

# Mind Reading by Machine Learning: A Doubly Bayesian Method for Inferring Mental Representations – supplementary material –

**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

## Quasi-ideal observer models

In order to be able to make inferences about subjective distributions based on subjects’ responses we need to construct a model of the process of generating these responses based on the information available to subjects and parameters that describe their behaviour (see Fig.1A in the main paper). Throughout the paper, we assume that the response variables  $r_i$  are discrete, although the methodology naturally extends to the case of continuous responses.

The first step of constructing an ideal observer model involves a Bayesian analysis of the task faced by subjects. A subject may entertain several alternative hypotheses, each corresponding to one of the  $R$  different available responses. The likelihood of hypothesis  $\mathcal{H}_j$ ,  $p(\mathcal{S}_i|\mathcal{H}_j, \mathcal{P}, \Theta_{\mathcal{O}})$ , expresses the probability of seeing stimuli  $\mathcal{S}_i$  in trial  $i$  in which response  $j$  is correct. Although the precise form of this likelihood depends on the particular task, it always depends on the subjective distribution  $\mathcal{P}$ , and this dependence will allow us to invert the model and infer  $\mathcal{P}$  from behavioural data. Importantly, as also noted in the paper, we introduce a number of parameters, captured in  $\Theta_{\mathcal{O}}$ , to model different sources of suboptimal behaviour so that the resulting quasi-ideal observer models will be appropriate statistical models of subjects’ responding.

Given the likelihood of hypothesis  $j$ , its posterior probability can be obtained by applying Bayes’ rule:

$$\mathbb{P}[\mathcal{H}_j|\mathcal{S}_i, \mathcal{P}] = \frac{p(\mathcal{S}_i|\mathcal{H}_j, \mathcal{P}) \cdot \pi_j}{\sum_{k=1}^R p(\mathcal{S}_i|\mathcal{H}_k, \mathcal{P}) \cdot \pi_k} \quad (\text{S1})$$

where  $\mathbb{P}[\mathcal{H}_j] = \pi_j$  describes the subject’s prior preferences over hypotheses. The first source of sub-optimality we introduce is that, unlike in most ideal observer analyses, we assume that these prior probabilities are truly subjective and thus they need not reflect the veridical fraction of trial types in which their corresponding responses are correct.

An ideal observer bases its decisions on the posterior

probabilities of the hypotheses it is entertaining. Therefore, the second key modelling step is to specify the precise functional dependence between the probabilities with which different options are chosen and the posterior probabilities of the hypotheses associated with them. We apply the flexible and commonly used (Sanborn & Griffiths, 2008; Orbán et al., 2008) soft-max rule:

$$\mathbb{P}_{\text{softmax}}[j] = \frac{\mathbb{P}[\mathcal{H}_j|\mathcal{S}_i, \mathcal{P}]^{\beta}}{\sum_{k=1}^R \mathbb{P}[\mathcal{H}_k|\mathcal{S}_i, \mathcal{P}]^{\beta}} \quad (\text{S2})$$

The rule is parameterised by the decision noise parameter  $\beta$ . We note that the soft-max rule contains probability matching (Vulkan, 2000) ( $\beta = 1$ ) and maximum *a posteriori* decisions ( $\beta \rightarrow \infty$ ) as special cases. For binary tasks ( $R = 2$ ,  $\pi_1 = \pi$ ,  $\pi_2 = 1 - \pi$ ) considered in the paper this rule becomes a simple logistic sigmoidal function of the log-likelihood ratio:

$$\mathbb{P}_{\text{softmax}}[1] = 1 - \mathbb{P}_{\text{softmax}}[2] = \frac{1}{1 + e^{-\beta \log \frac{\pi}{1-\pi} R}} \quad (\text{S3})$$

$$R = \frac{p(\mathcal{S}_i|\mathcal{H}_1, \mathcal{P})}{p(\mathcal{S}_i|\mathcal{H}_2, \mathcal{P})} \quad (\text{S4})$$

It is also possible, and desirable, to model other sources of sub-optimality in responding. For instance, a subject may misunderstand the instructions or become bored and start pressing the same button irrespective of the stimuli. To account for this, we apply a probabilistic mixture of an ideal and a ‘misbehaved’ subject, parameterised by the ‘attention’ parameter  $\kappa$ . Actual response probabilities thus take the following form:

$$\mathbb{P}[r_i = j|\mathcal{S}_i, \mathcal{P}, \Theta_{\mathcal{O}}] = \frac{1 - \kappa}{2} + \kappa \mathbb{P}_{\text{softmax}}[j],$$

