regardless of the task. Hence, we should be able to infer the same subjective distribution across multiple tasks.

As we mentioned earlier, we do not assume that subject's behave ideally during response generation. Hence, we introduce the first of our non-optimal behavioural parameters in Equation 2, π, the prior bias parameter.

$$\mathbb{P}\left[H_j|S_i,\mathcal{P},\pi\right] = \frac{p(S_i|H_j,\mathcal{P})\cdot\pi_j}{\sum_{k=1}^{R} p(S_i|H_k,\mathcal{P})\cdot\pi_k} \quad (2)$$

Where π_j measures how biased a subject is towards preferring H_j, independent of the likelihood assigned to it. In other words, the parameter describes a subject's prior preference over all hypotheses where $\sum_k \pi_k = 1$.

The prior preference term in Equation 2 is not the only source of noise we anticipate observing from our subjects. The next among these is a parameter which models a subject's internal representation of observation noise as a constant covariance Gaussian distribution:

$$p_n(s|s^*) = \mathcal{N}(s; s^*, \Sigma_N) \quad (3)$$

Equation 3 is a simple, but effective way to describe how the noise in a subject's perceptual process, controlled by Σ_N, would distort the true stimuli s^* into the stimulus actually experienced by the subject, s. The next behavioural parameter we introduce models the inverse attention noise level, κ, of the subject. This parameter is intended to quantify attention related misbehaviour exhibited by the subject during the task (e.g. lack of task understanding, boredom, etc.). Our model assumes that the average subject's behaviour lies somewhere between a completely ideal subject, which would generate responses based solely on the available data, and a completely non-ideal subject, which would generate responses randomly. When incorporated into our model, the κ parameter helps describe the mixture between these two types of subjects:

$$\mathbb{P}[r_i = j|S_i,\mathcal{P},\kappa] = \frac{\kappa}{R} + (1-\kappa)\mathbb{P}[r_i = j|S_i,\mathcal{P}] \quad (4)$$

Lastly, we include a parameter which models the discrepancy between the probability assigned to a hypothesis by the posterior, and the corresponding probabilities that a subject

actually chooses that hypothesis. In other words, we quantify how well subjects are generating the response that corresponds to the hypothesis with the highest posterior probability. This decision noise parameter, β, can be introduced by using the Softmax rule in place of the maximum operation when calculating the probability of a subject's response:

$$\mathbb{P}[r_i = j|S_i, \mathcal{P}, \beta] = \mathbb{P}_{softmax}[j] = \frac{\mathbb{P}[H_j|S_i,\mathcal{P}]^{\beta}}{\sum_{k=1}^{R}\mathbb{P}[H_k|S_i,\mathcal{P}]^{\beta}} \quad (5)$$

Having discussed all the parameters in θ_0, the overall response probabilities of our non-ideal subjects are modelled by:

$$\mathbb{P}[r_i = j|S_i, \mathcal{P}, \theta_0] = \frac{\kappa}{R} + (1-\kappa)\frac{\mathbb{P}[H_j|S_i,\mathcal{P}]^{\beta}}{\sum_{k=1}^{R}\mathbb{P}[H_k|S_i,\mathcal{P}]^{\beta}} \quad (6)$$

Where, once again, *R* indicates the number of discrete choices available in the task. With this model, we can compute a subject's response probability, $p(r_i|S_i, \mathcal{P}, \theta_0)$ given novel stimulus sets. A key assumption of our model is that $\mathcal{P}$ will remain constant over the course of the experiment and that the responses during experimental trials are conditionally independent. Under these assumptions, we would expect the following likelihood for our latent parameters:

$$p(r_{1:N}|s_{1:N}, \mathcal{P}, \theta_0) = \prod_{i=1}^{N} p(r_i\,|s_i, \mathcal{P}, \theta_0) \quad (7)$$

Applying Bayes' rule, we can infer the posterior over $\mathcal{P}$ *and* θ_0 given a feasible prior over the distribution and parameters:

$$p(P, \theta_0|r_{1:N}, s_{1:N}) = \frac{p(\mathcal{P})p(\theta_0)\prod_{i=1}^{N} p(r_i\,|s_i, \mathcal{P}, \theta_0)}{\prod_{i=1}^{N} p(r_i\,|s_i)} \quad (8)$$

Where $p(\mathcal{P})$ is the prior distribution over $\mathcal{P}$, $p(\theta_0)$ is the prior distribution over behavioural parameters θ_0, $p(r_i\,|s_i, \mathcal{P}, \theta_0)$ is the likelihood and $\prod_{i=1}^{N} p(r_i\,|s_i)$ is the marginal likelihood.

As the marginal likelihood term in Equation 8 makes calculation of this posterior analytically intractable, we implemented a Hamiltonian Monte Carlo algorithm to perform approximate inference. This approach will allow us to converge on the set of parameters for our model that best approximate $\mathcal{P}$ and θ_0 .

2.2 Mixture of Gaussians Model

To estimate a posterior for an underlying distribution we must first make certain assumptions about the nature of the distribution. For this study, we assumed $\mathcal{P}$ to be a mixture of Gaussians (MoG):

$$\mathcal{P}(S) = \sum_m \omega_m \mathcal{N}(s; \mu_m, \Sigma_m) \quad (9)$$

Where m indexes the number of mixture components, μ describes the means of the m components, ω is a 1 x m vector containing weights of the Gaussian components and $\sum_m$ contains the covariance matrix for each component.

We use the MoG model in its simplest form, setting a static number of Gaussians that define the subjective distribution up-front. While theoretically even the most complex distributions can be approximated with a MoG model, there are practical limitations to consider. In our case, the number of model parameters grows quadratically with respect to stimulus dimensions and linearly with respect to m. It is worth noting at this point that a MoG model might not be the optimal way to model the underlying distributions; however, we prefer it due to its simplicity and inherent flexibility.

2.3 Likelihood for the 2AFC Task

The 2AFC task presents subjects with two stimuli and instructs them to choose the one which is more familiar (see section 3 for more details). Here we will discuss the form that the

likelihood function defined in Equation 7 takes on for the 2AFC task. Please refer to Figure 1b for a graphical illustration of our 2AFC generative model.

In our model, s_i is a noisy version of the actual stimulus, s_i^* drawn from the familiar distribution $\mathcal{P}$, while the other observed stimulus s_j is a noisy version of a stimulus s_j^* drawn from a uniform distribution of non-familiar stimuli Q. This convenient assumption of uniformity makes the influence of Q negligible when comparing hypotheses and has been used in other studies (Sanborn et al., 2010; Orbán et al., 2008). Hence, the likelihood for a 2AFC task hypothesis can be stated as:

$$p(S_x|H_i,\mathcal{P}) = \int p_n(s_i|s_i^*)\mathcal{P}(s_i^*)\,ds_i^* \cdot \int p_n\left(s_j\middle|s_j^*\right) Q\,(s_j^*)ds_j^* \tag{10}$$

The likelihood for all hypotheses in the 2AFC task can be accurately computed using Equation 10 so long as our assumptions about Q hold true. This will depend largely on the subject's understanding of the task.

2.4 Likelihood for the OOO Task

The OOO task presents subjects with three stimuli and instructs them to choose the odd one out (see section 3 for more details). Here we will discuss the form that the likelihood function defined in Equation 7 takes on for the OOO task. Please refer to Figure 1c for a graphical illustration of our OOO generative model.

Our model for the task assumes that the subjective similarity of two things is gauged by evaluating the probability that both were produced using an identical generative process (Kemp et al., 2005). Therefore, the subject generates responses by identifying the stimulus pair that maximize this probability and then choosing the stimuli which is not part of that pair. It follows that the likelihood for an OOO task hypothesis can be stated as:

$$p(S_x|H_i,\mathcal{P}) = \int p_n(s_i|s_{odd})\,\mathcal{P}(s_{odd})\,ds_{odd} \cdot \int\left[\prod_{j\in\{a,b\}} p_n\left(s_j\middle|s_{rest}\right)\right] ds_{rest} \quad (11)$$

where $\{a,b\} = \{1,2,3\}/i$, and $\mathcal{P}(S)$ is defined as in Equation 9, and $p_n(s|s^*)$ is defined in Equation 3. The likelihood for all hypotheses in OOO task can be computed using Equation 11.

2.5 Prior Distributions

Having defined feasible likelihood functions for our tasks, the last quantities we need to define to perform Baysian inference are the prior distributions over our mixture of Gaussian model, and behavioural parameters. More specifically, this requires us to set priors over the latent variables in $\mathcal{P}$ and θ_0. We would like to indicate here that all stimuli were normalized between 0 and 1 before running the HMC. Hence, our priors are selected to work with the normalized versions of the data.

We already described θ_0 earlier in this section as including κ, β, π, and Σ_n. For θ_0, the prior over decision noise, β, was Wishart distributed with mean 1 and concentration 5 for both tasks. The inverse attention, κ, was Beta distributed with ∝=1, β=10 for both tasks. Observation noise, Σ_n, was Wishart distributed with means 0.02**I**, and concentrations 20 and 10 for the 2AFC and OOO tasks respectively. The prior bias , π, was drawn from a beta distribution with ∝=2, β=2 for the 2AFC task and a continuous von Mises distribution (also known as circular normal) with mean 0, and concentration of 1 for the OOO task.

We will now define the parameters which describe the prior distribution over $\mathcal{P}$ as θ_p. Following from the MoG assumption in Equation 9, θ_p will include:

- *K*: The total number of mixture components
- μ_m: The means of each component

- Σ_m: The covariance matrix of each component

For θ_p, we set the number of mixture components, *K*, equal to 2 for both tasks. The prior over means was Gaussian distributed with a mean of 0.5 for both tasks and variance of 0.05 for both tasks. We specified covariance values for mixture components using a Wishart prior with means 0.01**I** with concentration values of 4 for both tasks. As we anticipated that the posterior distribution over θ_p parameters would vary widely on an individual basis, it was difficult to select meaningful priors over these parameters.

2.6 Hamiltonian Monte Carlo

As we mentioned before, the posterior defined in Equation 8 is analytically intractable. Thus, we use the Hamiltonian Monte Carlo algorithm for approximating the posterior distribution by randomly generating parameter values that allow our model to converge on the true distribution. The efficiency of HMC is its primary attraction, requiring fewer (though more computationally expensive) steps to produce statistically independent samples than other MCMC methods. HMC is able to do this because it uses information from both the distribution as well as the gradient of the log probability distribution when making steps.

Hamiltonian Monte Carlo draws its name from classical mechanics, where a Hamiltonian is the sum of potential and kinetic energies of a system. For each model parameter θ_q HMC introduces a "momentum" parameter γ_q (Neal, 1996) and defines a Hamiltonian:

$$H = \varphi(\theta) + \lambda(\gamma_q) = -\log(\mathcal{P}(\theta)) + \sum \frac{\gamma_q^2}{2M_q} \quad (12)$$

In the above equation, φ is potential energy, λ is kinetic energy and M_q is an assumed mass. This Hamiltonian can be used to draw samples that are proportional to e^{-H}, and thereby sensitive to the gradients of the of the log probability distribution. This process begins with Gibbs sampling γ_q from the uncorrelated Gaussian, λ. Then, using what is known as the "leapfrog technique", it moves along some trajectory which keeps *H* constant, for a time *T*:

$$\gamma_q\left(t+\frac{\tau}{2}\right) = \gamma_q(t) - \frac{\tau}{2}\frac{d\varphi}{d\theta_q}\,\theta(t) \tag{13}$$

$$\theta_q(t+\tau)' = \theta_q(t+\tau) + \tau\frac{\gamma_q\left(t+\frac{\tau}{2}\right)}{M_q} \tag{14}$$

$$\gamma_q(t+\tau) = \gamma_q\left(t+\frac{\tau}{2}\right) - \frac{\tau}{2}\frac{d\varphi}{d\theta_q}\,\theta(t+\tau) \tag{15}$$

At time *T*, a metropolis acceptance criterion is applied to $\theta_q(t+\tau)'$ where $T = M\tau$ measures the total time, *M* is the number of steps, and τ is a time increment. The gradient equations used for our method were derived by Neil Houlsby but were not included in this thesis due to space limitations.

The samples from this approach provide enhanced exploration while ensuring that model parameters still converge to their appropriate values. After *M* steps, *M* samples from the parameter space $\theta^{1\ldots M}$ are returned by the HMC. The posterior distributions defined by each sample can then be averaged to estimate the value of the underlying distribution $\mathcal{P}$. This is more robust than averaging the parameter values themselves. For instance, assume we are trying to identify a bimodal underlying distribution using a single Gaussian mixture component. If the MCMC can explore both modes for sufficiently large *M*, then the average of posteriors would identify the two modes whereas the average of parameters would identify only the average of the two modes.

The implementation of the method, pre and post processing of data, plot generation, and simulations were all performed using Matlab and the lightspeed toolbox (Minka, 2006). The code for implementing the HMC was written in Matlab by Ferenc Huszár and Neil Houlsby.

3. Subjects and Experimental Procedure

3.1 Participant and Apparatus

For our study, Subjects performed 1000 trials of the OOO task followed by 1000 trails of the 2AFC task. The gap between tasks varied across subjects but was always less than 24 hours. In the 2AFC task, subjects were prompted with two facial stimuli and asked to choose the more "familiar" face. The OOO task presents subjects with three stimuli and asks the subject to choose the "odd one out". The specifics of the experiments are outlined in this section. The experimental parameters were kept constant across subjects and were validated as outlined in section 4.

3.1.1 Participants

Six volunteers and four paid subjects completed the experiment. Basic demographic information of our subjects is shown in Table 1. All subjects were naive to the purpose of the study and informed of their right to withdraw from the experiment at any time, without the need to state a reason. Subjects also gave written informed consent to participate in accordance with the requirements of the Psychology Research Ethics Committee of the University of Cambridge. The vision of our subjects was normal, or corrected to normal while performing the experiment.

Age (μ / range)	**Gender (% Male)**	**OOO Time in mins (μ / σ)**	**2AFC Time in mins (μ / σ)**
27.8 / 20	70%	75 / 35	2.58 / 0.55

Table 1. Subject demographic information. Starting from the left, the first entry shows the average and range of ages for our subjects. The second entry indicates the percentage of subjects that were male. The next two entries provide the mean and standard deviation of experimental length (not including breaks) in minuets for the OOO, and 2AFC tasks respectively.

