Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies

Payam Piray[1,*], Amir Dezfouli[2], Tom Heskes[3], Michael J. Frank[4] and Nathaniel D. Daw[1]

1. Princeton Neuroscience Institute, Princeton University
2. University of New South Wales Sydney
3. Institute for Computing and Information Sciences, Radboud University
4. Department of Cognitive, Linguistics, and Psychological Sciences, Brown University

* Corresponding author: Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA. Email: ppiray@princeton.edu.

Abstract

Computational modeling plays an important role in modern neuroscience research. Much previous research has relied on statistical methods, separately, to address two problems that are actually interdependent. First, given a particular computational model, Bayesian hierarchical techniques have been used to estimate individual variation in parameters over a population of subjects, leveraging their population-level distributions. Second, candidate models are themselves compared, and individual variation in the expressed model estimated, according to the fits of the models to each subject. The interdependence between these two problems arises because the relevant population for estimating parameters of a model depends on which other subjects express the model. Here, we propose a hierarchical Bayesian inference (HBI) framework for concurrent model comparison, parameter estimation and inference at the population level, combining previous approaches. We show that this framework has important advantages for both parameter estimation and model comparison theoretically and experimentally. The parameters estimated by the HBI show smaller errors compared to other methods. Model comparison by HBI is robust against outliers and is not biased towards overly simplistic models. Furthermore, the fully Bayesian approach of HBI enables researchers to quantify uncertainty in group parameter estimates, for each candidate model separately, and to perform statistical tests on parameters of a population.

Introduction

Across different areas of neuroscience, researchers increasingly employ computational models for experimental data analysis. For example, decision neuroscientists use reinforcement learning and economic models of choice to analyze behavioral and brain imaging data in reward learning and decision-making tasks [1,2]. The field of computational psychiatry uses these models to characterize patients and people at the risk of brain disorders [3–6]. Neuroimaging studies use models of neural interaction, such as dynamic causal modeling [7,8], as well as abstract models to analyze brain signals [1,9]. The success of these efforts heavily depends on statistical methods making inference about validity and robustness of estimated parameters across individuals, as well as making inference on validity and generalizability of computational models. A key theoretical and practical issue has been capturing individual variation both in a model's parameters and additionally in which of several candidate models a subject expresses, which may also vary from subject to subject.

Computational models usually rely on free parameters, for example learning rate in reinforcement learning models, which often capture quantities of scientific interest but typically vary across individuals and must be estimated from data. A dataset includes a number of subjects, and often the question of interest is to characterize parameters in a population: Is choice consistency altered in patients with attention-deficit hyperactive disorders? Do cognitive enhancers, such as Ritalin, enhance learning rate at the population level? These questions are most naturally framed in terms of hierarchical models, which characterize both the population distributions over a model's parameters, and also each individual subject's parameters given the population distribution. Since these two levels are mutually interrelated, they are often estimated simultaneously, using methods like expectation maximization or sampling (MCMC). For example, the hierarchical parameter estimation (HPE) procedure [10,11] regularizes individual estimates according to group statistics, producing better individual estimates and permitting reliable group-level tests. Because subjects typically share underlying structure, hierarchical Bayesian approaches can leverage this this structure to yield better individual estimates, and to provide better predictions for unseen data, compared to approaches that fit each subject separately [12].

A second, and seemingly logically prior, question is which of several candidate models provides the best explanation for the data. This is important both for providing the setting within which to do parameter estimation, and also for investigating questions of scientific interest. Are rodents' reaction times best explained by independent or competing accumulators? Do compulsive gamblers rely more on model-free reinforcement learning compared to controls? Importantly, in principle (and apparently in practice) the model

expressed might also vary from subject to subject; thus modern model comparison techniques rely on estimating which of several models obtains for each subject [13]. Estimating such variation is important since by assuming that the same model obtains across all individuals (treating model identity as fixed effect) inflates significance for model comparison and makes it sensitive to outliers [13]. To estimate this variation, in turn, depends on the likelihood of each subject's data given each model (and, thus, on each subject's parameters for each model).

Intuitively, evaluating whether a model is a good model for a subject's data precedes estimation of its specific parameter values; and indeed previous research has used separate tools to solve these two problems. But statistically, the two questions are actually interconnected, because individual parameters and hence individual fit depend on which subjects belong to the population that expresses the model. Here, we address this challenge from a fully Bayesian viewpoint. This work addresses issues of statistical inference over both parameters and models, which have remained elusive with the previous hierarchical methods.

Notably, although it is accepted (for the reasons discussed above) that the best-fitting model may vary from subject to subject, hierarchical parameter estimation (conducted separately) has typically assumed that the given model is expressed over all subjects, i.e. that it is a fixed effect. (And if multiple models are compared, these are each fit to the entire population.) This assumption biases parameter estimation, at both individual and group levels, because it entails that the estimated parameters for each individual subject *equally* affect group-level estimates, even though some members of the population may be better understood as expressing altogether different models. This same bias, in turn, affects the estimation of which subjects are best fit by each model.

In this work, we introduce a hierarchical and Bayesian inference method, which solves these problems by addressing both model fitting and model comparison within the same framework using variational techniques. Furthermore, our fully Bayesian approach enables us to assess uncertainty and provide a rigorous statistical test for making inference about parameters of a model at the population level, an issue that has been incompletely addressed in some previous hierarchical models. This paper is structured as follows. First, we highlight the main theoretical advances of our approach. A full formal treatment is given in the appendix. We then apply the proposed method to synthetic choice datasets as well as an empirical dataset to demonstrate its advantages over previous methods.

Theory

Consider a typical computational modeling study in which data of a group of subjects have been measured and a set of candidate models are considered as possible underlying computational mechanisms generating those data. Such studies have generally two main goals: 1) to compare model evidence across competing models; 2) to estimate free parameters of models for each individual and their group-level distributions. All this is typically characterized in terms of inference in a hierarchically structured model of the data, which captures how each subject's observations depend on their parameters, and the individual parameters on their group distribution.

The HPE procedure [10,11] employs a hierarchical approach to define the priors based on statistics of the group. This method typically assumes that for a particular model k, all individual parameters are normally distributed,

$$p(h_{kn}) = Normal(h_{kn}|\mu_k, V_k)$$

where h_{kn} is a vector of the free parameters of kth model for subject n, μ_k and V_k are the mean and variance parameters, respectively, indicating the prior distribution over h_{kn}.

HPE uses the expectation-maximization algorithm [14], a well-known iterative procedure, for obtaining estimating group parameters μ_k and V_k and individual parameters h_{kn}. Every iteration of this algorithm alternates two steps: 1) an expectation step in which the individual parameters are estimated in light of the group-level distribution; and 2) a maximization step in which the group parameters, μ_k and V_k, are updated given the current estimates of the individual parameters. Importantly, this update weights the individual subjects' estimates equally; for instance, the update for μ_k is given by the average of subject level mean estimates (denoted θ_{kn}) across all subjects:

$$\mu_k = \frac{1}{N}\sum_n \theta_{kn}$$

where N is the number of subjects.

