Mind Reading by Machine Learning: A Doubly Bayesian Method for Inferring Mental Representations

Ferenc Huszár (fh277@eng.cam.ac.uk)
Computational and Biological Learning Lab, Dept Engineering, U Cambridge, Cambridge CB2 1PZ, UK

Uta Noppeney (uta.noppeney@tuebingen.mpi.de)
Max Planck Institute for Biological Cybernetics, Spemannstrasse 41, Tübingen 72076, Germany

Máté Lengyel (m.lengyel@eng.cam.ac.uk)
Computational and Biological Learning Lab, Dept Engineering, U Cambridge, Cambridge CB2 1PZ, UK

Abstract

A central challenge in cognitive science is to measure and quantify the mental representations humans develop – in other words, to 'read' subject's minds. In order to eliminate potential biases in reporting mental contents due to verbal elaboration, subjects' responses in experiments are often limited to binary decisions or discrete choices that do not require conscious reflection upon their mental contents. However, it is unclear what such impoverished data can tell us about the potential richness and dynamics of subjects' mental representations. To address this problem, we used ideal observer models that formalise choice behaviour as (quasi-)Bayes-optimal, given subjects' representations in long-term memory, acquired through prior learning, and the stimuli currently available to them. Bayesian inversion of such ideal observer models allowed us to infer subjects' mental representation from their choice behaviour in a variety of psychophysical tasks. The inferred mental representations also allowed us to predict future choices of subjects with reasonable accuracy, even in tasks that were different from those in which the representations were estimated. These results demonstrate a significant potential in standard binary decision tasks to recover detailed information about subjects' mental representations.

Introduction

Cognitive science studies the mental representations humans (and other animals) develop and the way these representations are used to perform particular tasks. A central challenge is to measure and quantify such mental representations experimentally – in other words, to 'read' subjects' minds. A classical approach to this is to ask subjects directly to report their mental contents verbally. Unfortunately, this procedure is prone to introducing biases arising from verbal processing, and from the educational and cultural backgrounds of subjects (Ericsson & Simon, 1980; Russo et al., 1989). In order to eliminate these biases, an alternative approach is to limit subjects' responses to simple binary decisions or discrete choices that do not require conscious reflection upon their mental contents. However, it is unclear what such impoverished data can tell us about the potential richness and dynamics of subjects' mental contents.

A powerful computational framework formalises the goal of learning as estimating the probability distribution or density of stimuli (Hinton & Sejnowski, 1986; Dayan & Abbott, 2001). This motivates many formal theories of human learning and cognition to model the relevant mental content of a subject either implicitly or explicitly as a 'subjective' distribution over possible stimuli (Chater et al., 2006; Sanborn & Griffiths, 2008). In this study we adopted this representation, and our goal was to estimate subjects' subjective distributions solely from their responses in simple binary decision tasks without making any assumptions about the process by which those subjective distributions were acquired, i. e. learning.

Ideal observer models are widely used for explaining human behaviour in various psychophysics tasks (Geisler, 2003). They formalise (quasi-)optimal decision making strategies given the information available to subjects and their background knowledge about the task, which in our case includes their subjective distributions. While previous studies mostly used ideal observer models to determine optimal performance in particular tasks to which human performance could then be compared, we treat them as stochastic models formalising the link between subjective distributions (the unobserved variable), and test stimuli and responses (the observed variables). Our main observation is that such models can be used to provide the likelihood in a Bayesian statistical analysis of subjective distributions, thus enabling one to infer mental contents from task responses in a principled way.

We term our approach *doubly Bayesian*, as we assume that subjects act as quasi-ideal observers, which entails Bayesian inference on their side; and then we use these ideal-observer models in a Bayesian framework to infer a posterior distribution of possible subjective distributions.

Inferring subjective distributions

The graphical model (Koller & Friedman, 2009) in Fig. 1A describes our model of a subject's behaviour in a session of a psychophysics experiment. We assume that the subject entertains a subjective distribution $\mathcal{P}$ over possible stimuli, and that this distribution does not change over the analysed session. In trial i of the experiment, the subject is presented a set of test stimuli $\mathcal{S}_i$ and gives a response r_i. The value of r_i depends on the current stimuli $\mathcal{S}_i$, the subjective distribution $\mathcal{P}$, and 'link' parameters $\Theta_{\mathcal{O}}$ describing further aspects of observation and decision making, such as attention, perceptual noise, etc.