3.1.2 Experimental Apparatus

The experiment was conducted on a World of Computers PC (1.877 GHz, 3.25 GB RAM) running Windows XP. Subjects sat with their eyes approximately 60 cm from the display. Stimuli were presented on a Dell 1800FP 18-inch flat screen monitor with a native screen resolution of 1280 × 1024 pixels and a refresh rate of 75 Hz. Each face was presented at a size of approximately 300 x 300 pixels. Subjects indicated their response preference using a mouse click on a facial stimulus.

3.1.3 Basel Face Model

Facial stimuli for the experiment were generated using the Basel Face Model (BFM) published by the Computer Science department of the University of Basel (Paysan et al., 2009). The BFM allows for a range of faces to be rendered as linear combinations of 199 structural and colour based principal components in face-space. The model had been calculated by the original authors from registered 3D scans of 200 faces (half male and half female).

3.2 Facial State Space

Valentine (1991) first proposed an abstract structure for face recognition where faces are encoded as locations within a multidimensional space. The model assumes that typical faces tend towards the centre of the space and distinctive faces are located at the periphery. We use this concept of a face-space, to provide an intuitive way of mapping a subject's mental representation over faces. As data volume requirements for inference scale with the dimensionality of the data, we chose to limit the dimensions of our face-space to the first two structural principal components of the BFM data set. Together, these account for roughly 18% of the total variance in facial structure. The colour component of all stimuli was set to

the mean value of the BFM. The value of a stimulus in our dataset describes its distance from the mean stimulus value (0,0), in standard deviations (i.e. z-score). The range of the state space was chosen to be ±5 standard deviations from the mean along both structural components (see Figure 2). This was decided based on our own judgement of what looked to be within the "normal" domain of faces.

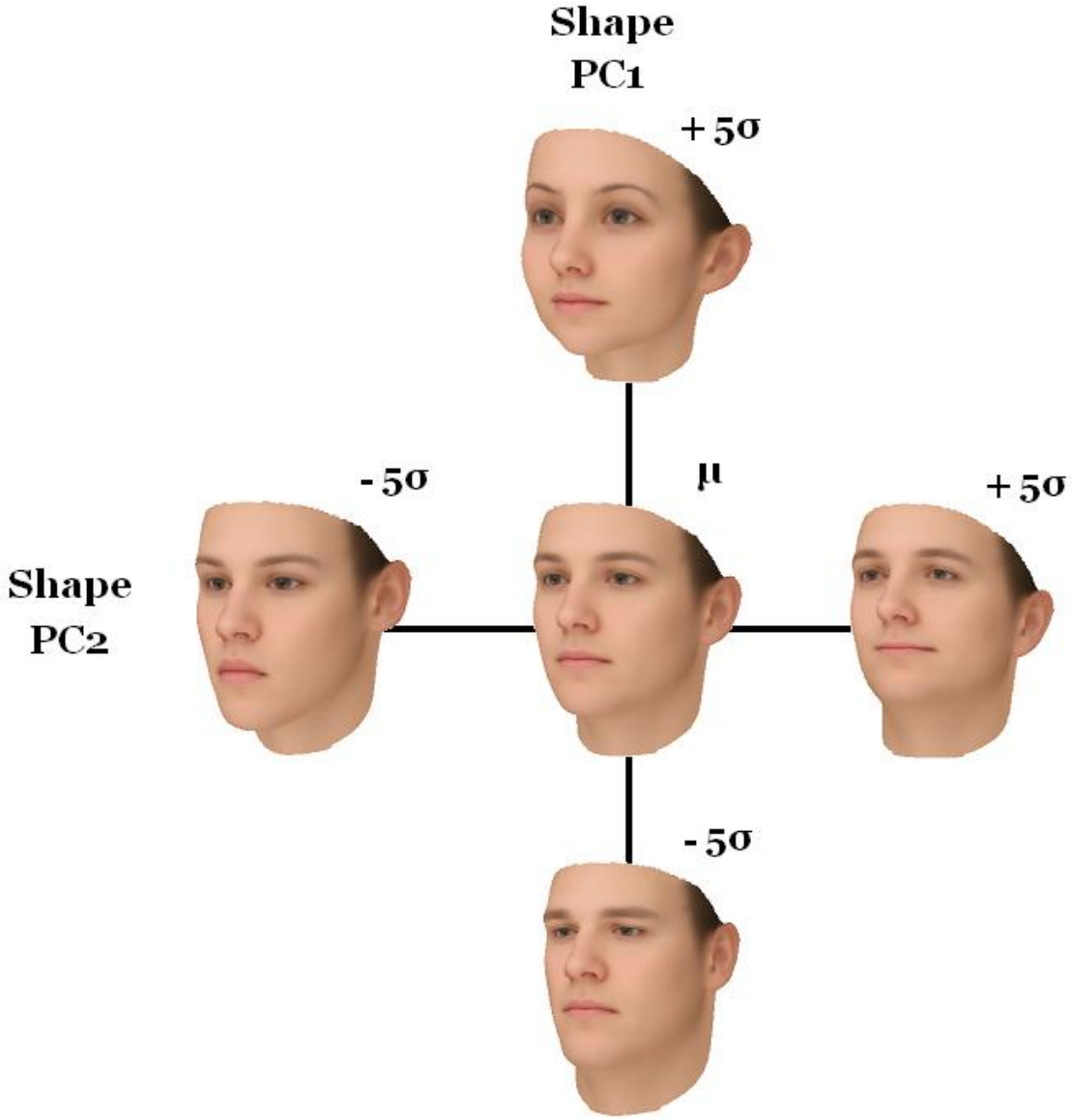


Figure 2. A depiction of stimuli mean and range. The mean face from the perspective of the experimenter according to the BFM and the structural changes that take place along either of the two principal components used to generate faces. Typical faces should cluster around the origin of face-space (i.e., the norm) and distinctive faces should be located in the periphery. The orientation of the faces was selected to increase the visual distinctiveness of each facial dimension according to our own judgement.

3.3 The Odd One Out Task

3.3.1 Setup

Before beginning the experiment subjects were given the following instructions.

"For this test we will show you 3 faces. Two people are from country A, one person is from country B. During each trial, click on the person from country B, the odd-one-out. Please press enter to verify that you have read and understood these instructions."

We chose to present the instruction in this way to prevent our subjects from making choices based on some sub-dimension of the space, such as gender. Subjects were asked to verbally verify that they had read and understood the presented instructions. After verification, subjects were presented with three adjacent facial stimuli for 1000 trials, and chose the odd-one-out by clicking the appropriate face. We recorded the length of each trial, from stimulus presentation to the time of selection. After every 100 trials subjects were given a short break. The stimuli and cursor were centred with respect to the screen. An example trial is shown in Figure 3.

Figure 3. A sample trial from the OOO experiment. The stimuli and cursor (not shown) are centred with respect to the screen for each trial.

3.3.2 Stimuli

For each trial, we generated facial stimuli by drawing two independent samples from a Gaussian distribution $\mathcal{N}(0,3)$. In combination, the draws correspond to a random point in the 2D face-space which we call the centre point. All centre points were bounded between ±3.5 in face-space. The actual stimuli presented to subjects for each trial were generated by moving a set distance along three randomly oriented vectors which originate at the randomly generated centre point. As all three vectors are 120 degrees apart, stimuli corresponded to the vertices of a randomly oriented triangle in face-space, centred on a point between ±3.5. See Figure 4 for a graphical illustration.

We generated centre points using a normal distribution centred on the origin because we assumed that the average subject's mental representation of faces will be biased towards the average (or most common) face and wished to explore this area with greater stimuli density. As the OOO contains more facial stimuli per trial than the 2AFC task, we also generated stimuli that spanned a larger area of the face-space for this task (increasing the domain and range by ±1). We assumed this would help minimize redundant stimuli and allow us to take advantage of the larger number of stimuli per trial in the OOO task.

3.3.3 Task

The first 100 trials of the task allowed us to gauge the sensitivity of our subjects to the distance between stimuli in face-space. They also provided a subtle way for us to train the subjects on how to perform the experiment. In these trials, subjects were provided with two stimuli which were close to one another in face-space, relative to a third stimulus. As trials progressed, the Euclidean distance between the three stimuli became closer to equal, eventually settling on a constant Euclidean distance of 1.5 between all three stimuli (the choice of this particular distance is explained later). In other words, for the first several trials

of the experiment there is an actual Euclidian odd-one-out that the subjects can choose. Figure 4 illustrates the difference between the early and late trials graphically.

The distances between stimuli in the remaining 900 trials are a constant 1.5. While our model is capable of accounting for differences in the distance between stimuli, equidistance was attractive because it encouraged subjects make choices on the basis of their mental representation alone as opposed to the Euclidian distances between the stimuli. Indeed, previous studies using faces have shown that high level subjective categories play a less significant role in decision making tasks when subjects are able to detect differences between stimuli in feature space (Kietzmann and König, 2010).

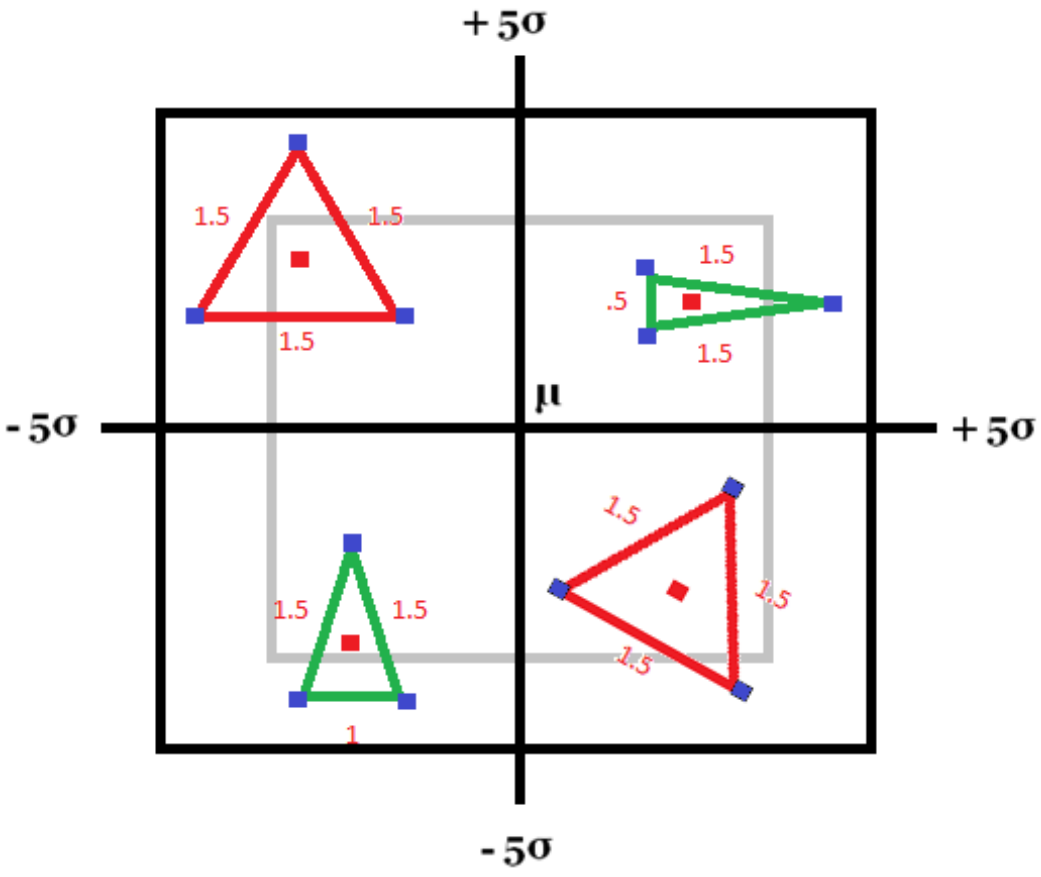


Figure 4. The generation of four stimulus sets for the OOO task. The boundary of the face-space is shown as a black box. The grey box shows the portion of the face-space in which centre points can be generated. The centre points are shown as red squares. Blue squares indicate the actual stimuli used for the experiment trials. Green triangles exemplify the first 100 experimental trials and red triangles exemplify trials 101-1000. The numerical distance between stimuli is shown in red.

Lastly, we note here that the stimuli from the last hundred trials of the task are permuted versions of the stimuli from trials 101-200. This provided a simple way to gauge the consistency of the subject during the experiment.

3.4 The Two Alternate Forced Choices Task

3.4.1 Setup

Before beginning the experiment subject were prompted with the following instructions.

"For this test we will show you two faces. Choose the face which is more familiar to you. Press enter to verify that you have read and understand these instructions."

Subjects were also asked to verbally verify that they had read and understood the presented instructions. After verification, subjects were presented with two adjacent facial stimuli for 1000 trials. After every 100 trials subjects were given a short break. Like the OOO task, the length of each trial was recorded. The stimuli were centred and the cursor position always began in the centre of the screen. An example trial is shown in Figure 5.

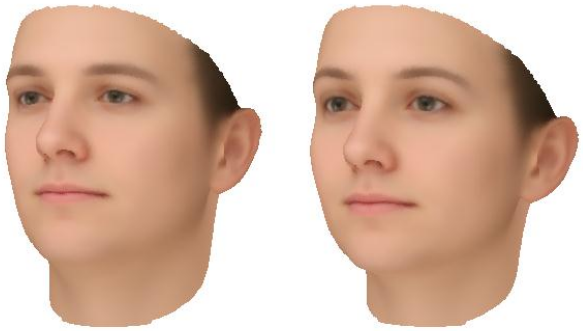


Figure 5. A sample trial from the 2AFC experiment. The stimuli and cursor (not shown) are centred with respect to the screen.

3.4.2 Stimuli

There were two types of trials for the 2AFC task, A and B. For trial type A, we generated stimuli by drawing two independent samples from a Gaussian distribution $\mathcal{N}(0,3)$ to use as a centre point in face-space. Stimuli were then generated along a randomly oriented vector in the space which crossed through this centre point. Stimuli were at an equal distance from the centre point, and a distance of 1.5 from one another in the feature space.

The second type of trial, type B, was generated from the subset of type A trials that had stimuli values outside the range of ±4 along either dimension. Those stimuli alone were regenerated using draws from $\mathcal{N}(0,3)$ and bounded between ±4. It follows that while all stimuli in the 2AFC task fall within the same range, type B stimuli are separated by a random distance in the state space, while type A are at a constant distance. We opted to vary the distance between stimuli in some trials to gauge the sensitivity of our subjects to Euclidean distances between stimuli in the 2AFC task. For a graphical illustration of the generation process for the two types of stimuli please see Figure 6.