Although HPE characterizes variation across subjects in the model parameters h_{kn} (that is, it treats those parameters as random effects), a critical assumption of the procedure is that the parameters for model k are estimated assuming that same model is responsible for generating data in all subjects. That is, the model identity is taken as a fixed effect, in contrast to random effects approach that assumes different models might be responsible for generating data in different subjects. The fixed effects assumption has two important implications: 1) for

parameter estimation, group parameters, the group mean μ_k and variance V_k, are influenced equally by all subjects, even those who would be better fit by some other candidate model $j \neq k$; 2) for model comparison, the straightforward procedure (e.g. iBIC from [10,11]) is to compare models according to the sum of individual model evidences over all subjects, i.e. again treating the model identity as a fixed effect. Note that while it is possible to submit individual model evidence values (per subject and model) derived from HPE to a separate model comparison procedure that treats model identity as a random effect (such as random effects model selection [13]), these will be biased both from having been fit under the fixed effects assumption and also due to the optimization of the free group-level parameters. Altogether, violations of the fixed effects assumption can adversely influence both parameter estimation and model comparison.

Here, we extend HPE's generative model with another level of the hierarchy, specifying for each subject which model generated their data. This is governed by a subject-specific multinomial random variable, itself drawn from a distribution controlling the proportion of each model in the population. This, in effect, merges the Bayesian model selection model from Stephan et al. [13] with HPE. We then lay out a procedure for joint inference over model identities and parameters, including quantifying the probability that each model is responsible for generating data for each subject. To achieve this goal, we take a full Bayesian approach in which the group parameters for each model, μ_k and V_k, are also random variables. This also gives us a straightforward way to quantify the level of certainty in group-level estimations. We use variational Bayes [15,16], an extended version of expectation-maximization [17], which is able to deal with multiple latent variables in a probabilistic model. Since HBI is a variational framework, the resulting algorithm (Appendix A.3) is an iterative algorithm. On every iteration, the HBI performs 4 steps: calculates the summary statistics, updates its estimates of the posterior over group parameters, updates its estimate of the posterior over each individual parameter and finally updates its estimates of responsibility of each model in generating each individual data. The algorithm and other important mathematical issues are given in appendix A. Here, we highlight three main results. The mathematical proofs are given in appendix B.

As noted above, the HBI method estimates the probability of each subject's dataset being generated by each model, or the *responsibility* of model k for generating data for subject n, r_{kn}, which is expressed as (expected) probability. Larger values of r_{kn} (i.e. close to 1) indicate that model k is likely to be the true underlying model of the nth subject. In contrast, smaller values of r_{kn} (close to 0) indicate that model k is unlikely to be the underlying model

for the *n*th subject. Based on the responsibilities, it is then possible to estimate number of subjects explained by each model, $\overline{N_k}$:

$$\overline{N_k} = \sum_{n=1}^{N} r_{kn}$$

Thus $\overline{N_k}$ is always less than number of subjects and indexes the predominance of model k in the population. Furthermore, the fraction $\overline{N_k}/N$ always lies between 0 and 1 and is a useful and intuitive metric for model comparison.

In practice, in many situations, researchers are interested in selecting a single "best" model (rather than relative comparisons among several) even in the face of variation in model identity across subjects. One way to accomplish this goal is to use $\overline{N_k}$ to compute the exceedance probability of each candidate model, a metric commonly used for model selection [13]. Exceedance probability is the probability that model k is more commonly expressed than any other model in the model space. Furthermore, the random effects approach enables us to quantify how likely the observed differences in model evidence is simply due to chance [18]. In this case, model selection is not statistically supported (although model comparison is valid). A metric called protected exceedance probability [18], which typically is more conservative than the exceedance probability, takes into account this possibility (see Appendix A.7). Altogether, the random effects approach results in a more robust model comparison and model selection, one less driven by outliers than fixed-effects methods. Note that previous attempts to do model selection at group level using exceedance probability assumed no hierarchy for parameter estimation, thus did not deal with the issue that parameter estimation was not properly conditionalized by group distributions based on model identity.

We noted above that an issue with the HPE is that the influence of subjects on the group parameters is equal. However, the comparable parameter in our approach, the mean of posterior distribution over μ_k, denoted by a_k, shows an important property: A subject's effect on this parameter depends on the degree to which the model is estimated to be the underlying model for that subject. Specifically, this parameter, a_k, is updated at each iteration as:

$$a_k = \frac{1}{1 + \overline{N_k}}(a_{0k} + \sum_{n} r_{kn}\theta_{kn})$$

where θ_{kn} is the mean of the individual posterior and a_{0k} is the prior mean over μ_k. The important point in this equation is that a_k is a *weighted* average of individual parameters, in which the weights are the corresponding responsibilities, r_{kn}. This is not specific to the group mean, but it is rather a general feature of our approach: contribution of model k to group

parameters is weighted according to the responsibility of model k in generating data in nth subject, r_{kn}.

As mentioned above, another issue that has been incompletely treated in HPE is related to inference on parameters of a fitted model at the population level. Statistically, one needs the uncertainty of the estimated group mean, μ_k, to be able to make inference on the corresponding parameter at the group level. Since parameters fitted by the hierarchical parameter estimation method are not independent but instead regularized according to the variance given by data, one cannot employ regular statistical tests, such as t-test, to test whether a specific model parameter is "significantly" different from zero. Using those tests on such parameters is biased in favor of generating a significant p-value (more false positives). The HBI framework solves this problem by quantifying uncertainty of the posterior over the group parameter. Specifically, it is possible to show that the posterior over the group parameter, μ_k, takes the form of standard Student's t-distribution centered at a_k with $n_k = 1 + \overline{N_k}$ as degrees of freedom. The resulting t-value takes an intuitive form:

$$t = \frac{\mu_k - a_k}{s_k/\sqrt{n_k}}$$

where s_k is the empirical deviance statistics. Therefore, $s_k/\sqrt{n_k}$ plays the role of standard error, which we call it *hierarchical error*. Note that the degrees of freedom of the test depend on the number of subjects (i.e. evidence) in favor of model k given by $\overline{N_k}$, not the total number of subjects. Other group statistics, a_k and s_k, are also weighted according to the responsibilities of model k in generating data of each subject (as formally obtained in Appendix A4).

Results

In this section, we apply the proposed HBI method to synthetic and empirical datasets and compare its performance with that of HPE, as well as with Laplace approximation procedure, a non-hierarchical inference (NHI) method estimating parameters for each subject independently according to some fixed, *a priori* priors [19–22]. The HBI is general and could be applied to any type of data, such as choice data, reaction times, physiological signals and neural data. Since we are primarily interested in models of choice data, we focus on decision-making experiments. The details of simulations are given in the Supplementary Materials.