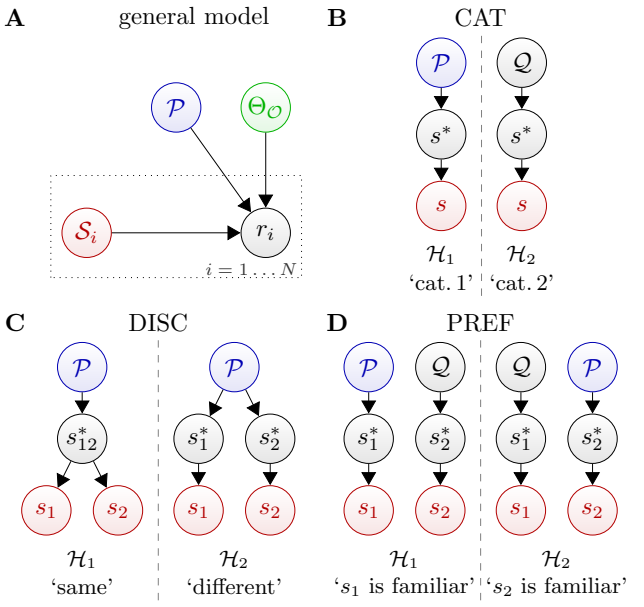

Figure 1: **A,** Graphical model describing the subject's behaviour in an experimental session of N consecutive trials. We assume that the subject represents a subjective distribution, $\mathcal{P}$, over possible stimuli, and in trial i their response r_i depends on the currently observed test stimuli $\mathcal{S}_i$, their subjective distribution, $\mathcal{P}$, and some other parameters influencing their responding, $\Theta_\mathcal{O}$. Our goal is to infer $\mathcal{P}$ and $\Theta_\mathcal{O}$ from the observed sequence of stimulus-response pairs. **B-D,** Generative models for the three task types (CAT, DISC, and PREF, see descriptions under 'Experimental data sets'). Subjects assume that their observations, s, are perceptual noise-corrupted versions of the 'true' stimuli, s^*, sampled by the experimenter from a distribution that is the same as their subjective distribution, $\mathcal{P}$, or an alternative distribution, $\mathcal{Q}$ (which is assumed to be uniform for tractability), depending on the particular hypothesis, $\mathcal{H}$.

In order to quantify the dependence between subjects' choices and their subjective distributions, response probabilities, $p(r_i|\mathcal{S}_i, \mathcal{P}, \Theta_\mathcal{O})$, were specified by quasi-ideal observer models. These models formalise subjects' choices as functions of the posterior probabilities of the two hypotheses corresponding to either response being correct.

Each hypothesis amounts to a different model of how stimuli might have been generated, and so the posterior over hypotheses is inferred by a Bayesian inversion of these generative models. Fig. 1B-D shows such generative models in three tasks considered later in this paper (for more detail, see the supplementary material[1]). Once posterior probabilities are available, the statistically optimal, although psychologically unrealistic, strategy would be to deterministically choose the response with the maximal posterior probability. As a more realistic model of human decision making we used a soft-max function (parametrised by $\Theta_\mathcal{O}$) of log posterior probabilities, that describes quasi-optimal decision making (Sanborn & Griffiths, 2008; Orbán et al., 2008).

Our goal is to estimate latent parameters $\mathcal{P}$ and $\Theta_\mathcal{O}$ from a series of stimulus-response pairs $\{\mathcal{S}_i, r_i\}_{i=1}^N$. As responses given in subsequent trials of the experimental session are assumed to be conditionally independent, the likelihood of latent parameters becomes

$$p(r_{1:N}|\mathcal{S}_{1:N}, \mathcal{P}, \Theta_\mathcal{O}) = \prod_{i=1}^{N} p(r_i|\mathcal{S}_i, \mathcal{P}, \Theta_\mathcal{O})$$

To allow for full Bayesian inference we specified prior distributions over the subjective distribution, $\mathcal{P}$, and link parameters, $\Theta_\mathcal{O}$. We chose to model subjective distributions as mixtures of Gaussians (MoG's). This parametric family of distributions is flexible enough to model complex subjective distributions in low dimensional feature spaces and allows for analytical computation of likelihood ratios in the binary tasks considered here. Importantly, this prior reflected no information about the distribution of stimuli with which subjects were trained (i. e. the distribution to which their subjective distributions could be expected to be close), except for the general domain of possible stimulus values. The MoG representation is not a vital part of our general approach: other representations and priors may be more appropriate in some cases.