A final source of stochasticity we consider in subjects’ behaviour is observation noise (or, more precisely, the subject’s internal model of her observation noise), describing the way ‘true’ stimulus feature values,  $s^*$ , controlled by the experimenter are distorted in the subject’s perception

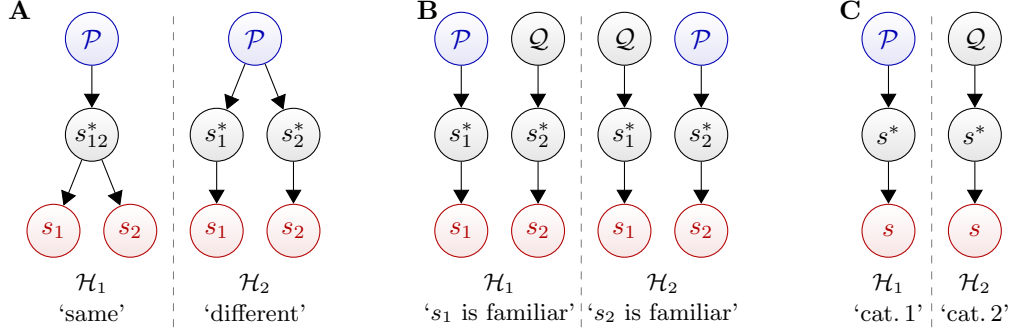


Figure S1: Graphical models describing alternative hypotheses in the psychophysical tasks considered in the paper (see also Fig. 1B-D in main paper). **A**, Alternative hypotheses in a trial of the discrimination task. Under  $\mathcal{H}_1$  (left), the two observed stimuli,  $s_1$  and  $s_2$  are noisy observations of the same underlying stimulus  $s_{12}^*$  that was drawn from  $\mathcal{P}$ , whereas  $\mathcal{H}_2$  (right) holds that  $s_1$  and  $s_2$  are noisy observations of two different stimuli  $s_1^*$  and  $s_2^*$ , respectively, that are independent draws from  $\mathcal{P}$ . **B**, Alternative hypotheses in a trial of the preference task. The subject is presented with two stimuli  $s_1$  and  $s_2$ . Under  $\mathcal{H}_1$  (left),  $s_1$  is familiar, i.e. it is a noisy observation of a stimulus drawn from  $\mathcal{P}$ , whereas  $s_2$  is a noisy observation of another stimulus drawn from some alternative distribution  $\mathcal{Q}$ . Under  $\mathcal{H}_2$  (right),  $s_2$  is the familiar stimulus. **C**, Alternative hypotheses in a trial of the categorisation task. The subject is presented with stimulus  $s$ . Under  $\mathcal{H}_1$  (left),  $s$  is a noisy observation from category-conditional distribution  $\mathcal{P}$ . Under  $\mathcal{H}_2$  (right),  $s$  is a noisy observation from  $\mathcal{Q}$ .

to yield observed features,  $s$ . We model observation noise as a Gaussian distribution around the true stimulus value with constant (i.e. stimulus-independent) covariance  $\Sigma_n$ :

$$p_n(s|s^*) = \mathcal{N}(s; s^*, \Sigma_n) \quad (\text{S5})$$

Of course, it is possible to incorporate more complex observation noise models when it is particularly required. A possible extension would be to allow the covariance of the Gaussian noise to be stimulus-dependent.

In summary, the resulting quasi-ideal observer models have the following free parameters,  $\Theta_{\mathcal{O}}$ : prior bias  $\pi$ , decision noise  $\beta$ , level of attention  $\kappa$ , and observation noise covariance matrix  $\Sigma_n$ . Importantly, these parameters are estimated from data rather than being fixed at some pre-defined values. Hyper-parameters for priors over them are defined below.

In the following we derive likelihood ratios,  $R$ , that are needed to compute binary choice probabilities based on eq. S3 for the three binary task types considered in the paper.

### One-back discrimination task

In a trial of a discrimination task (DISC) the subject has to decide whether two stimuli presented to them are identical or not. In the experiments presented in this paper, a special case of the discrimination task, the so-called one-back discrimination task was used, in which stimuli are presented sequentially, and the subject has to discriminate the current stimulus from the previously presented one. However, our ideal observer analysis generalises to any discrimination task.