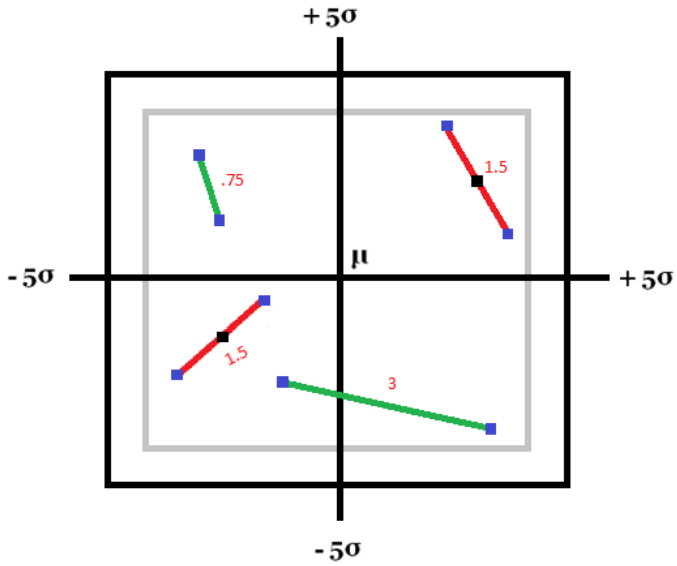


Figure 6. The generation of four stimulus sets for the 2AFC task. The boundary of the face-space is shown as a black box. The grey box shows the portion of the state space in which stimuli can be generated. Stimuli for type A trials are shown as red lines. The centre points of these trials are shown as black squares. Stimuli for type B trials are shown as green lines. There is no centre point for these types of trials and the stimuli are at a random distance from one another in face-space. For both trial types, blue squares indicate the actual stimuli used for the experiment and the numerical distance between stimuli is shown in red.

3.4.3 Task

The stimuli from the last 100 trials of the task were permuted versions of the stimuli from trials 1-101. Like the OOO task, this provided a simple way to gauge the consistency of the subject during the experiment.

3.5 Distance between Stimuli

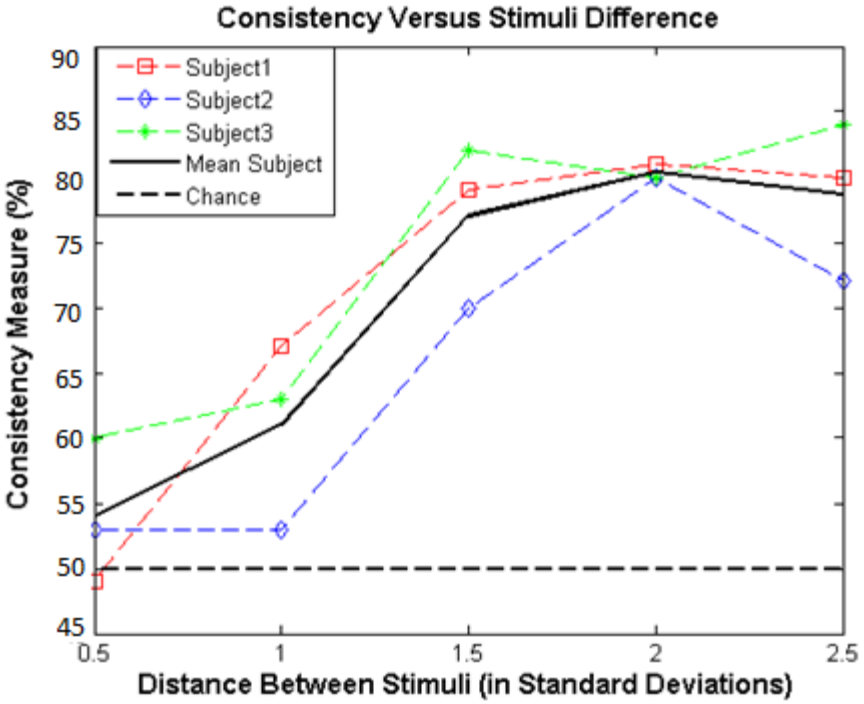
Prior to collecting data from actual subjects, we ran experiments using lab members to help optimize certain experimental parameters. As we mentioned earlier in this section, a majority of the stimuli presented to the subject in both the OOO and the 2AFC experiments are a distance of 1.5 from each other in face-space. This value was chosen based on a consistency experiment which mimicked the procedure of the OOO and 2AFC tasks. The goal of the experiment was to identify how to set the Euclidian distance between stimuli to minimize the behavioural noise of our subjects.

To gauge this, subjects were presented with 10 sets of 100 trials. The distance between stimuli in each trial set was constant, but the distance differed across sets. The last five trial sets were permuted versions of the first five sets. Otherwise, the stimuli generation process was identical to that outlined in the sections 3.3.2 and 3.4.2. To gauge consistency we measured how often the subjects made the same decision when presented with a permuted version of the same stimulus set. The results of the experiment (shown in Figure 7) indicate that subject consistency varied with the distance between stimuli in face-space. We chose the smallest distance which allowed for what we deemed an acceptable consistency value (1.5 x chance).

We observe In Figure 7 that the relationship between distance and consistency differs for each subject. Hence, it would be ideal to choose the distance between stimuli on a per-subject basis. Ultimately we decided against this as the combined length of a consistency,

OOO and 2AFC experiments made it experimentally unfeasible. Even though it does not produce the highest consistency according to Figure 7, we decided upon 1.5 as the distance between stimuli for the OOO and 2AFC experiments. We reason that responses to stimuli sets which are closer together in face-space convey more information about the subtleties of our subject's underlying distribution (especially if the distribution has low variance). Given a sufficient number of samples and a reasonable consistency value on the part of the subject, closer stimuli would allow us to construct underlying distributions which better represent what is in the minds of our subjects.

OOO:



2AFC:

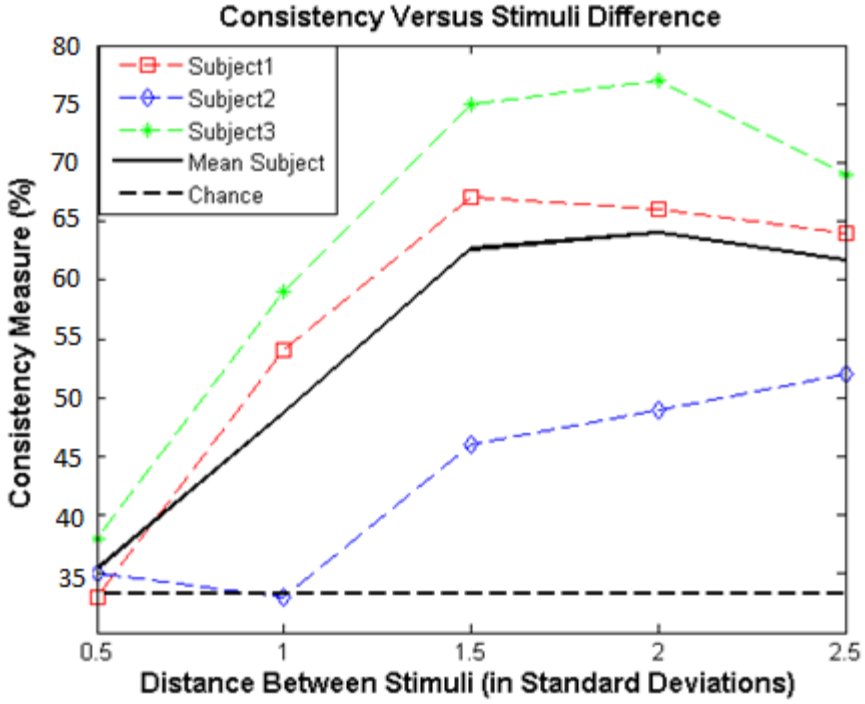


Figure 7. A consistency test for both tasks. The consistency of three non-naive subjects varied with the distance between presented stimuli. The mean of subjects is shown in black. According to the mean, subject consistency is acceptable for any distance from 1.5 – 2.5 but is best when the distance between stimuli is 2.

4. Validation

We performed extensive simulations of our tasks to validate the method's ability to approximate underlying subjective distributions. Several aspects of our experiment, including the number of stimuli and Markov chain steps were based on the results in this section. We will highlight only the most important subset of these simulations to demonstrate the integrity of our method and the chosen experimental parameters.

In this section, we simulate subject behaviour using a set of true model parameters, θ_0' and θ_p', which we then attempt to reconstruct using our method. The parameters in θ_p' corresponds to a randomly generated mixture of two Gaussian masses. The parameters in θ_0' correspond to a randomly generated set of behavioural parameters. We used the true parameters to generate responses to stimuli sets for 1000 trials according to the probabilities assigned by our model. Using the HMC, the stimuli and corresponding response vector were then used to reconstruct the true model parameters. This approach allowed us to gauge the efficacy of our method, assuming that real subjects also behave according to the underlying generative model.

4.1 HMC, Prior and Experimental Settings

Simulated subjects performed both tasks as described in section 3. Responses were generated in accordance with our probabilistic model of subject behaviour for each task. For all simulations, we used the same prior distributions over parameters as outlined in section 2. Stimuli sets were normalized between 0 and 1 before posterior estimation with the HMC. The Markov chain was set at a randomized state for each run with the following settings:

- Number of leapfrog steps: 20
- Parameter space step size: 0.03
- Number of Gaussian mixture components: 2

- Number of posterior samples: 50,000
- burn in = 5000

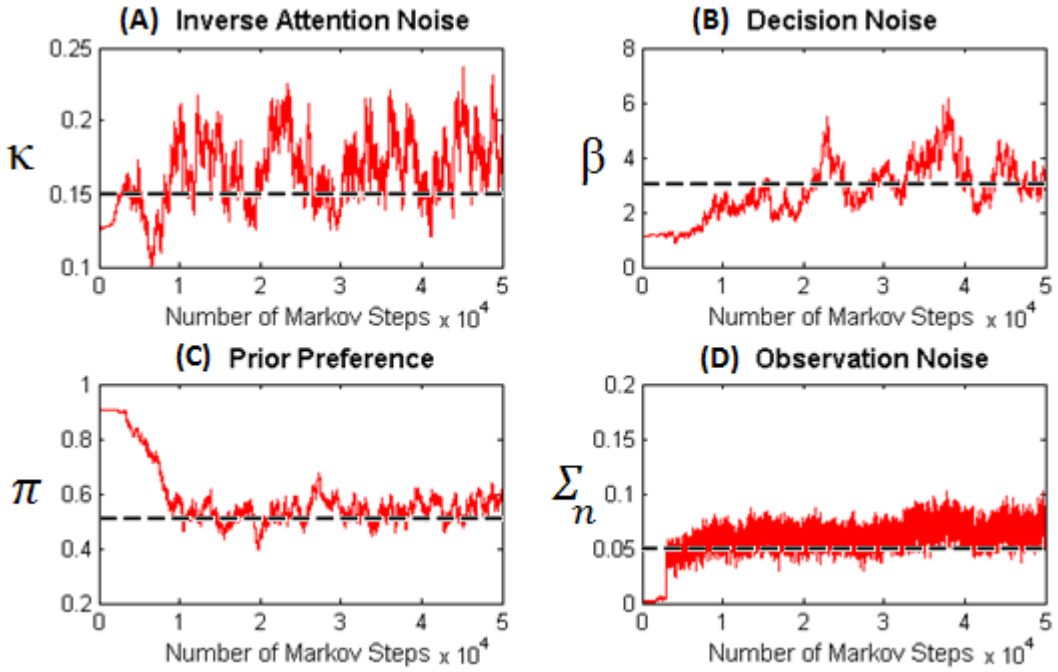
These settings placed our Markov chain's acceptance rate between 40% and 80% for any given run. The probability for each parameter being updated in a given Markov step was equal.

4.2 Convergence of the HMC

The first and most basic validation step we performed was to check if the HMC was sampling the parameter space in a sensible way for both tasks. The simulations in Figures 8 and 9 show the evolution of the reconstructed parameters in θ_p and θ_0 during a typical HMC run. Figure 8 A-D illustrates reconstruction of the κ, β, π and Σ_n parameters respectively using simulated 2AFC data while Figure 8 E-H illustrates reconstruction of the κ, β, π and $\Sigma_n$parameters respectively using simulated OOO data. The stable evolution of the parameters towards the true underlying values validated our ability to use the HMC to identify model parameters across tasks. Importantly, the results showed that parameter reconstruction using the OOO task requires more HMC samples to converge on certain parameter values, such as κ (Figure 8A).

Figure 9 A-B shows the evolution of the μ parameter in face-space during an HMC run using 2AFC data. Figure 9 C-D shows inference of the μ parameter using OOO data. In this case, the two modes (represented by black dashed lines) were successfully identified. The results were typical of our simulations where 2AFC values converged more quickly to the appropriate mean whereas OOO tended to perform more exploration of the parameter space. The variance of the sample density was clearly larger for parameter estimation using OOO data. The simulations assumed zero covariance and hence these values are not illustrated.

2AFC



OOO

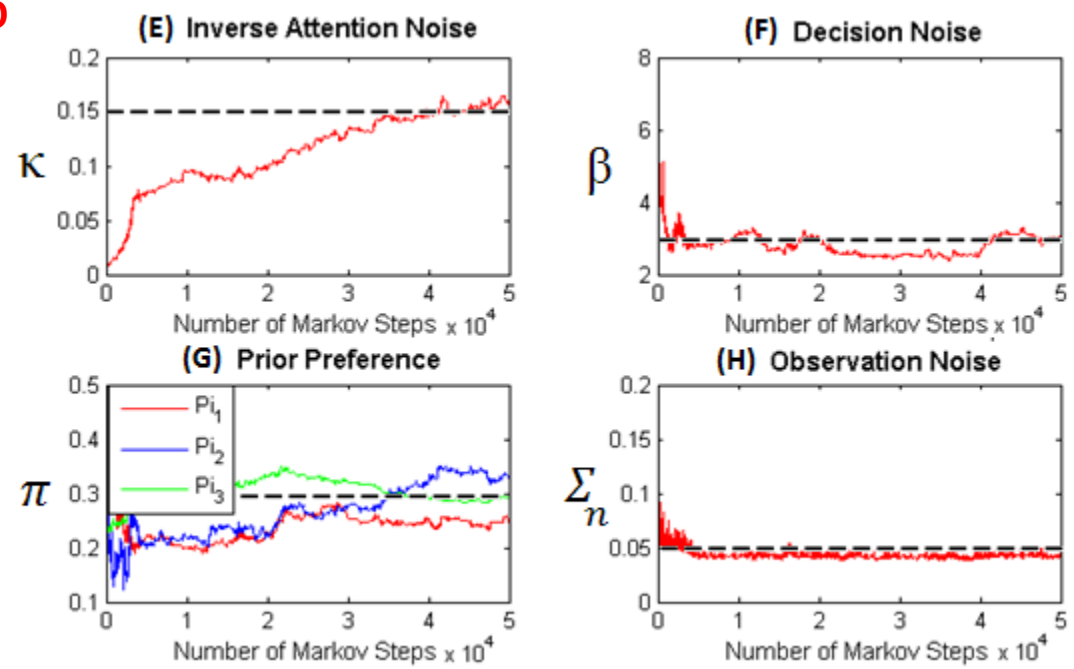


Figure 8. Inferring θ_0 parameters for a simulated OOO and 2AFC task. Solid lines represent the evolution of parameter states during HMC runs. Subplots A-D illustrate reconstruction of the κ, β, π and Σ_n parameters respectively using simulated 2AFC data. Subplots E-H illustrate reconstruction of the κ, β, π and $\Sigma_n$parameters respectively using simulated OOO data. The Markov chain ran for 55000 steps with a burn in period of 5000 steps. The underlying parameter values (shown as black dashed lines) were: κ = .15, β = 3 and Σ_n = .05 with uniform prior bias. We note here that the OOO did not converge on the κ parameter; this might indicate that more Markov steps were required.