Given the prior and the likelihood defined above, we inferred a posterior over $\mathcal{P}$ and $\Theta_\mathcal{O}$ via Bayes' rule:

$$p(\mathcal{P}, \Theta_\mathcal{O}|r_{1:N}, \mathcal{S}_{1:N}) \propto p(\mathcal{P})p(\Theta_\mathcal{O}) \prod_{i=1}^{N} p(r_i|\mathcal{S}_i, \mathcal{P}, \Theta_\mathcal{O})$$

Unfortunately, calculating the posterior exactly is intractable, so we have to resort to approximate inference techniques, for which we implemented a Hamiltonian Monte Carlo algorithm (Neal, 2010).

Experimental data sets

Two experimental data sets were analysed, each collected using simple visual stimuli and requiring binary responses from subjects.

One-dimensional feature space The first set of experimental data was the fish categorisation data set collected by Sanborn & Griffiths (2008). In this experiment, the stimuli used were schematic images of fish of fixed length and variable height, i. e. the relevant feature space was one dimensional (see Fig. 2A). Subjects were trained (with corrective feed-back) in a supervised binary categorisation task (CAT) to distinguish fish drawn from a Gaussian training distribution from fish drawn from a uniform distribution. The mean and variance of the

[1] available online at mlg.eng.cam.ac.uk/ferenc/mindreading

training distribution was varied across four conditions (Fig. 2B, red curves), with 9-11 subjects in each condition. Subjects also performed a stimulus preference task (PREF), in which they had to choose the stimulus which seemed more likely to be drawn from the training distribution. In this task, no feedback was provided. The experiment started with an initial block of 120 CAT trials (to train subjects) followed by four blocks of PREF task alternating with four blocks of CAT task, each block consisting of 60 trials. In a final block of CAT trials, no feedback was provided. For our analysis we neglected the initial training session. We used the next 180 PREF and 180 CAT trials to infer subjects' subjective distributions and reserved the last 60 PREF and 60 CAT trials for cross-validation.

Three-dimensional feature space The stimuli in the second experiment were trapezoids with three features varying systematically: colour (gray-scale), size, and shape (ratio of parallel sides), each parametrised by continuous values between 0 and 1 (Fig. 3A). This experiment involved one-back discrimination (DISC) and stimulus preference tasks (PREF). During DISC trials, which also served to train subjects on a particular distribution of stimuli, subjects were presented with one stimulus per trial, and had to judge (without feedback) whether it was the same or different than the one presented in the previous trial. In actuality, 10% of stimuli were exact repetitions of stimuli presented in the previous trial, the rest was sampled independently from the training distribution. Two different training distributions were used in the two conditions (Fig. 3B, left panels), with six subjects in each condition. During PREF trials subjects had to choose (without feed-back) the stimulus which appeared to be more familiar based on the stimuli they had seen during training. The experiment started with 300 DISC training trials, followed by 100 PREF trials and another 200 DISC trials. In our analysis we neglected the first 100 DISC trials, used 300 DISC and 50 PREF trials to infer subjective distribution and preserved 100 DISC and 50 PREF trials for cross-validation.

Results

Inferring subjective distributions After extensively validating out method on synthetic datasets (supplementary material[1]), we inferred human subjective distributions from the two experiments described earlier. Fig. 2B shows results on the experiment with a one-dimensional feature space. The inferred subjective distributions reflected qualitative aspects of the distributions of stimuli on which subjects were trained in different conditions. This match between inferred and training distributions became especially clear in the categorisation task.