Our construction of the ideal observer model for the discrimination task relies on the Bayesian analysis of similarity judgements put forward by Kemp et al. (2005). According to this analysis, the degree of subjective similarity of two objects is related to the probability of them having been generated by the same generative process. In our framework, the subject's belief about the process generating possible objects is captured by the subjective distribution  $\mathcal{P}$ . Based on this, we can derive the likelihood ratio of the two hypotheses (see Fig. S1A for further explanation):

$$R = \frac{\int p_n(s_1|s_{12}^*) p_n(s_2|s_{12}^*) \mathcal{P}(s_{12}^*) ds_{12}^*}{\int p_n(s_1|s_1^*) \mathcal{P}(s_1^*) ds_1^* \cdot \int p_n(s_2|s_2^*) \mathcal{P}(s_2^*) ds_2^*} \quad (\text{S6})$$

If the subjective distribution is a mixture of Gaussians, and the observation noise is Gaussian with constant covariance matrix  $\Sigma_n$ , then this likelihood ratio can be computed analytically.

### Stimulus preference task

In each trial of the stimulus preference task (PREF)<sup>1</sup>, two stimuli are presented to the subject, who then needs to choose the one that appears more familiar, or the one that appears more to be a member of a particular category.

The two hypotheses that the subject has to choose between are explained in Figure S1B. The likelihood ratio for this task is:

$$R = \frac{\int p_n(s_1|s_1^*) \mathcal{P}(s_1^*) ds_1^* \cdot \int p_n(s_2|s_2^*) \mathcal{Q}(s_2^*) ds_2^*}{\int p_n(s_1|s_1^*) \mathcal{Q}(s_1^*) ds_1^* \cdot \int p_n(s_2|s_2^*) \mathcal{P}(s_2^*) ds_2^*}, \quad (\text{S7})$$

<sup>1</sup>this task is often referred to as the two-alternative forced choice (2AFC) task

where  $\mathcal{Q}$  is the distribution of *non-familiar* objects. For convenience, we assume that the alternative distribution  $\mathcal{Q}$  is approximately uniform, in which case the response probabilities will not depend on  $\mathcal{Q}$ , and reduce to standard formulæ also used in Sanborn & Griffiths (2008); Orbán et al. (2008). Whether this assumption is reasonable, depends largely on experimental conditions and on the exact phrasing of the instructions given to subjects. In principle, one could treat  $\mathcal{Q}$  as another parameter of the ideal observer and infer it in the same way as we infer  $\mathcal{P}$ .

### Binary categorisation task

In each trial of the binary categorisation task (CAT), the subject is presented with a single stimulus, and has to decide whether it belongs to category 1 or 2. For this task, we assume the subject entertains category-conditional subjective distributions  $\mathcal{P}$  and  $\mathcal{Q}$  over stimuli. Then, the likelihood ratio of hypotheses takes the following simple form (see Figure S1C):

$$R = \frac{\int p_n(s|s^*) \mathcal{P}(s^*) ds^*}{\int p_n(s|s^*) \mathcal{Q}(s^*) ds^*} \quad (\text{S8})$$

In the experiments with one-dimensional feature space, we consider a special case, when the subject is instructed that  $\mathcal{Q}$  is uniform, i. e. every stimulus is equally probable to be an exemplar of category 2. We also assume that – prior to the analysed part of the experiment – the subject had already learned the support and magnitude of the uniform distribution, and we can thus concentrate on inferring  $\mathcal{P}$  only.

### Prior distributions

As part of Bayesian analysis, we have to define prior distributions over the unobserved variables, i. e. the subjective distribution  $\mathcal{P}$  and link parameters  $\Theta_{\mathcal{O}}$ . As explained in the paper, we chose to represent subjective distributions as mixtures of Gaussians (MoGs), that are parametrised by the number of components  $K$ , the component means  $\mu_k, k = 1 \dots K$  and covariances  $\Sigma_k, k = 1 \dots K$ . We refer to this set of parameters as  $\Theta_{\mathcal{P}}$ . The link parameters  $\Theta_{\mathcal{O}}$  include  $\pi, \beta, \kappa$  and  $\Sigma_n$ , as described in section . For ease of computation of derivatives, we chose simple exponential family distributions as priors over parameters  $\Theta_{\mathcal{P}}$  and  $\Theta_{\mathcal{O}}$ .