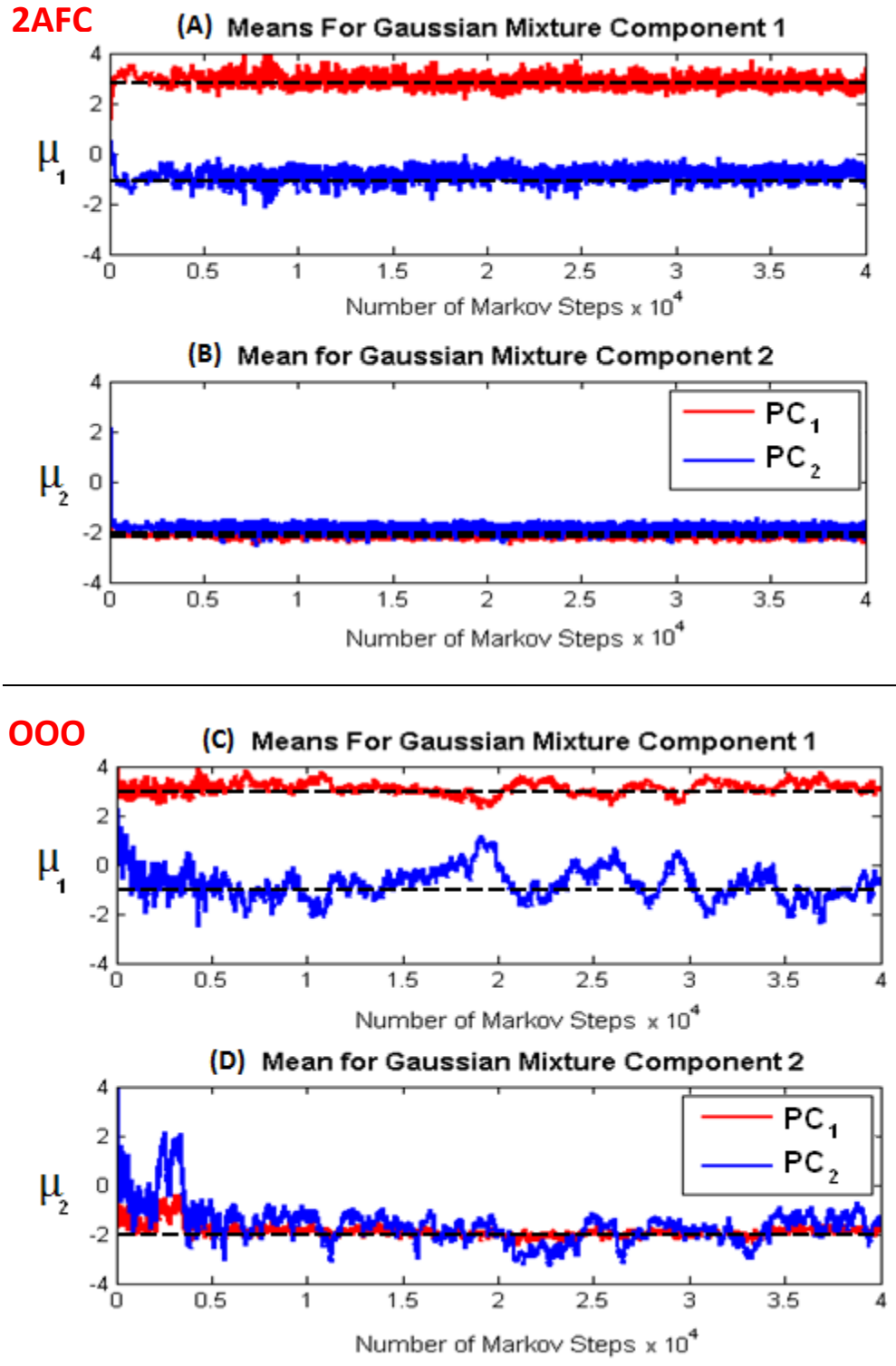


Figure 9. Inferring the two dimensional means for a simulated 2AFC and OOO task. Each solid line represents the evolution of a mean parameter in the two-dimensional space during an HMC run using both 2AFC (A-B) and OOO data (C-D). The Markov chain ran for 55000 steps with a burn in period of 5000 steps. The underlying distribution was bimodal with $\mu_1 = (3, -1)$ and $\mu_2 = (-2, -2)$. The underlying parameter values are represented by black dashed lines.

4.3 Parameter Reconstruction

We performed several simulations of both the 2AFC and OOO task with a large variety of underlying parameter values. The results of these simulations are shown in Figures 10 and 11. The simulations often progressed as expected, with the HMC sampling the parameter space in a sensible way and parameter values converging across both tasks. We note here that we did observe instances where the Markov Chain became stuck in a particular state, and hence failed to produce acceptable samples.

We quantified the performance of our method by performing posterior estimation according to the OOO and 2AFC task using 100 simulated subjects, each with unique underlying distributions and behavioural parameters. The accuracy of the extracted distribution, and parameter estimates were gauged using Jenson-Shannon (JS) divergence as well as linear regression. Figures 10 and 11 show the results of the simulations. Table 2 outlines the results of linear regression on the data and discusses the correlation between the estimated and the true parameter values.

The results make clear that our method is better at extracting some latent parameters and worse at others. Furthermore, it illustrates that the effectiveness of parameter estimation differs across tasks. We immediately notice two main issues with these results. Firstly, our method struggles to reconstruct the κ parameter (Figure 11a) for values larger than 0.65. This is no surprise as our prior over kappa strictly prevents the model from classifying all behavioural noise as being attention related. The second issue that the simulations make clear is that the method performs poorly at estimating the prior bias of our subjects using OOO data (Figure 11c). The inability of the method to reconstruct high values of κ, and the OOO task's inability to reconstruct the prior bias was noted for our analysis of actual subjects in the next section.

For values less than 0.65, the 2AFC task is the clear victor in κ parameter reconstruction. While neither task does a great job at estimating the β, σ^2 or Σ_n parameters, the 2AFC task is better at β estimation (Figure 11b), while the OOO task is better at σ^2 (Figure 10d) and Σ_n (Figure 11d) estimation. In general, both tasks perform decently at the reconstruction of the mean (Figure 10 a and c), with the OOO showing more error, on average, than the 2AFC. The mean JS divergence for posterior estimates from the OOO and 2AFC were 0.63 and 0.52 respectively. To understand these values we compared them to the average JS divergence between simulated posterior estimates across subjects, which was 0.66. This indicates that both estimates, on average, perform above the baseline with the overall 2AFC posterior estimates tending to be more accurate than the OOO posterior estimate. We also correlated the JS divergence measures with underlying behavioural parameter values for both tasks to gauge the effects of these parameters on the posterior extraction (See Table 3). While most of the correlations were not statistically significant, the direction of correlation is certainly what we expected to see. In a general sense, these results indicate that our posterior estimates improved as subjects behaved more ideally.

Parameters	**Slope from Linear Regression (2AFC / OOO)**	**Average R^2 error (2AFC / OOO)**
κ	0.86 / 0.29	0.001 / 0.033
β	0.14 / 0.05	8.33 / 8.81
π_1	0.66 / 0.024	0.008 / 0.18
Σ_n	-0.069 / 0.12	0.0023 / 0.0012
μ_1	(0.55 , 0.52) / (0.77, 0.53)	(2.24 , 2.07) / (2.12 , 3.01)
σ_1^2	(0.05 , 0.005) / (0.086 , 0.23)	(0.20 , 0.24) / (0.29 , 0.29)

Table 2. Efficacy of parameter estimation according to 100 simulations. Column 1 of the table lists parameters. Column 2 shows the slope of the reconstructed parameter values according to linear regression. Note that ideally, this slope should be 1. Column 3 shows the average squared error of the estimates. Linear regression on κ only included points where the underlying κ was less than 0.65.

(correlation / P-value)	**κ**	$\boldsymbol{\Sigma_n}$	**β**	**π**
OOO	.2 / .056	.12 / .24	-.12 / .24	.23 / .02
2AFC	.14 / .18	.097 / .36	-.14 / .17	-.02 / .78

Table 3. Correlation between JS-Divergence and behavioral parameters

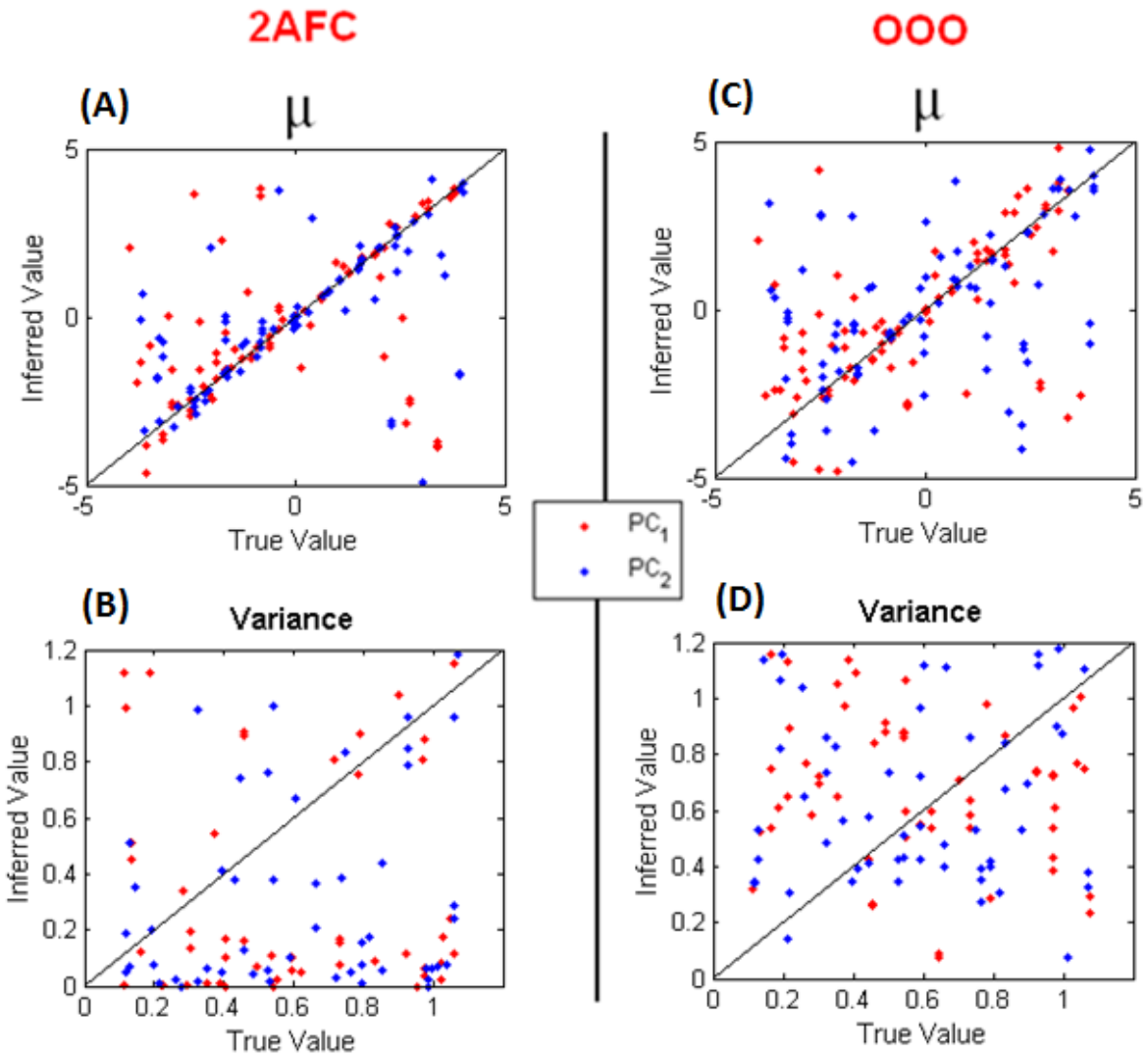


Figure 10. Simulation of 100 subjects with posterior estimates for mean and variance. The scatter plots illustrate the efficacy of our method at parameter reconstruction across both tasks. The means (A and C) and variance estimates (B and D) along both dimensions are illustrated. The 2AFC based estimates are shown on the left (A and B) and the OOO based estimates are shown on the right (C and D). For each parameter, we plot the "true" value versus the inferred posterior mean estimate. Each plot contains red and blue points. Red points correspond to the estimates along the first dimension of the feature space and blue points correspond to the estimate along the second dimension. The diagonal line (shown in black) represents an ideal reconstruction of parameter values.