Fig. 3B shows results on the experiment with a three-dimensional feature space. These results suggest that subjects did not learn the training distribution in this experiment very well (see also below), although some resemblance between training and inferred subjective distributions were recovered for a few subjects (e. g. subjects 1, 4, 11 and 12). The subjective distributions inferred for the same subject in the two different tasks also revealed some consistency of these distributions.

Figs. 2-3B illustrate the primary goal of our study: to provide a method for inferring and visualising subjective distributions based on subjects' responding in psychophysics experiments. However, as subjective distributions cannot be observed or measured directly, there is no obvious way to assess the degree to which these inferences are 'correct'. One possibility, pursued above, is to compare the inferred distributions to the distributions subjects were trained on (assuming that subjects are approximately ideal learners and decision makers). While a match between the inferred subjective distribution and the training distribution (Fig. 2B) can be taken as indicative of valid inferences, a lack of match (Fig. 3B) is harder to interpret. In particular, one cannot distinguish between the algorithm giving incorrect results or subjects behaving sub-optimally (because of a failure to learn, or a failure to use learned information to direct choices). Therefore we sought to establish the quality of the inferences of our method in a more reliable way.

Predicting human behaviour A standard way to assess the quality of a statistical model of a data set is to test its predictive performance in cross-validation: infer its parameters (hidden variables) based on a subset of the data, and measure how well it predicts the held-out part of the data set. Our method is readily amenable to this cross-validation approach since it defines an explicit statistical model for predicting subjects' responses based on the stimuli they see (Fig. 1A). Making such predictions is not only important for validation purposes in the context of the present study, but may also be relevant in its own right in applications in which e. g. customer choices need to be predicted based on their previous choices.

For cross-validation, we inferred subjects' subjective distributions and link parameters from the first blocks of trials of a task and based on the inferred model predicted their responses in the final block of trials in the same task (Fig. 4, *double Bayes*). Ideally, subjective distributions are independent of the type of task subjects are performing, and hence one would even expect to be able to infer the subjective distribution from behaviour in one task and, based on that, predict choices in an other task. Thus, we also performed a stronger cross-validation test in which we measured such across-task predictive performance (Fig. 4, *double Bayes-CT*).

Subjects' responding is inherently stochastic, therefore the absolute predictive performance of our model is not particularly informative in itself. In order to establish some relevant baseline performance, we implemented al-

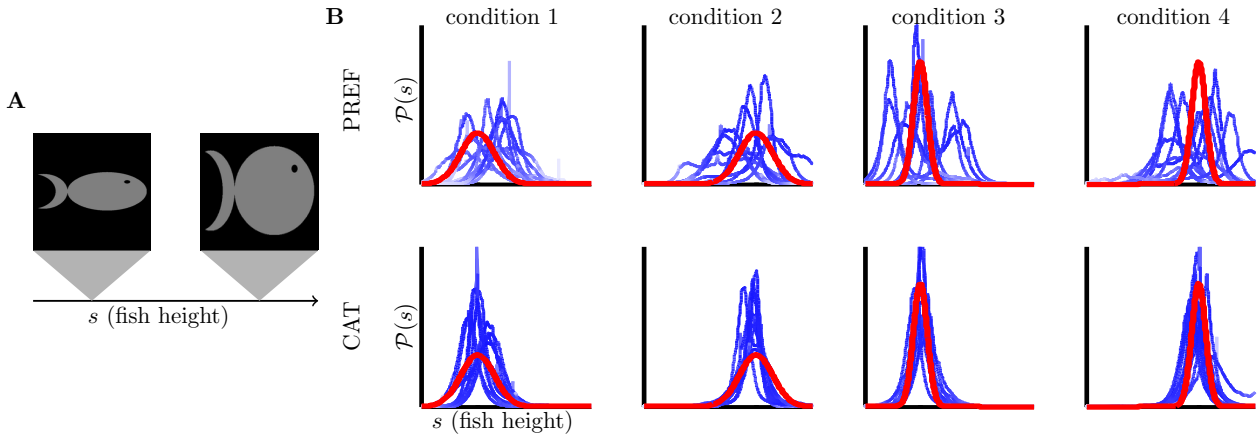

Figure 2: **A,** One dimensional stimuli used in the first set of experiments. **B,** Subjective distributions in a one-dimensional feature space. *Rows* correspond to task types, *columns* correspond to experimental conditions using training distributions with different means (1 & 3 vs. 2 & 4) or variances (1 & 2 vs. 3 & 4). *Red lines* show training distributions, *blue lines* show the posterior mean subjective distribution of each subject. *Shading of blue lines* indicates point-wise marginal posterior uncertainty: lighter means higher uncertainty (s.e.m. divided by the mean).