$\Theta_{\mathcal{P}}$ : In the one-dimensional experiments fish height (the only relevant stimulus feature) ranged from 0.3 to 8.35 cm. Accordingly, we set the prior over component means to be Gaussian with mean 4.16 cm and variance 1 cm<sup>2</sup>. Component covariances were given a Wishart prior with mean 0.1 cm<sup>2</sup> and concentration 4. This way, training distributions corresponding to different conditions were approximately equally probable under the prior. We set the number of components to be 6.

In the second experiment with three-dimensional feature space, the stimulus features were normalised so that they took values between 0 and 1. To enforce subjective distributions to be concentrated in this effective stimulus set, we set the prior over component means to be Gaussian with mean  $[0.5, 0.5, 0.5]^T$  and covariance  $0.4\mathbf{I}$ . We put a Wishart prior over component covariances with mean  $0.1\mathbf{I}$  and concentration 7. The prior average Kullback-Leibler divergences from the two training distributions were approximately equal. We set the number of components to be 8.

$\Theta_{\mathcal{O}}$ : We used weak priors: Gamma with mean 1, concentration 3 for decision noise  $\beta$ ; Beta with mean 0.5 concentration 3 for the prior bias  $\pi$ ; Beta with mean 0.9 and concentration 3 for attention  $\kappa$ ; and Wisharts with means 0.02 and 0.02 $\mathbf{I}$ , and concentrations 2 and 4 for the observation noise covariances  $\Sigma_n$  in the two experiments, respectively.

### Results on synthetic data

We have extensively validated our method on simulated data. For this, we fixed the subjective distribution  $\mathcal{P}$  (or, more precisely, its parameters,  $\Theta_{\mathcal{P}}$ ) and link parameters  $\Theta_{\mathcal{O}}$ , and simulated responses of the corresponding quasi-ideal observer to a given set of test stimuli. We then used our Hamiltonian Monte Carlo algorithm to reconstruct the values of hidden variables  $\mathcal{P}$  (or  $\Theta_{\mathcal{P}}$ ) and  $\Theta_{\mathcal{O}}$ . We found that our method successfully estimated the true underlying parameters in most cases (see Fig. S2). However, due to known non-identifiability of our ideal observer-based likelihood (Sanborn & Griffiths, 2008), parameter  $\beta$  cannot be reliably inferred from PREF trials (Fig. S3).

We also investigated the effect of the distribution of test stimuli on our algorithm. This is important because the distribution of test stimuli used in experiments often resembled the training distribution to what we compared our estimated subjective distributions (expecting subjects' subjective distributions to reflect the distribution of stimuli on which they had been trained). While the choice of test stimuli substantially affected posterior uncertainty, it had almost no effect on the mean of the posterior, that is our results were not biased by the choice of test stimuli (Fig. S4). Thus, any match found in experimental data between training distributions and estimated (mean) subjective distributions could be taken as real and were not artifactual consequences of the overlap between training and test distributions.

### Predicting human responses

The problem of predicting human responses in simple binary decisions tasks can be interpreted as a problem of binary categorisation (mapping the feature values of a test stimulus, or of a pair of test stimuli, to one of two possible