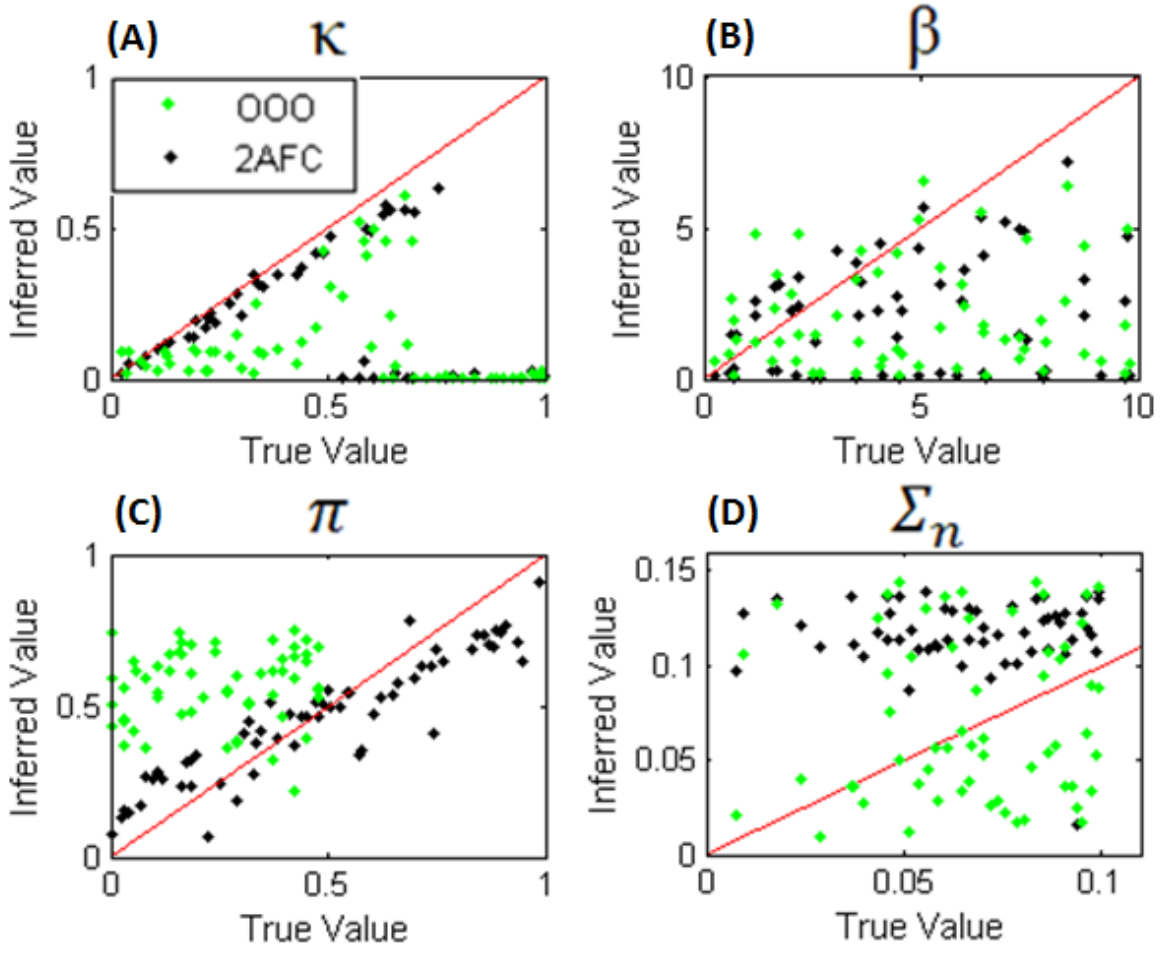


Figure 11. A simulation of 100 subjects with posterior estimates for θ_0 parameters. The scatter plots illustrate the efficacy of our method at behavioural parameter reconstruction across both tasks. Subplots A-D illustrate reconstruction of the κ, β, π and Σ_n parameters respectively. Each plot contains data from 100 2AFC and OOO simulations. The 2AFC based estimates are shown in black, and the OOO based estimates are shown in green. For each parameter, we plot the "true" value versus the inferred posterior mean estimate. The diagonal line (shown in red) represents an ideal reconstruction of parameter values.

A visual inspection of the reconstructed distributions of simulated subjects with low behavioural noise (Figure 12) also demonstrates this fact. Indeed, our method was able to reconstruct similar subjective distributions from both tasks when non-ideal behaviours were minimized.

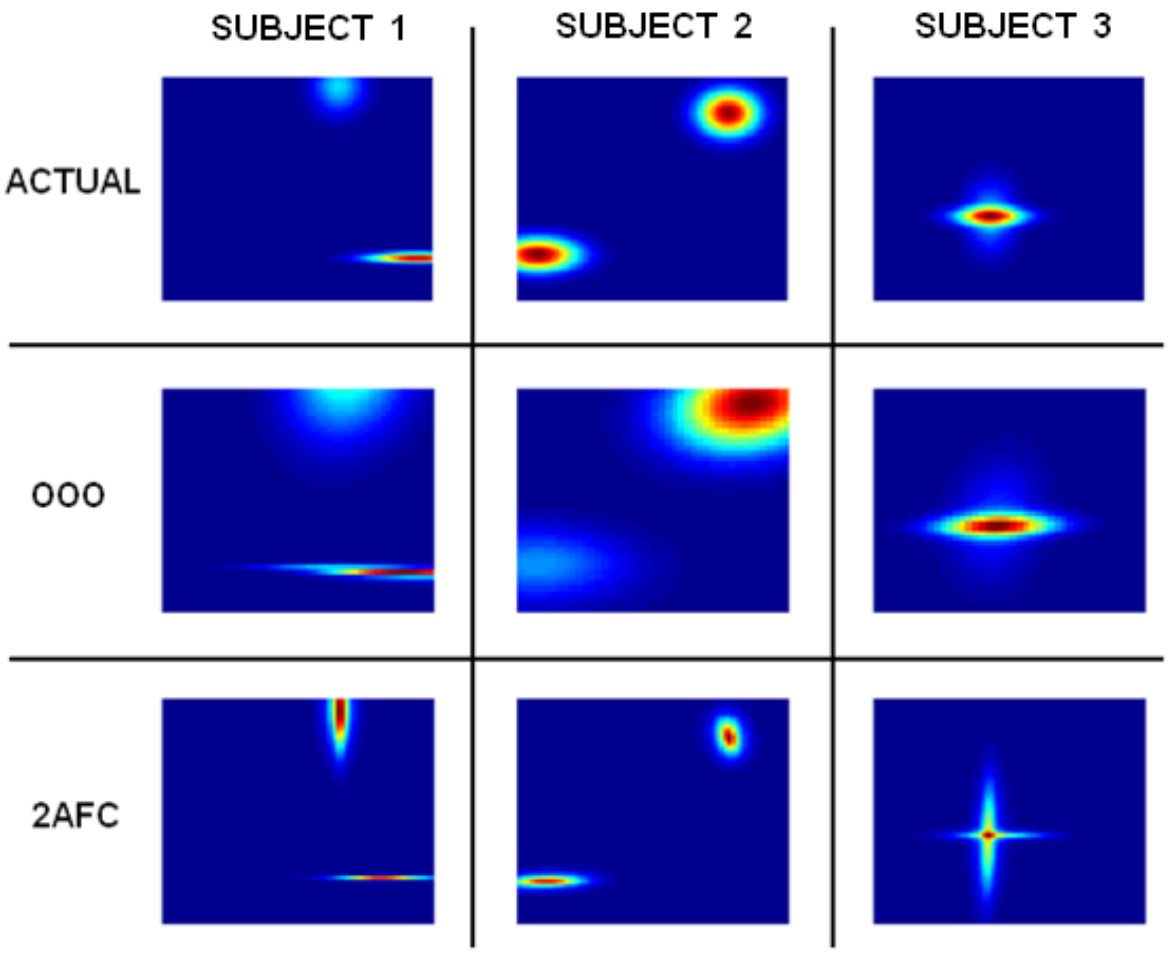


Figure 12. Results of three independent subject simulations (columns) with low behavioural noise using 1000 experimental trials, κ = 0.15, β =10, Σ_n = 0.05 and uniform prior bias. These results illustrate that the HMC is better at reconstructing the parameters of an underlying distribution from both OOO and 2AFC data when subjects exhibit lower behavioural noise.

There are practically two ways to gauge the overall efficacy of our results thus far: bias level and variance. It seems that in most cases the estimates are heavily biased. As the accuracy of our estimate requires the HMC to converge (which is dependent on m), we expect the results from our simulations would be improved with additional stimuli sets. As we mentioned before however, we were most interested in performing simulations that allow us to gauge the method's ability to extract model parameters for an actual human subject. Thus, we limited the number of simulated experimental trials to 1000, fearing that additional trials would bore an actual subject (i.e. increase κ) and impact our ability to extract the underlying distribution from the experimental data as a whole.

4.4 Effects of Test stimuli Distribution

We wanted to determine if the distribution of test stimuli would bias our method's ability to infer the underlying subjective distribution. To investigate this, we constructed an artificial subjective distribution and generated stimuli for both tasks that spanned:

1. The entire space (as presented to actual subjects)
2. A portion of the space not occupied by the subjective distribution

We then investigated how this impacted subjective distribution extraction. For this experiment we simulated a relatively well-behaved subject (β = 10 , Σ_n= 0.05 , κ = 0.15 and uniform prior bias) with a bi-modal underlying distribution parameterized by μ_1 = (-3,3) μ_2 = (3,-3). Stimuli for part 1 (Distribution b in Figure 9) of the experiment were generated as outlined in section 2, where centre-points were independently sampled from $\mathcal{N}(0,3)$. Stimuli for the second portion (Distribution a in Figure 9) of the experiment were generated in the same way as outlined in section 2, except that we changed the variance of the centre-point generating distribution from 3 to 0.5. This provided a scenario where a majority of test stimuli had no overlap with the underlying distribution.

Responses to the stimuli for both tasks were generated according to our model. We performed posterior estimation 25 times using both tasks and computed the Kullback-Leibler (KL) divergence between the average posterior estimate and the true underlying distribution. The distribution of stimuli and underlying representation are illustrated in Figure 13. A simple visual inspection clearly demonstrates that both methods are sensitive to the distribution of test stimuli, with the OOO task being far more sensitive than the 2AFC. Hence, we opted to distribute our stimuli as outlined in section 2. Based on these results, we must acknowledge that our method will encounter difficulty if a subjective distribution falls outside the area of the face-space where we have generated test stimuli.

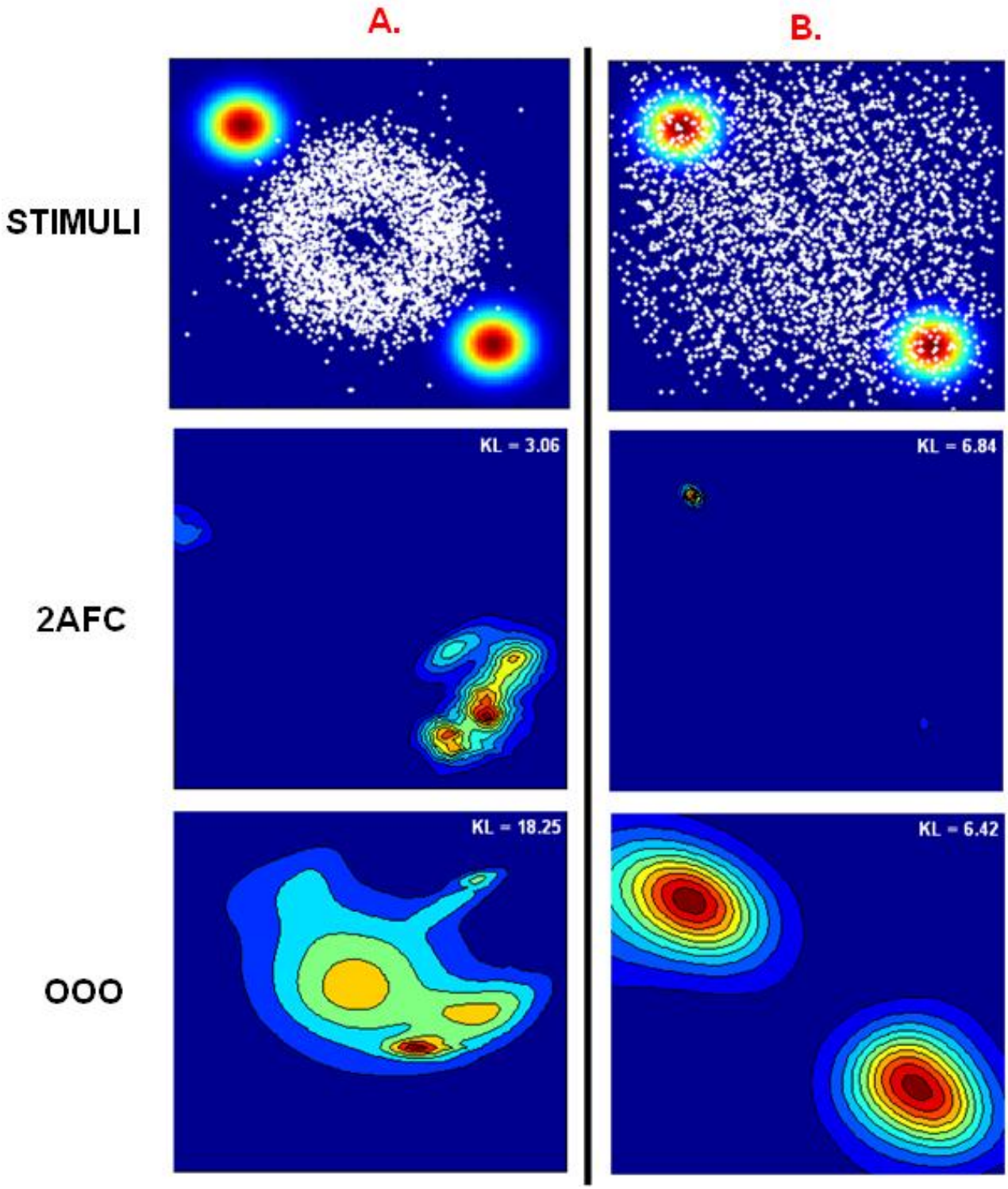


Figure 13. The effects of test stimuli distributions on posterior estimation for both tasks. Distribution A and B (row 1) show stimuli as white specks overlaid on the underlying subjective distribution. The images in rows 2 and 3 show the average posterior mean estimate across 25 HMC runs according to 2AFC and OOO tasks respectively. The KL divergence value between the estimated and actual posterior is shown in the top right hand corner of each image in rows 2 and 3.

5. Results

Having validated our method's ability to reconstruct subjective distributions on synthetic data, we now apply our method on actual subjects. We present in this section the reconstructed subjective distributions of 10 subjects according to data from the OOO and 2AFC tasks. All experimental parameters are discussed in section 3, and validated in section 4. The prior distributions for all parameters are identical to those outlined in section 2. The length of our Markov chain, burn in period, and other MCMC related settings are defined in section 4.1

To gauge the accuracy of reconstructed distributions we use them to predict subject's responses to novel stimuli sets both within and across tasks. Hence, we randomly selected 800 of the available 1000 experimental trials for posterior estimation leaving the remaining 200 trials for within-task predictions. We used all 1000 for cross-task predictions. Additionally, we present in this section an analysis of our subjects' behaviour during the task itself. This includes gauging relationships between reaction time and experimental parameters (such as the distance between stimuli), the values of our model parameters (such as prior bias), and the probabilities of stimuli assigned by the model itself. We feel confident that our results should reflect the underlying distribution of our subjects, assuming they behave according to the statistical model we outlined in section 2, with reasonable levels of non-ideal behaviour.