Model comparison and parameter estimation

First, we simulated a dataset including 40 artificial datasets using two different learning models and a randomly generated reward sequence (binarized Gaussian random-walk). Both models maintain a value for each of the two possible actions and calculate a prediction error signal, δ, representing the difference between the seen reward and predicted value. On every trial, the action value gets updated according to the product of δ and a learning rate. The first model is a reinforcement learning model, in which the learning rate is a constant free parameter, α. The second model is a Kalman filter model in which the learning rate gradually decreases on every trial. The decreasing rate depends on a positive free parameter (representing observational noise), ω. Both models employ a softmax function together with an inverse-temperature parameter, β, to calculate the probability of each action according to corresponding expected values. Therefore, both models contain three free parameters and neither of them is nested within the other one.

The reinforcement learning and Kalman filter models were then used to simulate 10 and 30 artificial datasets, respectively. Parameters of these models were drawn randomly from normal distributions. Since parameters of these models have theoretical constrains, we used appropriate functions (sigmoid or exponential) to transform these randomly generated parameters. Using this procedure, we constructed a dataset of 40 artificial subjects, in which the true underlying model is known. We applied the HBI to this dataset to estimate parameters and model evidence given the sequence of actions. Simulations were repeated 20 times.

Figure 1 shows the results of the hierarchical Bayesian inference on this dataset. We first reported protected exceedance probability (Figure 1A), which represents the probability that each model is the most likely model across all subjects taking into account the null possibility that differences in model evidence is due to chance. This analysis revealed that the HBI has correctly identified the Kalman filter as the most likely model across the artificial datasets in all simulations with probability close to 1. Furthermore, HBI has indeed attributed about 10 and 30 artificial subjects to the reinforcement learning and Kalman filter models, respectively (Figure 1B). We then examined the performance of HBI in assigning the correct model at the individual level (Figure 1C). First, we found that the HBI has assigned the correct model to about 90% of all subjects (Figure 1C, inset). We then looked into the average of responsibilities for those artificial subjects whose underlying model was correctly assigned and for those cases whose model was erroneously assigned (Figure 1C). We found that the average of responsibilities estimated by HBI is about one for correctly identified cases and it is only slightly above chance for the rare cases that HBI failed to recognize the correct model.

This means that the HBI method was quite certain when it was successful in identifying the true model and uncertain in cases in which it failed to recognize the true model.

We then compared performance of the HBI with the HPE method of Huys et al. [10,11] and NHI. In the latter, the model evidence across the group was quantified using random effects model comparison [13,18], which uses approximate individual evidence quantified by the Laplace approximation to compute group evidence. The model comparison of HPE is essentially fixed effects, in which evidence in favor of each model is equal to the sum of individual evidence measures quantified using local Laplace approximation and the penalty due to fitting group parameters [23] (see Supplementary Materials for details). In this set of simulations, all methods performed well in recognizing the most likely model (i.e. the Kalman filter) across all samples (Figure 1d), although the HPE performed worse than the other two models (failing 15% of simulations). In the next section, we examine limitations of HPE for model comparison more thoroughly.

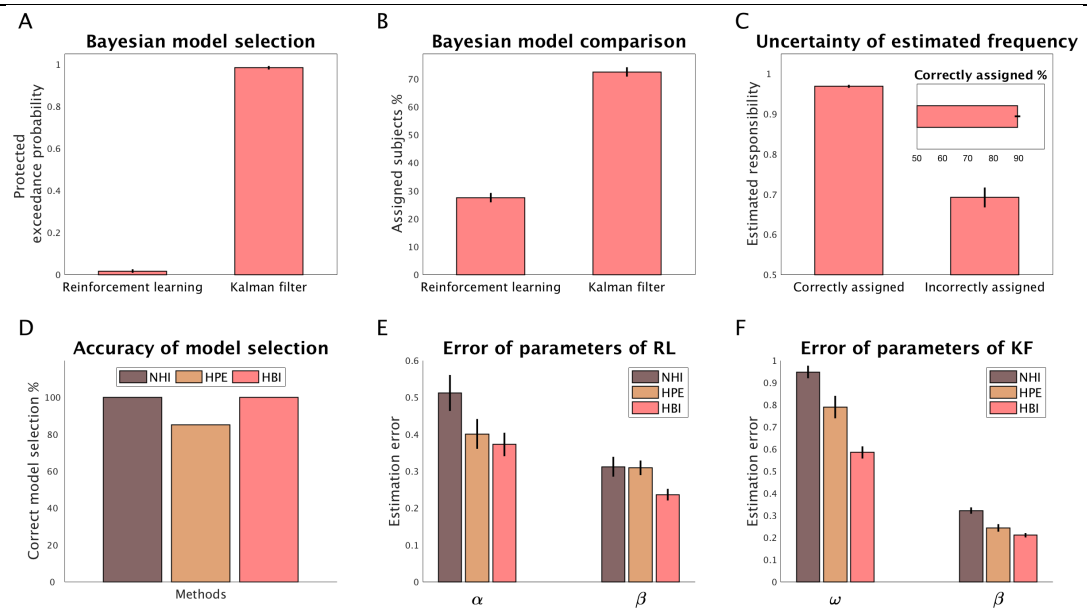


Figure 1. Performance of the HBI in a synthetic dataset. 10 and 30 artificial subjects were generated according to the reinforcement learning and Kalman filter models, respectively. A) Bayesian model selection using protected exceedance probabilities; B) Percentage of subjects explained by each model, estimated by the HBI. C) Uncertainty of HBI in estimation of responsibility for the correctly– and incorrectly– assigned subjects; Inset: percentage of correct assignment of the model by the HBI at the individual level. D) Comparison of accuracy of model selection with HPE and NHI; E, F) Error in estimating individual parameters of the reinforcement learning (E) and the Kalman filter model (F). Note that the errors are computed on the normally distributed parameters (not the transformed ones). The

estimation error is defined as the absolute difference between estimated parameters and the true parameters. In all plots, error-bars are standard errors of the mean obtained across simulations 20 times. Abbreviations: HBI, hierarchical Bayesian inference; HPE, hierarchical parameter estimation; NHI, non-hierarchical inference; RL, reinforcement learning; KL, Kalman filter.

We then investigated performance of these methods in parameter estimation. Estimation error, defined as the absolute difference between estimated parameters and true parameters used for generating data, was calculated. For both models and all parameters, the average error in parameter estimation by HBI was smaller than those by HPE and NHI (Figure 1E and 1F). Furthermore, HPE performed better than NHI in estimation across all parameters. These results were indeed theoretically expected. Unlike NHI, both HPE and HBI use group statistics to regularize parameter estimation for each individual. However, while HPE uses all subjects equally to regularize group parameters of a model, HBI weights individuals according to its belief that that model is responsible in generating each individual dataset. Further simulation analyses, in which the ratio of subjects expressing each model were different, confirmed these results (Supplementary Figure 1-2).