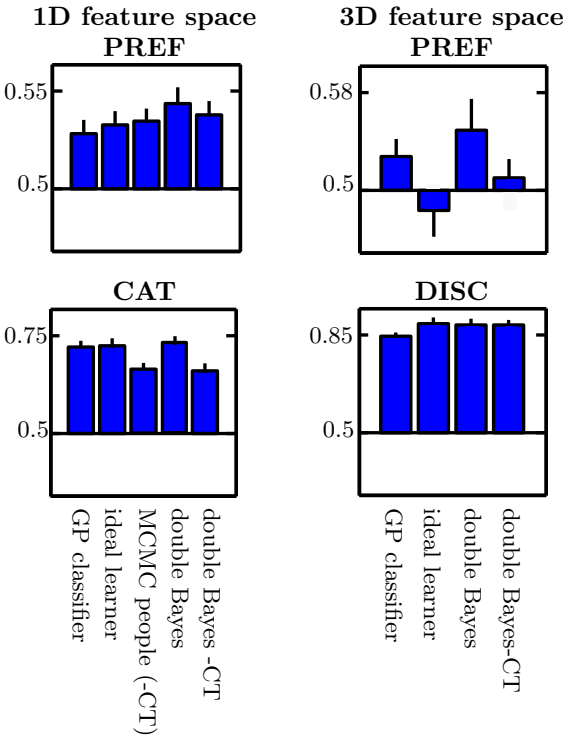

Figure 4: Predicting human responses by alternative methods. *Bars* show across-subject averages (± s.e.m.) of probabilities of correct predictions. In the PREF tasks our method *double Bayes* significantly outperformed both the *GP classifier* and the *ideal learner* in both experiments and also *MCMC people* in the 1D experiment ($p < 0.05$). In the CAT task, the *MCMC people* method was used for across-task predictions (*-CT*).

ternative models for predicting subjects' responses. Since the task of predicting responses based on the stimuli that subjects see is formally equivalent to a binary classification task (see supplementary material[1]), we implemented a Gaussian process classifier (Fig. 4, *GP classifier*) (Rasmussen & Williams, 2006). The GP classifier is a particularly powerful algorithm applicable for such classification tasks, but it is also a black-box model in the sense that it has no explicit notion of subjective distributions. Therefore, it provides an interesting baseline by giving about the best predictive performance that can be achieved without modelling subjects' mental representations.

As an alternative method that did have an explicit notion of subjective distributions, we implemented an 'ideal learner' version of our model, which has the training distribution as its subjective distribution for all subjects, but its link parameters (parametrising stochasticity in decision making) are still fitted to each subject's data individually (Fig. 4, *ideal learner*). This model controls for the importance of individual differences in the inferred subjective distributions in our method, and also tests the validity of the assumption that subjects act as ideal learners in these experiments.

Finally, we also implemented as an alternative method a previously published algorithm ('MCMC with people') to infer subjective distributions (Sanborn & Griffiths, 2008). Although this algorithm can only be applied to specifically designed stimulus preference experiments, one of our data sets includes data from such an experiment, so we tested the performance of the algorithm on that data set by performing both within-task and across-task cross-validation (Fig. 4, *MCMC people (-CT)*).