For each of our subjects, we present the parameter estimates, a graphical illustration of the subjective distribution according to data from both tasks, and the face which corresponds to the mean of the strongest weighted Gaussian mode**.** Figure 14 visually illustrates that the distributions extracted from OOO task data are similar to those extracted from 2AFC data for many of the subjects. To understand these results in a more detailed way,

we will turn our attention to the estimated and measured behavioural patterns presented in Table 4, and its corresponding p-value correlation matrix in Figure 15.

Before we delve into the analysis of our results however, we must remind the reader of the two potential issues outlined in our validation section and discuss how we attempted to address them when analysing actual subject data. First, we noticed while performing validation that our method fails to reconstruct values of κ exceeding 0.65. As we mentioned before, this is due to our strong prior over κ. The problem is clearly illustrated in Figure 11, where higher values of κ tend to be forced to values less than 0.05. Given this known issue, a low inferred κ value could in fact indicate that a subject was acting very noisily. To address this issue, we discarded subjects whose κ parameter measures were less than 0.05 when analysing correlations in subject data related to κ.

The second issue of concern from our validation section was the poor performance of the method at prior bias reconstruction from OOO data. As opposed to removing the measure completely, we manually set the prior bias of subjects to reflect the distribution of their choices during the experiment. We justify this course of action as the inferred prior bias parameter did not correlate with the distribution of responses for our subjects. Having mentioned both of the issues, we now proceed to discuss our results.

5.1 Extracted Distributions

The JS divergence values between the OOO and 2AFC related distributions (Table 4) reinforce what we visually observe in Figure 14. The average JS divergence value for extracted distributions was 0.69, with a standard deviation of ±0.19. When we compare the JS divergence of individual subjects to the average JS divergence between randomly generated modes (0.66), it is clear that the efficacy of our method varied strongly across subjects. According to the mean values of extracted behavioural parameters, we conclude that, on

average, our subjects behaved reasonably during the experiment. In the OOO task, which subjects performed first, we observed that reaction time correlated with the choice bias measure (-0.59, $p < 0.1$). The measure indicates a subject's bias towards each hypothesis according to the proportion of his/her responses during the experiment. This is a relatively intuitive correlation; a biased subject makes decisions faster than a non-biased subject. We also noticed that subjects generate responses faster, on average, during the 2AFC task.

Most interestingly, we observed a negative correlation between 2AFC reaction time and JS divergence (-0.56, $p < 0.1$), i.e. the extracted distributions are more similar when subjects took their time on the 2AFC task. As subjects performed the 2AFC task following the OOO task, we believe this is due to an increase in behavioural noise stemming primarily from experimental fatigue. Indeed, we observe in Table 4 that there is, on average, an increase in observation and decision noise for the 2AFC task. Additionally, we observe an average decrease in the consistency value, with the subjects performing at 18.9% above chance for the 2AFC task and 28.06% above chance for the OOO, which supports our claim of increased behavioural noise on the 2AFC task.

Figure 15 reveals that subjects' consistency measure in one task is strongly correlated with their consistency in the other (0.90, $p < 0.05$). Additionally, we noticed that the consistency measure correlates well with the β (0.78, $p < 0.05$) , Σ_n (-0.91, $p < 0.05$) and κ (-0.78, $p = 0.067$) parameters for the 2AFC task. This result is very encouraging, as it indicates that the parameters extracted from 2AFC data are indeed reflecting the behavioural noise of our subjects in a sensible way. Within the OOO task, there was a statistically significant correlation between subject consistency and the observation noise parameter Σ_n (-0.99, $p < 0.05$). The lack of statistically significant correlations with the attention and decision noise parameter is not surprising given what we observed in the validation section. This result

seems to verify that the method is not able to extract non-ideal behavioural parameters from the OOO task as well as it can from the 2AFC task.

Estimated behavioural parameters (2AFC/OOO)											
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**μ**
β	**2.25**	**2.9**	**1.56**	**2.3**	**2.6**	**.28**	**0.22**	**0.12**	**0.84**	**3.58**	**1.67**
	1.18	**0.93**	**1.32**	**5.11**	**5.06**	**2.6**	**0.9**	**1.28**	**1.78**	**4.87**	**2.5**
κ	**0.135**	**0.217**	**0.26**	**0.03**	**0.12**	**0.03**	**0.21**	**0.01**	**0.001**	**0.07**	**0.17**
	0.29	**0.06**	**0.12**	**0.23**	**0.25**	**0.12**	**0.04**	**0.004**	**0.01**	**0.25**	**0.19**
Σ_n	**0.08**	**0.0891**	**0.035**	**0.085**	**0.094**	**0.14**	**0.13**	**0.16**	**0.13**	**0.1**	**0.10**
	0.034	**0.03**	**0.07**	**0.02**	**0.03**	**0.09**	**0.1**	**0.11**	**0.15**	**0.022**	**0.07**
π x 100	**0.38**	**0.46**	**0.30**	**0.52**	**0.5**	**0.84**	**0.84**	**0.08**	**0.31**	**0.45**	**0.47**
	42/10	**40/6**	**43/10**	**50/7**	**43/6**	**18/59**	**49/21**	**29/33**	**9/46**	**47/9**	**37/21**
Measured Behavioural Parameters (2AFC / OOO)											
Consistency (%)	**84**	**74**	**79**	**92**	**86**	**43**	**67**	**45**	**28**	**91**	**68.9**
	72	**71**	**70**	**76**	**72**	**55**	**45**	**39**	**33**	**81**	**61.4**
Reaction Time (sec)	**2.75**	**3.46**	**2.54**	**3.20**	**2.52**	**2.45**	**1.58**	**1.82**	**3.05**	**2.45**	**2.58**
	5.41	**6.20**	**3.06**	**3.05**	**4.81**	**6.4**	**5.24**	**3.55**	**6.39**	**3.2**	**4.73**
Choice Preference (%)	**0.28**	**0.3**	**0.44**	**0.54**	**0.48**	**0.59**	**0.56**	**0.42**	**0.38**	**0.71**	**0.47**
	32/21	**23/24**	**47/24**	**44/28**	**28/41**	**26/64**	**44/24**	**48/20**	**44/37**	**49/42**	**38/33**
JS	**0.747**	**0.832**	**0.554**	**0.552**	**0.576**	**0.604**	**0.991**	**0.936**	**0.389**	**0.728**	**0.69**

Table 4. Estimated and measured behavioural parameters. This table summarizes the estimated and measured behavioural parameters for each subject. The table is partitioned into two sections: The first section (top half) shows estimated behavioural parameters, which includes β, κ, Σ_n and π. The second section (lower half) shows the measured behavioural parameters which includes: subject consistency as defined in section 3; the reaction time, which indicates the average trial length in seconds; and the choice preference measure, which indicates subjects' bias towards each hypothesis according to the proportion of their responses during the experiment. For each parameter, OOO estimates are shown in white rows, and 2AFC estimates are in grey rows. We show the choice and prior bias measure towards the face on the right for the 2AFC task and towards the right / left face for to the OOO task. The column on the right most side, displays the mean across subjects for each parameter. κ values in red were not included in calculations for the average or correlation between parameters. The JS divergence (green row) is also displayed in this table. Estimated parameters were computed using the mean of the last 10,000 of 55,000 Markov steps.

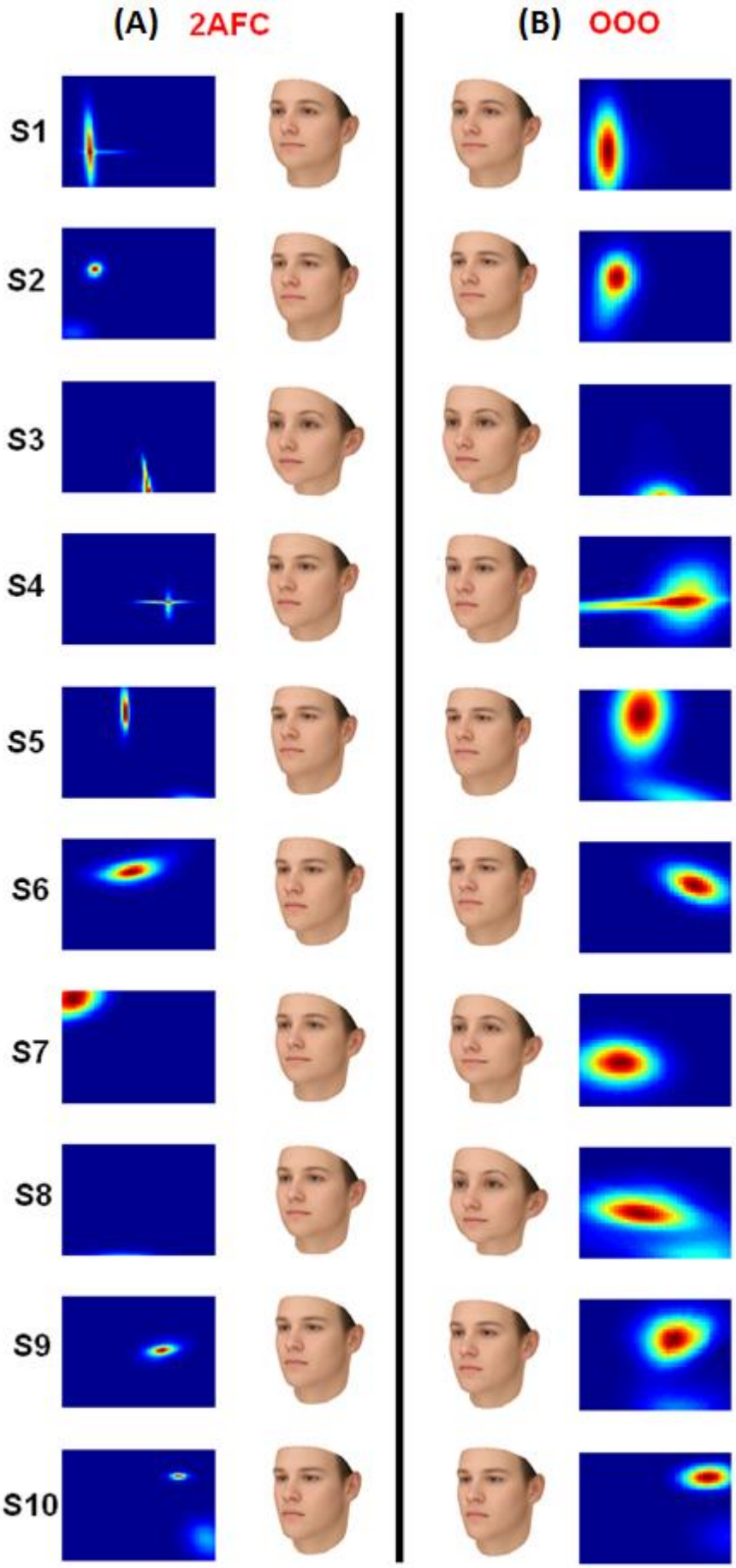


Figure 14. Extracted subjective distributions and corresponding faces. Column A on the left shows the posterior estimate and corresponding mean face according to 2AFC data. Column B on the right shows the posterior estimate and corresponding mean face according to the OOO data. Each row corresponds to a subject.

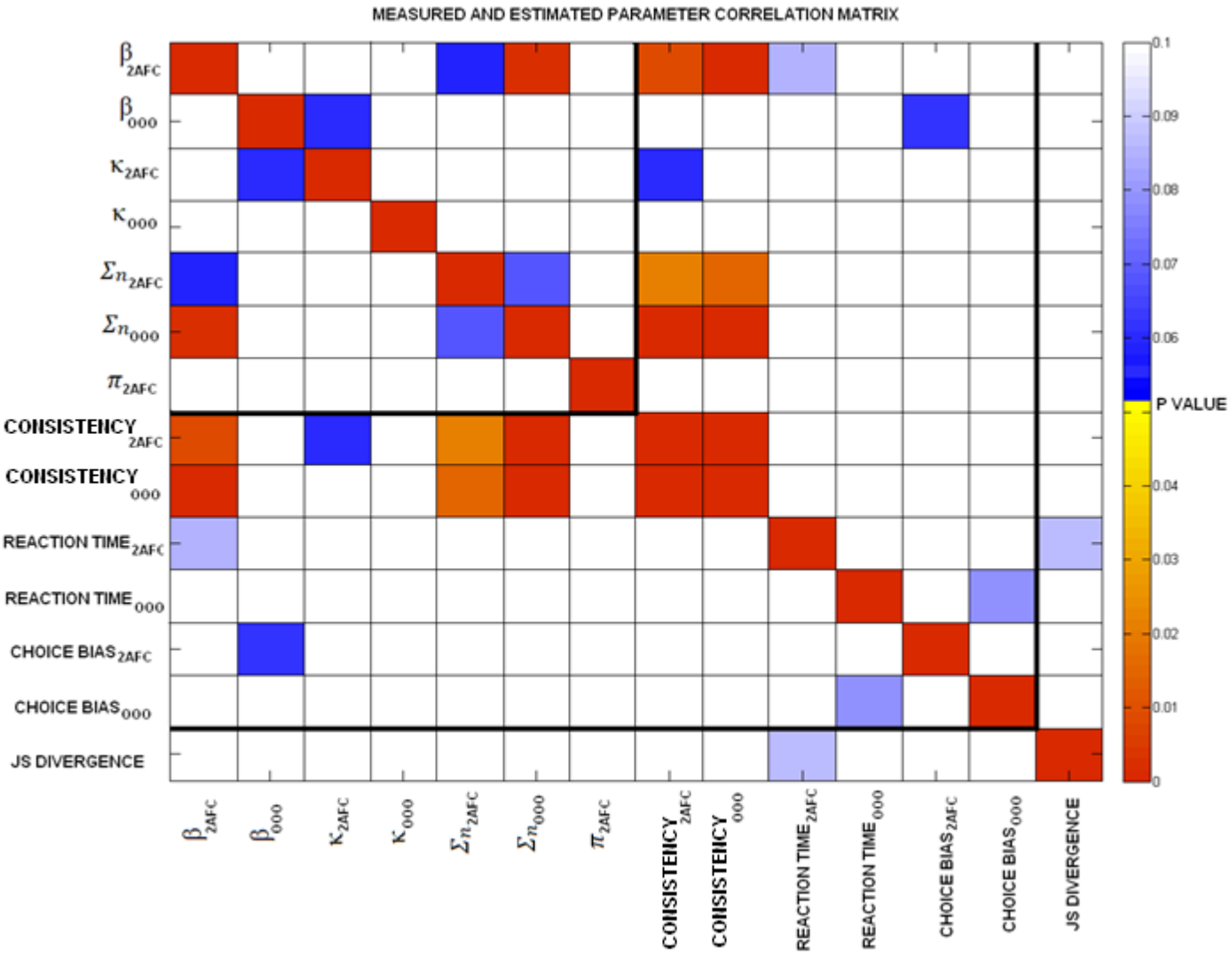


Figure 15. *p-value* correlation matrix for measured and estimated parameters. Parameters which exhibit statistically significant correlations are shown in colour. Values from yellow to red have *p-values* below 0.05 while values from white to blue have *p-values* below 0.1. The prior bias measure for the OOO was not included in this analysis.