Robustness of model comparison to outliers

We noted before that fixed effects model comparison using HPE is very sensitive to outliers. This is because fixed effects approaches sum up evidence across all subjects. If a few outlier subjects show large evidence in favor of a model, those usually impact model comparison adversely. In contrast, the HBI takes a random effects approach, in which the contribution of every subject in favor of each model is normalized according to the corresponding responsibility, which is a relative evidence measure with a maximum of one. In this section, we show a simulation analysis to demonstrate this point.

We took the same datasets generated in the previous simulations by the reinforcement learning and Kalman filter models. We then identified one outlier subject in that dataset that showed a large evidence in favor of the reinforcement learning model. This outlier subject was then used to create datasets with 1, 2 or 3 outliers by copying it 1, 2 or 3 times, respectively, and adding those copies to the original dataset.

We then compared the performance of NHI, HPE and HBI. Note that while NHI and HBI perform random effects model comparison, HPE perform a fixed effects model comparison. As shown in Figure 2, whereas the performance of HPE is very sensitive to outliers, the random effects model comparison of NHI and HBI are robust. Note that although

NHI performs well in model selection here, we will demonstrate its limitations for model comparison in the next section.

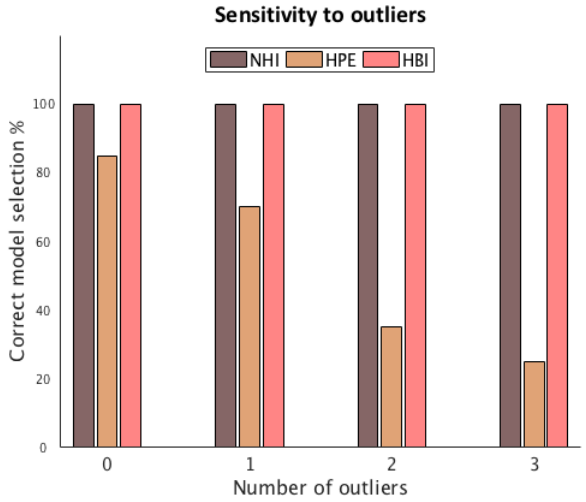


Figure 2. Sensitivity of model selection to outliers. The same 20 datasets simulated in the previous section were used as the base datasets (i.e. 0 outliers) and the effects of adding 1, 2 or 3 outliers to each dataset were examined. The HPE is severely sensitivity to outliers, while the other two (random effects) methods are robust.

Model comparison and parameter estimation in nested models

We then considered a challenging problem in which the number of free parameters in two models is different and one model is a special case of the other one. Such problems are ubiquitous in studies using computational models and inference using hierarchical approaches is typically even more advantageous in this setting, as the variance explained by such models are more likely to overlap.

The first model was again assumed to be a reinforcement learning model with a constant learning rate parameter, α. The second model, however, was assumed to contain two different learning rates depending on whether the prediction error is positive or negative (dual-α reinforcement learning, commonly used to assess asymmetries in learning from positive vs negative prediction errors [24,25]). Both models use the same choice function, i.e., a softmax function with an inverse-temperature parameter, β. The reinforcement learning and the dual-α reinforcement learning models were then used to simulate 10 and 30 artificial datasets, respectively. Note that the reinforcement learning model is a nested case of the dual-α reinforcement learning, in which $\alpha^{+} = \alpha^{-}$.

As Figure 3 shows, the HBI method was successful in model selection (i.e. recognizing the most likely model, Figure 3A) and attributed about 10 and 30 artificial subjects to the

reinforcement learning and dual-α reinforcement learning models, respectively (Figure 3B). At the individual level, HBI assigned the correct model to each individual in 95% of all subjects and was also quite certain when it was successful in selecting the right model (Figure 3C). In contrast, in those rare cases in which HBI failed to recognize the correct underlying model, it assigned responsibility that was only slightly above chance.

Next, we compared performance of the HBI with that of NHI and HPE. Here, NHI fails to choose correctly the most likely model in 75% of simulations. This is because non-hierarchical methods typically over-penalize more complex models, because they neglect the structure of the data. In particular, the issue is that a model with one additional parameter adds one independent free parameter per subject in the non-hierarchical case, which carries an excessive overfitting penalty, whereas these parameters are pooled by being drawn from a common distribution in the hierarchical setting, ensuring less overfitting and a more moderate complexity penalty. The HPE method performs much better, with a correct model selection in 80% of simulations. The HBI is successful in model selection in all simulations. We again examined performance of these methods in parameter estimation by calculating the absolute difference between estimated parameters and true parameters used for generating data. We found that errors in parameters estimated by HBI were smaller than those estimated by NHI or HPE for all parameters and both models, albeit the degree of improvement varies for different parameters. Further simulation analyses, in which the ratio of subjects expressing each model were different, confirmed these results (Supplementary Figure 3-4).

Note that the estimation errors of HBI are much smaller than those of HPE. Consider, for example, the learning rate parameter of the reinforcement learning model, α (Figure 3E). In generating the datasets for this analysis, α was assumed to be smaller than the learning rate parameters of the dual-α reinforcement learning model. Since the HPE uses average statistics across all subjects (even those generated by the dual-α model) to constrain parameters, the group average estimate of α by HPE was much larger than the true average. Therefore, the individual estimates of α by HPE are also tend to be larger than the true parameters, resulting in larger estimation error. The HBI does not have this problem because the group statistics are estimated using a weighted average, in which the weights are the corresponding responsibilities of models.