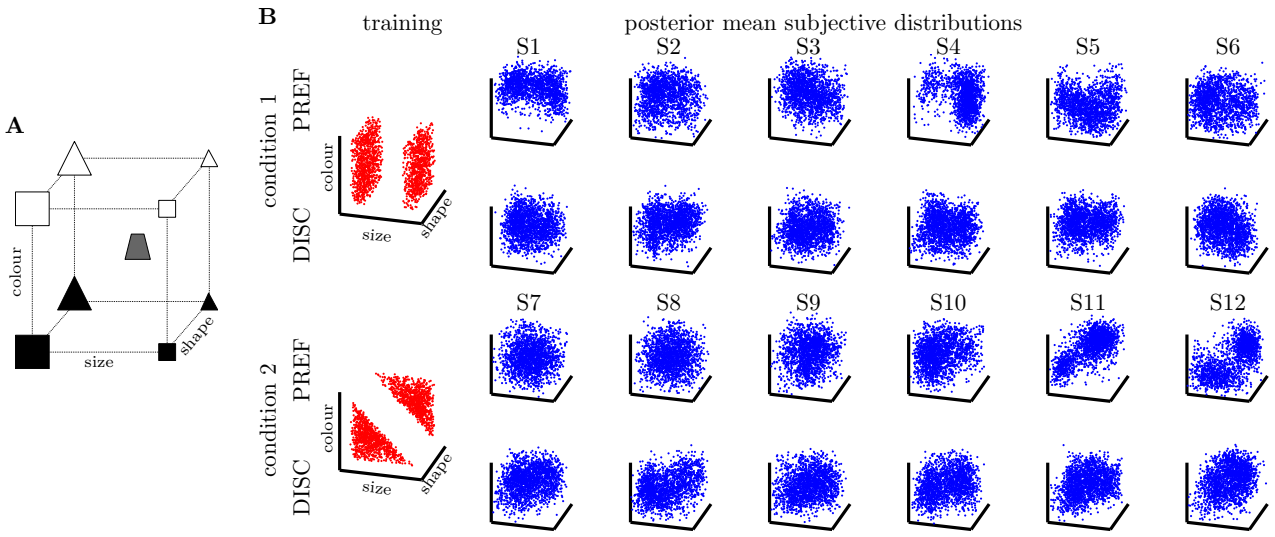

Figure 3: **A,** Stimuli used in the second experiment had three continuous features, size, colour and shape. **B,** Subjective distributions in the three-dimensional feature space. *Left, in red:* training distributions; *right, in blue:* posterior mean subjective distributions *(blue)* for each subject individually *(columns, S1-S12). Rows* correspond to different conditions, using different training distributions, and different task types to test subjects' subjective distributions. In the discrimination task, subjects 7 and 8 responded irrespective of stimuli.

Fig. 4 shows the predictive performances of these methods. The absolute difficulty of predicting responses greatly varied across tasks, the discrimination task and the categorisation tasks being considerably easier than the preference task, but the relative performances of the different methods showed consistent patterns across the different experiments and tasks. When comparing within-task predictive performances, our method was the best, or among the best, in all tasks. Notably, it outperformed the 'MCMC with people' method even in the case when that method was applicable at all.

In most cases, the three subjective distribution-based methods (*double Bayes*, *ideal learner*, *MCMC people*) outperformed the GP classifier, showing that making predictions about subjects' responding benefits substantially from representing and inferring subjective distributions explicitly. This is especially true in across-task cross-validation which is impossible with a GP classifier in lack of any parameter that could be shared between tasks. Yet, in two out of four cases our method had higher accuracy even when comparing its across-task performance against within-task performance of the GP classifier.

Methods using subject-specific subjective distributions (*double Bayes*, *MCMC people*) also performed at least as well as the ideal learner, confirming the validity of the individual differences in subjective distributions these methods inferred, and showing that the poor match found between training and subjective distributions in some cases (Fig. 3B) were real and not a failure of our algorithm to recover 'better looking' subjective distributions.

Discussion

We have presented a new computational method for inferring subjects' mental representation of stimuli from their responses on simple binary decision tasks. Since Bayesian inference was intractable, we implemented a Hamiltonian Markov chain Monte Carlo method for numerical analysis, which we have extensively validated and tested on real-world data sets. We found that the method was able to recover subjective distributions of humans when they were trained on stimuli with known structure and to predict future responses better than other model-based and 'black-box' methods. We have also shown that – using our method – information gained in one type of task could be transferred and applied to predict responses in another task which we take as further evidence for the veridicality and task-invariance of the mental representations we inferred. These results also offer a way to reconcile cognitivism with behaviourism inasmuch as they demonstrate that even when the only goal is to predict responses from stimuli, modelling mental representations explicitly is quantifiably useful.