What we have discussed up to this point are the statistically significant correlations within our results. We will now discuss other results that are less significant but still provide interesting insight into the efficacy of our method. While there is no statistically significant correlation between the 2AFC π parameter and any other parameter, it is worth mentioning that it correlated most strongly with the 2AFC choice bias measure (0.47, p = 0.167). As both parameters measure the same quantity, we expected this correlation to be much stronger. We

can only interpret this as an indication that the method is sub-optimal in 2AFC prior bias estimation.

We had expected the consistency measure to strongly correlate with the JS divergence values but observed no such relationship. We are unsure why this correlation did not arise. One possibility is that the last 100 experimental trials are not a good indication of the average consistency of the subjects during the entire task. We expect that a subject's attention noise would increase over the course of the experiment (Loewenstein, 1992); hence, the consistency measure is more realistically an indication of the subjects at their most noisy. In future studies, it would be useful to extend this consistency period, and intersperse the consistency trials across the entire run as opposed to a solid block towards the end of the experiment.

Not surprisingly, we encountered in this section some of the same issues we noticed during our validation of the method. Specifically, the OOO task does not perform behavioural parameter estimation as well as the 2AFC task. Having acknowledged this, we wish to reiterate that the results in Figure 14 are actually very encouraging. We strongly believe that the experiment is worth repeating, with some of the changes already discussed in this and the following section.

5.2 Validation According to Predictive Power

The results we have presented thus far visually indicate that the 2AFC and OOO task are capable of extracting the same underlying mental representations from our subjects. To more rigorously gauge the similarity of the reconstructed distributions shown in Figure 14, we performed cross-task validation. This process entails predicting a subject's responses to novel stimuli sets both within and across tasks using posterior estimates. For within-task validation, we randomly selected 800 of the 1000 experimental trials for posterior estimation.

The remaining 200 trials were then used to check within-task predictions against an actual subject's choices. For cross-task validation we used the posterior mean estimate from one task to predict the subject's responses in all 1000 trials of the other task. We present the results of this analysis in Figure 16.

Our choice overlap measure represents the overlap between the responses of highest probability according to our model, and those of the actual subjects. We notice that the overlap, on average, is well above chance for all within- and cross-task predictions. The posterior extracted from OOO data, on average, predicts 2AFC responses at 20.3% above chance while the posterior extracted from 2AFC data predicts OOO responses at 12% above chance. This result affirms that knowing the posterior estimate allows the method to predict the future behaviour of the subject. The OOO posterior is better at predicting 2AFC responses because of the higher variance estimates of the OOO distribution (see Figure 10).

If we examine the geometric mean of the predictive power assigned by our model to the choices made by subjects, it provides a more stringent analysis of the model's performance. Even then, the method still performs well above chance on within-task predictions. Where we begin to notice deficiencies is with cross-task predictions. As with the choice overlap metric, the posterior extracted using the OOO is a better predictor of behaviour in the 2AFC task than the 2AFC posterior is as a predictor for the OOO task. While the average predictive performance according to the geometric mean is close to chance, the large standard deviation of 14.5 when using OOO to predict 2AFC indicates that the performance varied greatly across subjects. Subjects 2, 5 and 10 for instance, all exhibited cross-task predictive performance above chance. Given the nature of our generative model in Equation 6, this result is not terribly surprising. The noisier the model believes a subject to be, the more uniform the probabilities it will assign to the possible responses. In contrast, when the model believes that a subject is behaving with low noise, then it is more likely to

assign high probabilities to responses that correspond to high probability regions of the posterior. For subjects with low κ values, the model will become confident in answers it provides.

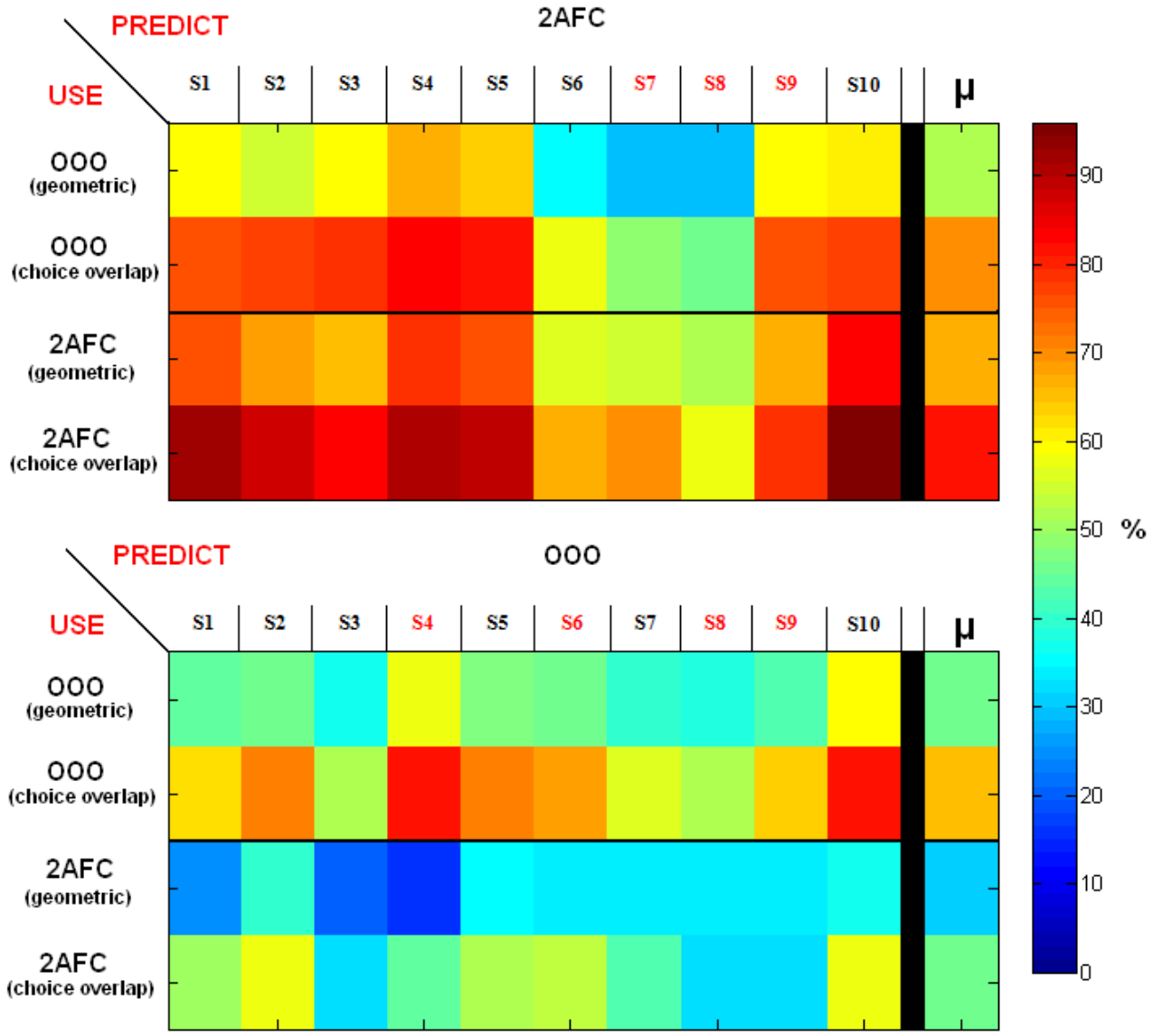


Figure 16: Model predictive power within and across tasks. The table is divided into two main sections. The first section (top half) indicates how well we predicted behaviour on the 2AFC task using the inferred OOO and 2AFC posteriors. The second section (bottom half) shows how well we predicted behaviour on the OOO task using the inferred OOO and 2AFC posteriors. For each trial we use the posterior to assign probabilities to each of the responses. We then calculate two metrics, a geometric mean of the probabilities corresponding to the subject's actual choices, and a choice overlap value, which represents the overlap between the predicted, and actual responses. The rightmost column shows the average predictive power across subjects using both metrics. Subjects shown in red had failed κ reconstructed (i.e. $\kappa < 0.05$).

Given the known problems with κ value miscalculation, it is therefore not surprising that several of the subjects with κ values near 0 also exhibit poor predictive performance when using the geometric mean. We noticed that if we exclude subjects with values of κ close to zero ($\kappa < 0.05$) then we increase the average predictive performance by 5.2% (using OOO predict 2AFC) and 1% (using 2AFC to predict OOO).

If nothing else, the visual presentation of the results in Figure 14 should provide a level of confidence that we were able to successful extract our subjects' mental representation over faces. Differences we observe in these distributions most likely stem from the behaviour of the subjects in each task, and deficiencies in parameter estimation.

5.3 Analysis of Reaction Times

In addition to extracting the mental representations over faces of our subjects we also performed a Pearson's correlation between our subjects' reaction times, the response probability of the selected stimuli, and the Euclidian distances between the stimuli.

Given models of choice reaction time in the literature, we expected reaction time for the responses to decrease as the probability assigned to the chosen hypothesis increased (Stone, 1960). Surprisingly, we did not find a statistically significant relationship between these measures for any of our subjects. We anticipate this may be due to noise-related factors.

Additionally, we investigated the Pearson correlation between our subjects' reaction times and the Euclidian distances between stimuli. The results are presented in Figures 17 and 18. While the result was not uniform across subjects, we observed a clear negative correlation between distance and choice time for the well-behaved subjects in the 2AFC task. In other words, the less noisy subjects took longer to generate responses in the 2AFC task when presented with stimuli that were closer together in the Euclidian space.

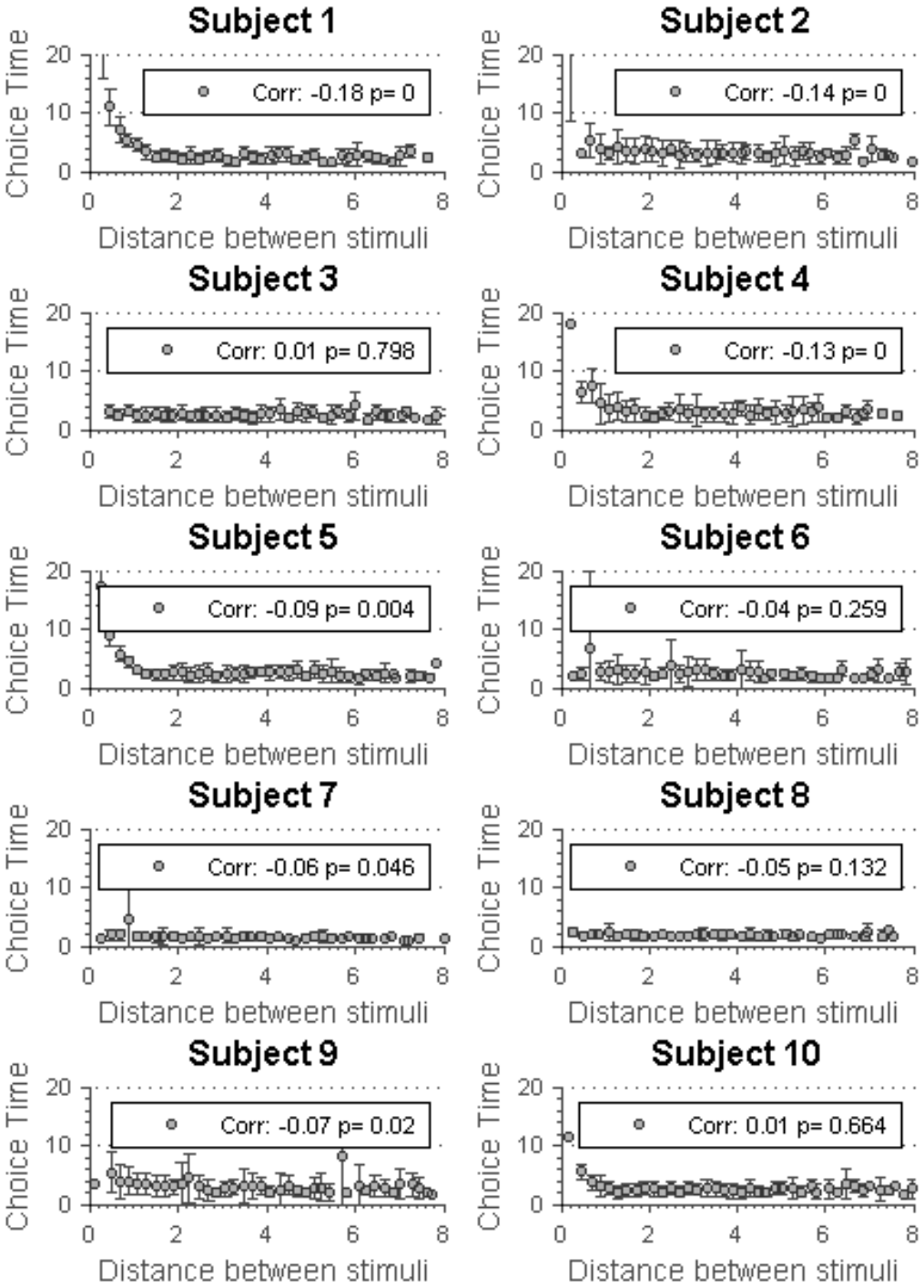


Figure 17. 2AFC stimuli distance versus reaction time. The error bars represent standard deviation. Choice time is in seconds.