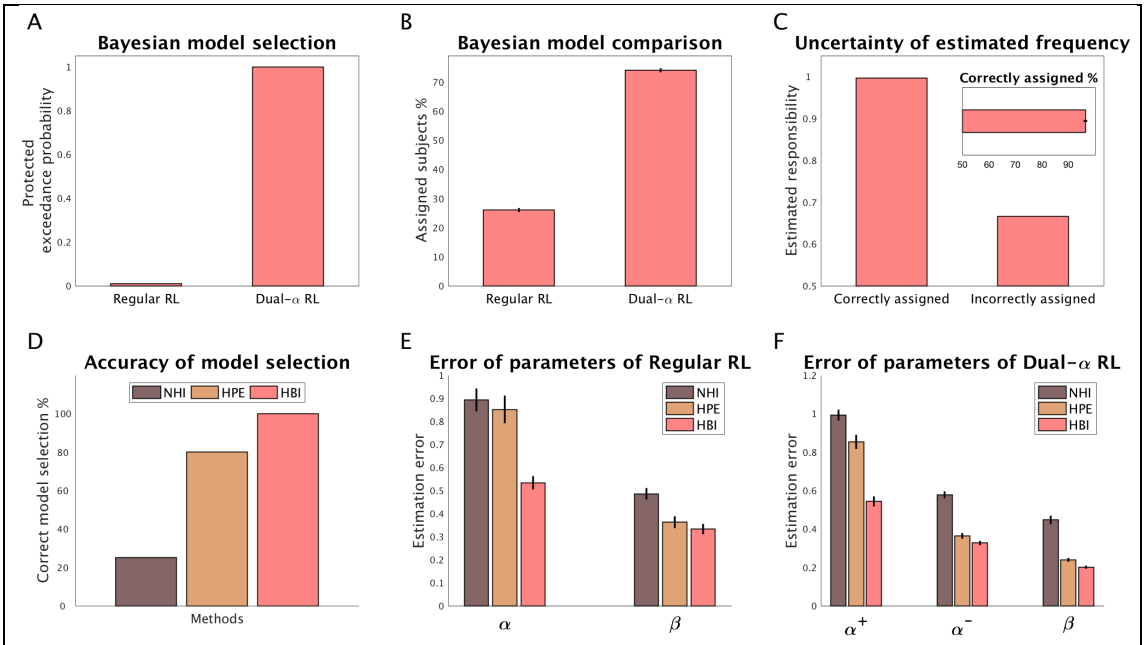


Figure 3. Performance of the HBI in a synthetic dataset. 10 and 30 artificial subjects were generated according to the reinforcement learning and dual-α reinforcement learning models, respectively. A) Bayesian model selection using protected exceedance probabilities; B) Percentage of subjects explained by each model, estimated by the HBI; C) Comparing uncertainty of HBI in estimation of responsibility for the correctly– and incorrectly– assigned subjects; D) Comparing accuracy of model selection by NHI, HPE, and HBI; E, F) Error in estimating individual parameters of the reinforcement learning (E) and the dual-α reinforcement learning model (F). Note that the normalized estimation errors are computed on the normally distributed parameters (not the transformed ones). The estimation error is defined as the absolute difference between estimated parameters and the true parameters. In all plots, error-bars are standard errors of the mean obtained across simulations 20 times. Abbreviations: HBI, hierarchical Bayesian inference; HPE, hierarchical parameter estimation; NHI, non-hierarchical inference; RL, reinforcement learning.

It is also important to note that all these methods are sensitive to amount of within-subject data (i.e. the number of trials). Importantly, the HBI is even more useful when there are limited number of trials (Figure 4). In this case, non-hierarchical methods, such as NHI, over-penalize complex models even more, as there are less data-points per subject to justify additional parameters. Furthermore, in this case, the HPE model selection performance is even more sensitive to outliers as it is more likely to have outliers when data per subject is limited (Figure 4A-C).

Note that hierarchical methods are also sensitive to amount of between-subject data (i.e. the number of subjects expressing each model). The difference between HPE and HBI is

marginal in this case, although the greater benefit of HBI is when there are limited number of subjects as HPE is more likely to be impacted by outliers in those cases (Supplementary Figures 5). Furthermore, our simulations showed that when there is only one model in the model space, both HPE and HBI show (similar amount of) benefit compared with NHI in estimating parameters when there are less amount of within-subject data (Supplementary Figures 6), as reported in previous works [12].

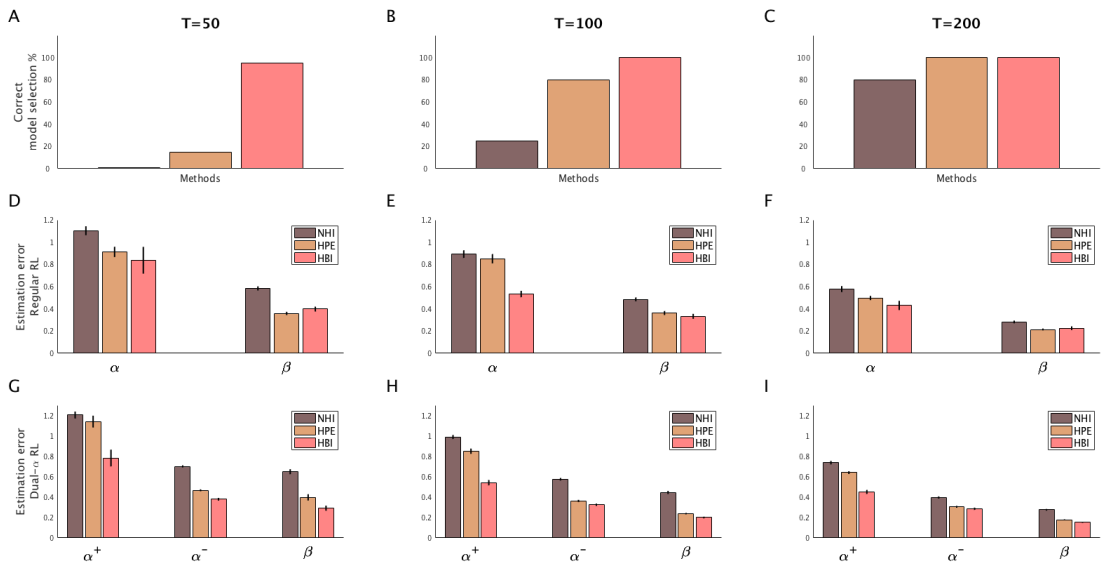


Figure 4. Performance of the HBI as a function of number of trials (T). 10 and 30 artificial subjects were generated according to the reinforcement learning and dual-α reinforcement learning models, respectively. These simulations were repeated 3 times with different number of trials per subject. A-C) Accuracy of model selection by NHI, HPE, and HBI for T=50, T=100 and T=200 trials; D-F) Error in estimating individual parameters of the reinforcement learning. G-I) Error in estimating individual parameters of the dual-α reinforcement learning. Note that the estimation errors are computed on the normally distributed parameters (not the transformed ones). The estimation error is defined as the absolute difference between estimated parameters and the true parameters. In all plots, error-bars are standard errors of the mean obtained across simulations 20 times. Abbreviations: HBI, hierarchical Bayesian inference; HPE, hierarchical parameter estimation; NHI, non-hierarchical inference; RL, reinforcement learning.

Inference about model parameters at the population level

We then focused on inference about parameters of a fitted model at the population level, and tested performance of the HBI using two simulation analyses. We focus on an example

that represents a typical inference problem at the population level for parameters of a computational model.