There is a long tradition in experimental psychology and cognitive science to use simple statistics of task performance, such as percent correct rates, or reaction times, as indices of learning (Gallistel, 1993). These 'naïve' methods, even in their statistically most sophisticated forms (Gallistel et al., 2004; Kakade & Dayan, 2002; Smith et al., 2005; Preminger et al., 2009; Katkov et al., 2007), boil down to estimating a single (time-dependent) scalar measure of memory strength, i. e. the degree of

match between subjects' mental representations and that required by the experimenter (which would presumably allow subjects to perform perfectly). However, by reducing mental contents to simple memory strength measures, these methods fail to provide a detailed picture of structured mental representations which is what we aimed to achieve in the present study.

While structured probabilistic models of cognition have become mainstream more recently (Chater et al., 2006), they have mostly been used in normative theories to account for general, qualitative principles of learning (e. g. patterns of generalisation) rather than to quantitatively estimate individual subjects' mental representations in specific experiments. Our approach is complementary to these as it makes no assumptions about learning itself.

Our work is most closely related to more recent work by Paninski (2006) and Sanborn & Griffiths (2008) who both used ideal observer models to infer subjective distributions. In the paper by Paninski (2006) continuous decision tasks were considered (in which subjects' responses are analogue rather than discrete), and the method developed there does not seem to generalise well to the binary decision tasks considered here (and used extensively in experimental psychology), because the linear programming problem that needs to be solved becomes seriously under-constrained. Our analysis of the preference task is taken from previous work by Sanborn & Griffiths (2008), but they used it to construct a particular kind of stimulus preference task in which subjects' responding itself implements a Markov chain Monte Carlo sampler. This is a most elegant idea, but does not translate in any obvious way to other task types, or indeed to preference tasks which were not constructed according to their particular rules. Our method does not suffer from these limitations because of its doubly Bayesian nature: once ideal observer behaviour based on Bayesian analysis is formalised, the method offers an automatic and principled way of inferring subjective distributions.

A natural way to extend our work in the future will be to consider dynamical priors over subjective distributions in order to track their temporal evolution, inferring changes brought about by learning. The machine learning literature offers powerful tools for carrying out inference in such dynamical models.

Acknowledgements

We thank P. Dayan, C. Rasmussen, and A. Sanborn for useful discussions and K. Jucicaite and S. Ölschläger for help with acquiring psychophysics data. We are grateful for A. Sanborn and T. Griffiths for providing access to their experimental data. This project was supported by the Wellcome Trust (FH, ML), the Gatsby Charitable Foundation (FH), and the Max Planck Society (UN).

References

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends Cogn Sci*, *10*(7), 287–291.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience.* Cambridge, MA: The MIT Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychol Rev*, *87*, 215–251.

Gallistel, C. R. (1993). *The organization of learning.* Cambridge, MA: The MIT Press.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proc Natl Acad Sci USA*, *101*(36), 13124–13131.

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

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In *Parallel distributed processing.* MIT Press.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psych Rev*, *109*(33), 533–544.

Katkov, M., Tsodyks, M., & Sagi, D. (2007). Inverse modeling of human contrast response. *Vision Res*, *47*(22), 2855–67.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques.* MIT Press.

Neal, R. M. (2010). MCMC using hamiltonian dynamics. In S. Brooks *et al.* (Ed.), *Handbook of Markov chain Monte Carlo.* Chapman & Hall / CRC Press.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proc Natl Acad Sci USA*, *105*(7).

Paninski, L. (2006). Nonparametric inference of prior probabilities from Bayes-optimal behavior. In Y. Weiss *et al.* (Ed.), *NIPS 18.* MIT Press.

Preminger, S., Blumenfeld, B., Sagi, D., & Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proc Natl Acad Sci USA*, *106*(13).

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning.* The MIT Press.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, *17*(6), 759–769.

Sanborn, A., & Griffiths, T. (2008). Markov chain Monte Carlo with people. In J. Platt *et al.* (Ed.), *NIPS 20.*

Smith, A. C., Stefani, M. R., Moghaddam, B., & Brown, E. N. (2005). Analysis and design of behavioral experiments to characterize population learning. *J Neurophysiol*, *93*(3), 1776–1792.