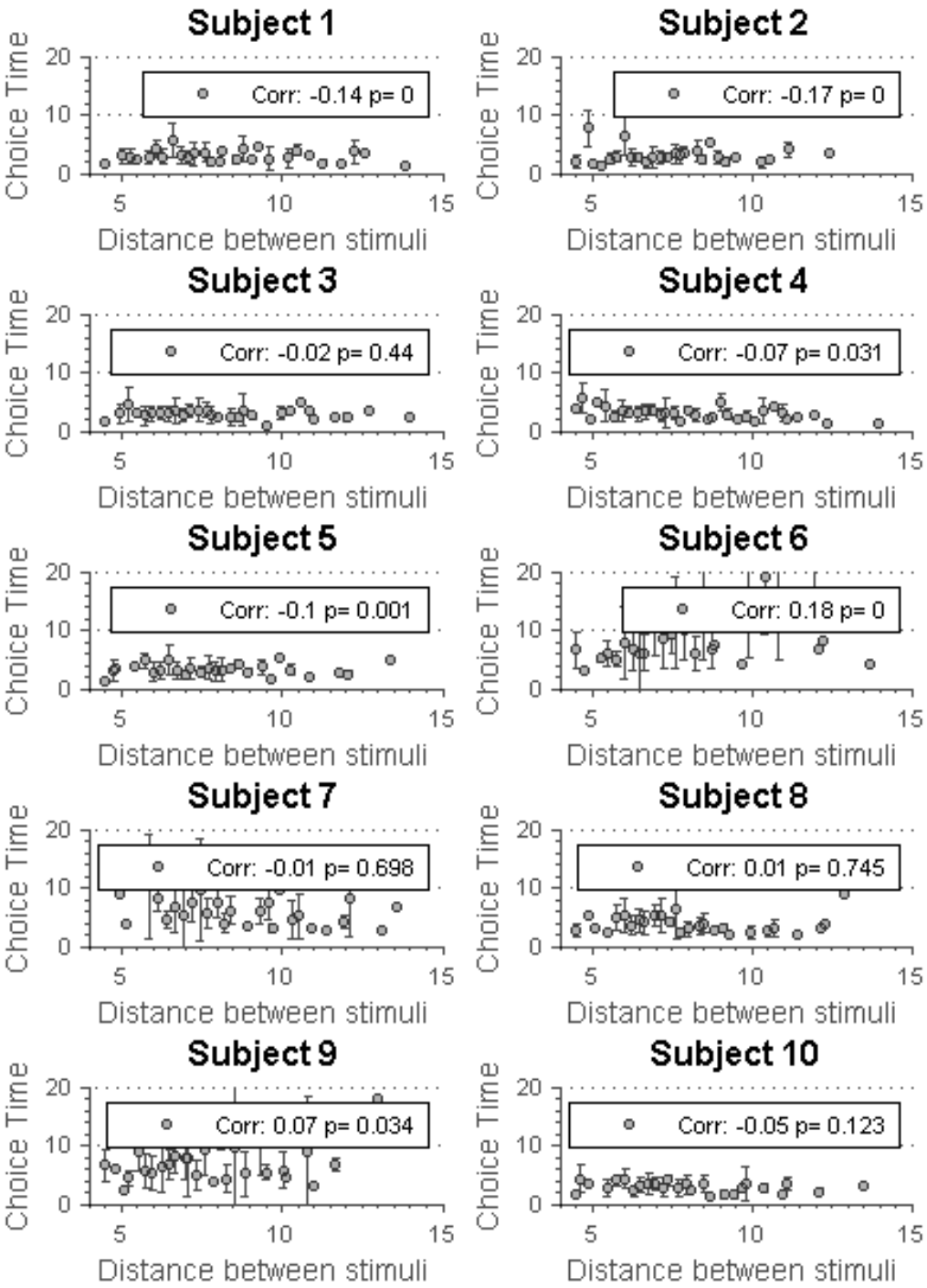


Figure 18. Impact of OOO stimuli distance on reaction time. The error bars represent standard deviation. Choice time is in seconds.

We observe a similar relationship between stimuli distance and reaction time for the well-behaved subjects in the OOO task, although it is admittedly noisier. Most interestingly, subjects who exhibit this correlation for both tasks also tended to have lower JS divergence values between their extracted distributions. The results and those presented in Figure 7 lead us to believe that subjects have an optimal Euclidian distance between the presented stimuli which minimizes their behavioural noise and increases their comfort level with the experiment. In future studies, we would encourage others to gauge this distance as part of the experimental design.

6. Discussion

6.1 Summary

While the performance metrics of traditional experimental psychology such as reaction time, choice bias, or percentage correct rates can provide some intuition about mental representations, they are fundamentally incapable of providing the level of detail that probabilistic models of cognition provide. Indeed, the application of probabilistic models for complex problems in the cognitive and brain sciences is growing ever more popular (Chater et al., 2006).

The central aim of this study was to see identify if we could infer a subject's mental representations over faces using discrete decision tasks. To do this, we outlined a statistical model of subject behaviour during the tasks, and estimated the likelihood of subjective distributions using a subject's responses and a flexible prior. As Bayesian inference was intractable, we implemented a Hybrid Monte Carlo algorithm for numerical estimation of the posterior, which we validated on simulated datasets. Our results show that we were able to use our tasks to extract detailed information about a subject's underlying mental representation over faces and make predictions about future behaviour both within, and across tasks with reasonable accuracy. We take these results as evidence that the subjective distribution of interest was actually inferred.

It should be stressed here that our goal was to parameterise not only subjects' mental representation, but also their behavioural characteristics. For this reason we also analysed our subjects' response patterns, reaction times, and consistency to help us gauge the accuracy of inferred behavioural parameters of our subjects during the task. The correlations between several of our inferred and measured behavioural parameters discussed in section 5, provides further confidence in the general integrity of our method. Ultimately, we feel that the results

provide sufficient evidence in support of our experimental design, and the model's ability to infer complex subjective distributions.

6.2 Related Work

The idea that humans harbour mental representations over faces is not a new concept. There are other studies that have also performed investigation into the mental representation over faces (Martin et al., in press 2011; Tanaka et al., 1998; Nishimura et al., 2009). Our study is the first we are aware of that uses statistical modelling of subject behaviour in discrete decision tasks to map the mental representations of faces. Our effort builds on the works of others that used statistical models to infer subjective distributions in a more general sense (Huszár et al., 2010; Paninski, 2006; Sanborn and Griffiths, 2008).

As in our study, Sanborn & Griffiths (2008) defined a subjective distribution as a static quantity which can be inferred. To extract the underlying distributions of their subjects, they presented them with two stimuli and asked them to choose the one which was more familiar. The chosen stimuli were then presented as one of the two stimuli in the next trial. This sequential design relates the abstract concept of stimuli familiarity with stimuli probability according to the subjective distribution. The subject response vector was used as an acceptance criterion for a Metropolis sampling algorithm with the stimuli themselves acting as samples from the underlying distribution. It should be noted that the Sanborn and Griffiths approach relies on a particular task (2AFC) to interpret choice probabilities as acceptance/rejection in a Markov chain. One clear advantage of our method is that it applies to multiple tasks and therefore also allows for cross-task analysis. On the other hand, Sanborn and Griffiths method is nonparametric, which would likely work better in high dimensions or structured stimulus spaces, where a mixtures of Gaussians may be a poor model of the subjective distribution. We should mention that there are other methods, such as reverse

correlation, that could be used for recovering some form of quantitative description of mental representations. Besides this, we are not aware of other methods directly targeted at estimating our particular object of interest.

As we have mentioned before, our method is compatible with multiple task types. We ultimately settled on the OOO because Kemp (2005) demonstrated that the subjective similarity of two quantities can be understood as the probability that both quantities were drawn from the same underlying distribution. This fact allowed us to construct the likelihood function for the OOO task.

The work of Huszár (2010) is similar to our own work in many ways. In fact, our work functions as an extension of their study. They also extracted subjective distributions using discrete decision tasks, and performed cross-task validation in the same way. The similarity between our studies makes Huszár (2010) a good point of reference and we are happy to note that we found similarity between the results presented in their study and our own.

6.3 Known Issues and Possible Improvements

While the results shown in section 5 are encouraging, there is certainly room for improvement and further investigation. We believe that the sub-optimal results from several of the tested subjects arise from confusion about the experimental instructions. We reiterate, however, that the wording of instructions was intended to prevent subjects from evaluating the faces according to sub-dimensions in the face-space, such as gender, as opposed to the concepts of interest (familiarity and odd-one-out). This, of course, risks elimination of gender based discrimination altogether which is certainly not ideal. Additionally, the stimuli presented to our subjects were exclusively Caucasian in colouring. This may have inhibited the performance of subjects whose subjective distribution falls outside of the Caucasian

portion of face-space. We anticipate that including faces outside the Caucasian spectrum would increase the performance of certain subjects.

Given the use of the HMC algorithm, and the large volume of data (1000 trials) we expected the extracted distributions to exhibit much stronger similarity than what was observed in this study. It is possible that we may have seen better results if we had relaxed the prior over the κ and prior bias parameters for the OOO further.

Our original intention in this study was to map the subjective distribution of our subjects across five dimensions of face-space, running the two-dimensional study as a proof of concept. When we found that the method struggled to reconstruct the underlying distributions in initial validation using five dimensional datasets, we decided to use the lower dimensional feature space. One of the greatest weaknesses of the method is the heuristic way in which we generate stimuli in the feature space. An active learning framework, where stimuli in subsequent trials depend on the responses from previous trials, would allow us to generate better quality stimuli sets, collect better quality data, and enhance the accuracy our method. This in turn may allow us to increase the dimensionality of our data sets.

In our study, we selected the number of mixture components to be static, which inhibits the flexibility of the model. While we are able to work around this by running multiple Markov chains in parallel, it would be advantageous to implement a variational Bayesian mixture model, especially for a higher a dimensional feature space (Ghahramani and Beal, 2000).

Lastly, while the results in this thesis stem from the 2AFC and OOO tasks, there are several other psychometric tasks which would be interesting to compare against those outlined in this study. We admit that the results we present lack a solid baseline for comparison; it would have been useful to follow the example of Huszár (2010) and compare the predictive performance of our method against other methods, such as a Gaussian process

classifier model. We hope that others continue to improve on the methods, and experimental design outlined in this document and apply it to increasingly complex data sets.

7. Acknowledgements

I want to begin by humbly thanking Professor Daniel Wolpert, my supervisor, for his patience and support throughout my year in Cambridge. His faith and guidance during tough times have led me from confusion to confidence, for which I will always be grateful and indebted. I owe an identical debt of gratitude to Dr. Máté Lengyel, one of the most patient and honest men I have ever had the pleasure of working with. His relentless pursuit of the truth inspires and fascinates me.

I must also thank Professor Zoubin Ghahramani and Dr. Carl Rasmussen for teaching me all about machine learning. Ferenc Huszár and Neil Houlsby deserve recognition for coding the HMC, refining the methods, and tirelessly answering my many questions. I also wish to thank Gergő Orbán, who served as my mentor, guide, and friend during tough times. Also, I thank the University of Cambridge, and more specifically, the Computational and Biological learning group. Lastly, I give a very special thanks to the Gates Cambridge Trust, which supported my studies during an amazing year of learning and discovery.

8. Bibliography

Beale, J.M. and Keil, F.C. (1995). Categorical effects in the perception of faces. *Cognition, 57*(3):217-239.

Chater, N. and Manning, C.D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*(7): 335-344.

Chater, N., Tenenbaum, J.B., Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7): 287-291.

Craik, K.J.W. (1967). *The nature of explanation*. Cambridge University Press.

Dayan, P. and Abbot, L.F. (2001). *Theoretical neuroscience: Computational and mathematical modelling of neural systems*. The MIT Press

Duane, A. (1987). Hybrid Monte Carlo. *Physics Letters B, 195*(2): 216-222

Ericsson, K.A. and Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87*(3): 215.

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The Visual Neurosciences*. The MIT Press.

Ghahramani, Z. and Beal, M.J. (2000). Variational inference for Bayesian mixtures of factor analysers. *Advances in Neural Information Processing Systems, 12: 449-455*

Gold, J.I. and Shadlen, M.N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36*(2): 299-308

Huszár, F., Noppeney, U., Lengyel, M. (2010). Mind reading by machine learning: A doubly Bayesian method for inferring mental representations. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2810-2815.

Kemp, C., Bernstein, A., Tenenbaum, J.B. (2005). A generative theory of similarity. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society.*

Kietzmann, T.C. and König, P. (2010). Perceptual learning of parametric face categories leads to the integration of high-level class-based information by not to high-level pop-out. *Journal of Vision, 10*(13): 1-14.

Körding, K.P. and Wolpert, D.M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10*(7): 319-326.

LaBerge, D. (1962). A recruitment theory of simple behaviour. *Psychometrika, 27*(4): 375-396.

Laming, D.R.J. (1968). *Information theory of choice-reaction times*. Oxford, England: Academic Press.

Loewenstein, G. (1992). *Choice over time*. Russel Sage Foundation Publications.

Martin, J.B., Griffiths, T.L., Sanborn, A.N. (in press, 2011). Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories. *Cognitive Science.*

Minka, T. (2006). The Lightspeed Matlab toolbox, efficient operations for. Matlab programming.

Neal, R.M. (1996). *Bayesian learning for neural networks*. New York: Springer-Verlag

Nishimura, M., Maurer, D., Gao, X. (2009). Exploring children's face-space: A \ multidimensional scaling analysis of the mental representation of facial identity. *Journal of Experimental Child Psychology, 103*(3): 355-375

Orbán, G., Fiser, J., Aslin, R.N., Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, 105*(7): 2745.

Paninski, L. (2006). Nonparametric inference of prior probabilities from Bayes-optimal behaviour. In Weiss, Y., Schölkopf, B., & Platt, J. (Ed.), *Advances in Neural Information Processing Systems*, 18: 1067-1074. MIT Press.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. *IEEE Computer Society*, 296-301.

Pinker, S. (1999). How the mind works. *Annals of the New York Academy of Sciences, 882*(1): 119-127.

Rhodes, G., Brennan, S., Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces* 1. *Cognitive Psychology 19*(4): 473-497.

Sanborn, A. and Griffiths, T. (2008). Markov chain Monte Carlo with people. *Advances in Neural Information Processing Systems,* 20: 1265-1272.

Sanborn, A. N., Griffiths, T.L., Shiffrin, R.M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology, 60*(2): 63-106.

Steyvers, M., Griffiths, T.L., Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in cognitive sciences, 10*(7): 327-334.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*(3): 251-260.

Tanaka, J., Giles, M., Kremen, S., Simon, V. (1998). Mapping attractor fields in face space: the atypicality bias in face recognition. *Cognition, 68*(3): 199-219.

Tenenbaum, J. B., Griffiths, T.L., Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7): 309-318.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*(2): 161-204.