Consider a situation in which subjects should learn stimulus-action-outcome contingencies. The subject's task is to either to make a go-response by approaching the stimulus or to do nothing (i.e. no-go response). Furthermore, assume that the stimulus is either emotionally appetitive or aversive (e.g. a happy or an angry face cue), but the outcome value is independent of the emotional content of the stimulus. A question of interest is whether the emotional content (happy versus angry) of stimuli induces opposite biases in making a go response, regardless of action values (a form of Pavlovian to instrumental transfer). This is easy to test using a reinforcement learning model with one additional bias parameter, b (we call this model biased reinforcement learning). The bias is assumed to be $+b$ for the emotionally appetitive stimulus and $-b$ for the emotionally aversive stimulus. Thus, for larger values of b, the subject has a tendency to choose a go response after seeing the emotionally appetitive stimulus and a no-go response after seeing the emotionally aversive stimulus.

We simulated a dataset including 20 artificial subjects using this model and a randomly generated reward sequence (binarized Gaussian random-walk). Importantly, we assumed that the null hypothesis was true at the group level by drawing the bias parameter, b, randomly from a normal distribution with zero mean and variance of 1. A collection of 1000 datasets (each containing 20 artificial subjects) was simulated. The HBI method estimates the uncertainty of group mean parameters and corresponding degrees of freedom and gives the posterior belief about the group mean of parameters according to those estimates. One can then calculate the probability (p-value) that the estimated bias at the population level is generated by the null hypothesis. Note that in this example, all subjects are assigned to the biased reinforcement learning model as the model in the model space.

Under the null hypothesis, the HBI theory indicates that the t-statistics, $t = \frac{a_k}{s_k/\sqrt{n_k}}$, in which a_k is the posterior mean of group parameters, s_k is the empirical deviance statistics and $n_k = 1 + \overline{N_k}$, has a standard t-student distribution with n_k as degrees of freedom. Therefore, we expect that the corresponding p-value takes a uniform distribution between 0 and 1. This simulation analysis showed that the resulting p-value for the bias parameter indeed took a uniform distribution (Figure 5A).

Now we consider situations in which there is more than one model in the model space. Here, the HBI first infers the number of subjects explained by each model, $\overline{N_k}$, and then quantifies hierarchical errors and degrees of freedom according to $\overline{N_k}$ and the empirical group

mean and deviance statistics. Therefore, we considered the same stimulus-action-outcome learning experiment as above and simulated a dataset including 40 artificial subjects. Data for half of subjects were generated using the same biased reinforcement learning model and data for the other half were generated using the dual-α reinforcement learning model explained in previous simulations. We again assumed that the null hypothesis is true at the group level by drawing the bias parameter randomly from normal distribution with zero mean and variance of 1. Again, a collection of 1000 datasets (each containing 40 artificial subjects) was simulated and the distribution of corresponding *p*-value was generated. Consistent with the theory, this simulation analysis showed that the resulting *p*-value for the bias parameter well follows a uniform distribution (Figure 5B).

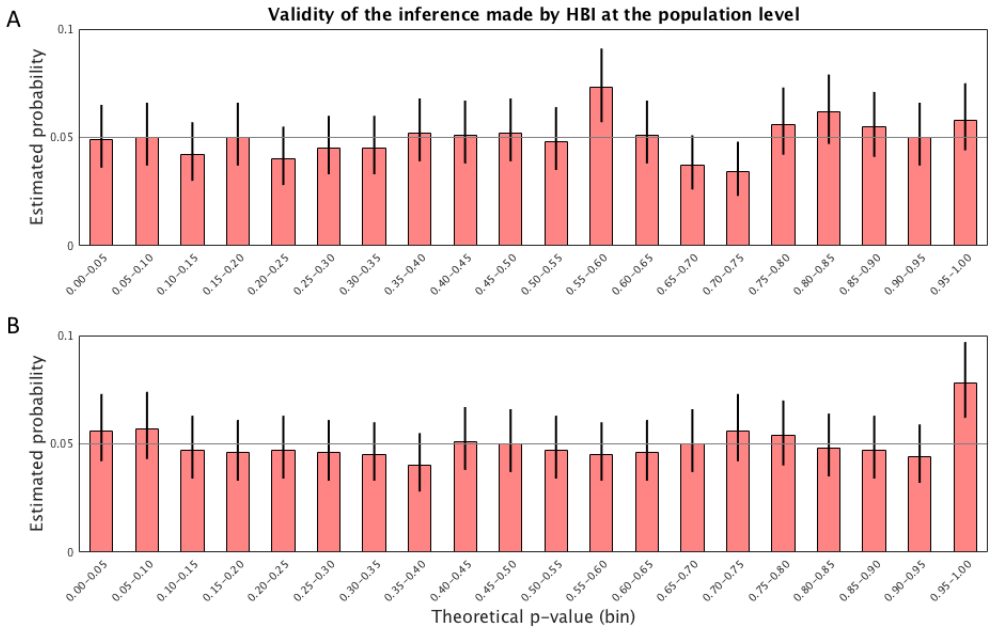
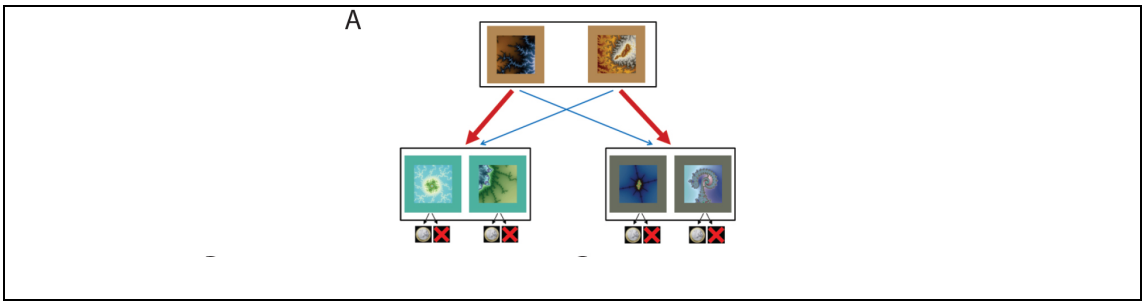


Figure 5. Performance of the HBI for statistical inference about model parameters at the population level. A bias parameter was generated under the null hypothesis at the group level in two simulations where there is only one model in the model-space (A) or there are more than one model in the model-space (B). Under the null-hypothesis, the resulting *p*-value of the group mean across all simulations is theoretically expected to have a uniform distribution. The distribution of estimated *p*-values by the HBI are plotted as a function of theoretical *p*-value in bins with the length of 0.05. The error-bars are 95% confidence intervals for the binomial distribution.