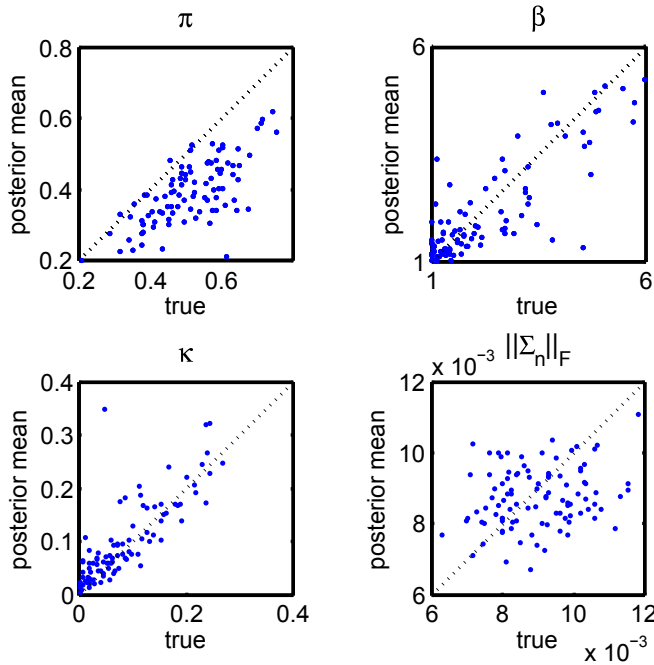


Figure S2: Scatter plots visualising the reconstruction of true model parameters  $\Theta_{\mathcal{O}}$  from simulated datasets in the discrimination task. The panels correspond to model parameters  $\pi$ ,  $\beta$  and  $\kappa$ , and  $\Sigma_n$  ( $\|\bullet\|_F$  denotes Frobenius norm). In the panels, posterior mean estimates are plotted against the true parameter values for 100 random experiments. We found significant positive correlations between true and estimated values ( $R_{\pi} = 0.72, p = 3.44 \times 10^{-17}$ ;  $R_{\beta} = 0.78, p = 5.99 \times 10^{-22}$ ;  $R_{\kappa} = 0.83, p = 8.62 \times 10^{-27}$ ;  $R_{\Sigma_n} = 0.25, p = 0.0099$ ).

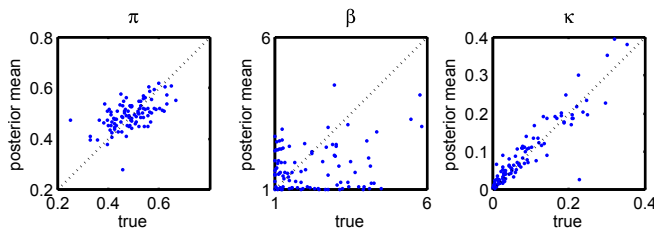


Figure S3: Scatter plots visualising the reconstruction of true model parameters  $\Theta_{\mathcal{O}}$  from simulated datasets in the preference task. The estimates of  $\pi$  and  $\kappa$  correlate significantly with the true values ( $R_{\pi} = 0.75, p = 2.51 \times 10^{-15}$ ,  $R_{\kappa} = 0.92, p = 9.44 \times 10^{-28}$ ), but  $\beta$  is estimated poorly ( $R_{\beta} = 0.17, p = 0.099$ ), which is due to known non-identifiability of this parameter under our likelihood model.

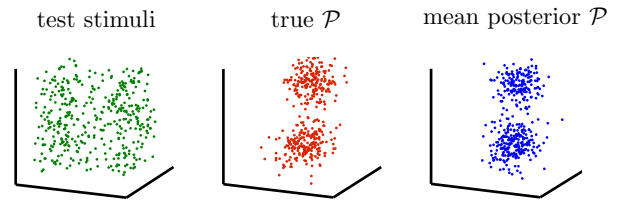


Figure S4: Illustration of the effects of test stimuli: We used the ideal observer model with a fixed ‘true’ subjective distribution (*middle*) to simulate  $n = 200$  responses to the stimuli that was presented to subject 1 in the 3-D experiment (*left panel*). We then inferred the subjective distribution (*right*) from the stimulus-response pairs obtained this way. We found that posterior resembled the ‘true’ subjective distribution irrespective of how the stimuli were chosen.

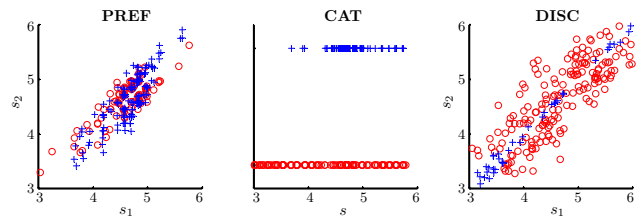


Figure S5: Prediction of human responses viewed as a binary classification task. *Left*: in the preference task, the subject is presented a pair of stimuli, and selects either  $s_1$  (*blue crosses*) or  $s_2$  (*red circles*) as being more familiar (human responses from the one-dimensional experiment). *Middle*: in the categorisation task, the subject is presented a single stimulus and classifies it as belonging to either category 1 (*blue crosses*) or 2 (*red circles*) (human responses from the one-dimensional experiment). *Right*: in the discrimination task, a pair of stimuli is presented in succession, and the subject decides whether they seem to be different (*blue crosses*) or identical (*red circles*) (simulated data).

choices). Figure S5 shows the categorisation problems inherent in the three psychophysics tasks considered in the paper.

## References

- Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the twenty-seventh annual conference of the cognitive science society*.
- 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).
- Sanborn, A., & Griffiths, T. (2008). Markov chain Monte Carlo with people. In J. Platt *et al.* (Ed.), *NIPS 20*.
- Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118.