Empirical dataset: HBI reveals meaningful individual differences

We then applied the HBI method to empirical choice data from 31 subjects performing the two-step Markov decision task introduced by Daw et al. [26]. This task is a well-known paradigm to distinguish two behavioral modes, model-based and model-free learning [27–29]. Previous works have shown that there are important individual differences in this task [26,30], especially in the degree to which people employ a model-based strategy, and those differences are related to neuroanatomical [30,31], psychological [32], genetic [33] and psychiatric [34] trait scores. Daw et al. [26] have proposed three reinforcement learning accounts, a model-based, a model-free and their hybrid (which nests the other two and combines their estimates according to a weight parameter), to disentangle contribution of these two behavioral modes on choices. Here, we skip the details of the models and focus on application of the HBI to a model space consisting of model-free, model-based and hybrid accounts. The dataset used for this analysis have been reported elsewhere [30].

Before analyzing the empirical dataset, we did a simulation analysis of this task and model space. We verified that the HBI recovers the parameters of the models better than alternative methods. In particular, the critical weight parameter of the hybrid model, which determines the degree that each account influences behavior, was significantly better recovered by the HBI than the other methods (in all 20 simulations, HBI did better than both HPE and NHI, Supplementary Figure 7). We then applied the HBI on the empirical dataset of this task (Figure 6). As Figure 6B shows, consistent with previous findings on this task, the hybrid model accounts best for choices in this task with a protected exceedance probability close to 1. About 25 (%81 of all subjects) and 6 subjects have been assigned to the hybrid and model-based models, respectively, while the model-free took no responsibility (Figure 6C). The estimated group mean and the corresponding hierarchical errors of the hybrid model are plotted in Figure 6D. Table 1 shows the weighted average and standard deviation computed by the HBI for the hybrid model in which the weights are given by responsibilities taken by the hybrid model in explaining data of each individual subject.



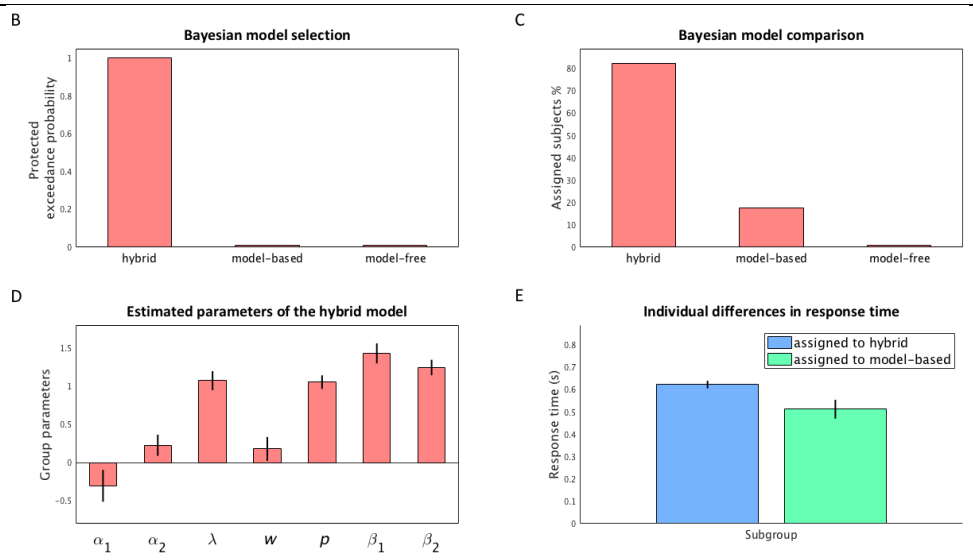


Figure 6. Performance of the HBI in an empirical dataset. A) the two-step Markov decision task [26]: every trial of the task comprises two steps. On each trial, subjects first made a choice between two fractal stimuli leading to one of the two different second-stage sets represented by different colors. Subjects then made another choice between two stimuli presented in the second-stage set. Each stimulus at the second-stage was associated with a specific probability of delivering a monetary reward. Critically, each choice at the first stage led predominantly (70%) to one of the two sets at the second stage (common transition), and less frequently (30%) to the other set (rare transition). This feature of the task allows for distinguishing contribution of model-based and model-free accounts on the first level choices. B) Bayesian model selection among the three models using HBI. Protected exceedance probabilities are plotted; C) Bayesian model comparison using HBI. Percentage of subjects explained by each model are plotted. HBI assigned almost 80% of subjects to the hybrid model. D) Estimated group mean parameters of the hyrbid model are plotted. Errorbars are corresponding hierarchical standard errors. The hybrid model [26] contains seven parameters: two learning rates, α_1 and α_2, for the first and second stage choice; an eligibility trace parameter, λ; a weight parameter, w, mixing model-based and model-free values (logit-transformed values plotted for all these parameters); a perserveration parameter p; and two decision noise parameters, β_1 and β_2 (log-transformed values are plotted), for the first and second stage choices. E) Response time (mean and standard errors) for the first choice for the two subgroups that were assigned to hybrid or model-based account by HBI. Those who were assigned to hybrid showed slower response time.

We also performed further analysis testing whether individual differences found by the HBI generalize to individual differences in conceptually related, yet independent, data. We reasoned that subjects showing a hybrid strategy might be slower in their choice, as the hybrid model requires combining of model-based and model-free values (which in some trials might be in conflict). Therefore, we looked at the median of response time across all first-level choices for each subject and tested whether there is a difference in response times between those subjects who employed a hybrid strategy and those who employed a model-based strategy as estimated by the HBI. As Figure 6E shows, the subgroup attributed to the hybrid model by the HBI showed slower response time compared to those subjects attributed to the model-based account ($p=0.03$, Wilcoxon test). A similar analysis using individual differences found by the NHI revealed no significant difference in response times. These results suggest that HBI reveals meaningful individual differences generalizing to unseen data.

Discussion

In this work, we have introduced a novel method, a hierarchical and Bayesian inference framework, for parameter estimation and model comparison. The HBI framework is hierarchical in the sense that parameters at the individual level are regularized by statistics across all individuals in the group. The HBI framework is Bayesian in the sense that all uncertainties at both individual and group levels are represented by probability distributions. The HBI framework has major theoretical advantages over current state-of-the-art methods, mainly because it combines two sorts of inference (about model identity and model parameters) in a single hierarchical model, which are interdependent but have previously been treated separately. Our simulation results demonstrated these advantages experimentally.

In this work, we took an empirical Bayes approach [35,36], in which priors are constructed based on data. In other words, parameters at the individual level are regularized by statistics across all individuals in the group. Furthermore, we took a so-called random effects approach to model identity [13], which indicates that different models might underlie data in different subjects. This is in contrast to previous hierarchical methods for model fitting, which assume the same model underlie data in all subjects (fixed effects assumption). The random effects approach to hierarchical inference has important consequences for both parameter estimation and model comparison. Moreover, we took a fully Bayesian approach by quantifying uncertainty at the group level, which enabled us to develop statistical tests about group parameters and to quantify corresponding statistical errors.

Empirical Bayes methods play an increasing role in modern statistics. These methods essentially take a hierarchical approach, by assuming that individual data are generated based on the probabilistic properties of population. This hierarchical approach has important consequences. The most important consequence is that they provide a promising solution to the classical problem of priors in Bayesian statistics by providing informative, yet objective, priors at the individual level. Furthermore, unlike non-hierarchical methods, model comparison based on these methods is not biased towards too simple models. This is because non-hierarchical methods assume that extra parameters of a complex model are independent. For example, consider a model space in which the more complex model has one extra free parameter and there are 40 subjects in the dataset. Fitting the dataset with the complex model using non-hierarchical methods introduces 40 additional independent free parameters, driving an excessive penalty for overfitting. The hierarchical approach, however, assumes that the individual parameters are dependent, as they are all generated according to the same distribution. Modeling this hierarchical dependency enables those methods to avoid penalizing complex models excessively. Our simulation results demonstrate this point experimentally (Figure 3D). While the non-hierarchical method failed to select the correct model with one additional parameter, the HBI was successful in selecting the correct model (Figure 3D).

The HBI method introduced in this paper is built based on the random effects view that different models might underlie data in different subjects. Taking this view enabled us to address problems caused by taking the model identity as a fixed effect in some hierarchical parameter estimation procedures. For parameter estimation, the fixed effects assumption biases the group parameters because it assumes that all subjects contribute equally to the group parameters. The proposed HBI framework solves this problem by weighting contribution of each subject to group statistics by the degree to which that model is likely to be the true underlying model for that subject (Figures 1 and 3). For model comparison, the fixed effects assumption leads to oversensitivity to outliers [13] as the evidence across the group is driven by the sum of individual evidences. Our simulation results (Figure 2) showed that only a few outliers change lead to incorrect model selection inference made by the fixed effects assumption. The proposed HBI method solves this problem by normalizing individual evidence across all candidate models. Specifically, the HBI framework quantifies the responsibility of each model k in generating each subject data, a metric lying between 0 and 1. For every subject, the responsibility sums up to 1 across all candidate models as it partitions probability space among those models (see [13,18] for a similar approach). It is then easy to compare models by enumerating responsibilities across the group in favor of each model or by estimating the most likely model.

Another major contribution of this paper is to provide a statistical solution to the inference problem at the group level using hierarchically fitted parameters. For models fitted by a non-hierarchical method, such as maximum likelihood or Laplace approximation, it is statistically valid to use conventional statistical tests on fitted parameters to make inference at the group level. However, for datasets fitted by a hierarchical method in which the individual fits are regularized according to statistics of the group data, conventional statistical tests are in some cases not valid, because the parameter estimates are non-independent from subject to subject. Our fully Bayesian approach enabled us to address this issue. This is because the HBI infers the posterior distribution of group parameters. Our method provides an intuitive solution to this problem in the form a *t*-statistic, in which all the group statistics are computed according to the estimated responsibilities of the corresponding model in generating each individual data. Thus, the HBI quantifies the uncertainty of the group parameters and thereby the corresponding hierarchical errors. Simulation analyses (Figure 5) highlighted this point experimentally by showing that *p*-values computed by the HBI under the null hypothesis (i.e. when the group parameter is normally distributed around zero) follow a uniform distribution. Therefore, the HBI framework enables researchers to make statistical claims about parameters at the group level.

In addition to model comparison, the HBI framework can also be used for model selection in situations where the goal is to select one of the models as the best model across the group. Exceedance probability is a metric proposed [13] to perform model selection using a random effects approach. An important revision of this metric called protected exceedance probability [18] also takes into account the possibility that none of the models in a model space is supported sufficiently by data, i.e. the differences in model evidence are due to chance. As the HBI framework treats model identity as a random effect, it is possible to compute exceedance and protected exceedance probabilities.

There are increasing efforts to exploit advances in computational modeling for understanding mental disorders [3–6]. Recent works, however, have started to tackle challenges related to quantifying uncertainty in diagnosis and also in evaluation of treatment effects. For example, hierarchical unsupervised generative modeling, have used Monte-Carlo and variational methods to identify cluster of subjects showing similar patterns of neural connectivity [37,38]. HBI also offers a promising solution by quantifying uncertainty in assigning models to data generated by a single case. Our simulation (Figure 1C and 2C) showed that HBI assigns probabilities that are close to chance in cases in which model identification goes wrong. This can help us to move towards better diagnosis and precise evaluation of different treatments [39].

In summary, the HBI framework proposed in this work rests on a hierarchical view of both hypothesis testing (i.e. model comparison) and parameter estimation for multi-subject studies and thus provides a generic framework for statistical inference. Moreover, the HBI framework runs fully automatically and it does not rely on hand tuning of parameters. Therefore, we expect this method to be useful for a wide range of studies testing different hypotheses in a multi-subject setting. This includes not only computational models of learning and decision making, but also any statistical models of brain or behavior.

References

1. Daw ND, Doya K. The computational neurobiology of learning and reward. Curr Opin Neurobiol. 2006;16: 199–204. doi:10.1016/j.conb.2006.03.006

2. O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci. 2007;1104: 35–53. doi:10.1196/annals.1390.022

3. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. 2011;14: 154–162. doi:10.1038/nn.2723

4. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012;16: 72–80. doi:10.1016/j.tics.2011.11.018

5. Friston KJ, Stephan KE, Montague R, Dolan RJ. Computational psychiatry: the brain as a phantastic organ. Lancet Psychiatry. 2014;1: 148–158. doi:10.1016/S2215-0366(14)70275-5

6. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016;19: 404–413. doi:10.1038/nn.4238

7. Friston KJ, Harrison L, Penny W. Dynamic causal modelling. NeuroImage. 2003;19: 1273–1302.

8. Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ. Ten simple rules for dynamic causal modeling. NeuroImage. 2010;49: 3099–3109. doi:10.1016/j.neuroimage.2009.11.015

9. Cohen JD, Daw N, Engelhardt B, Hasson U, Li K, Niv Y, et al. Computational approaches to fMRI analysis. Nat Neurosci. 2017;20: 304–313. doi:10.1038/nn.4499

10. Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, et al. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol. 2011;7: e1002028. doi:10.1371/journal.pcbi.1002028

11. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol. 2012;8: e1002410. doi:10.1371/journal.pcbi.1002410

12. Wiecki TV, Sofer I, Frank MJ. HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Front Neuroinformatics. 2013;7. doi:10.3389/fninf.2013.00014

13. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. NeuroImage. 2009;46: 1004–1017. doi:10.1016/j.neuroimage.2009.03.025

14. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B Methodol. 1977;39: 1–38.

15. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. In: Jordan MI, editor. Learning in Graphical Models. Springer, Dordrecht; 1998. pp. 105–161. doi:10.1007/978-94-011-5014-9_5

16. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

17. Neal RM, Hinton GE. A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants. In: Jordan MI, editor. Learning in Graphical Models. Springer, Dordrecht; 1998. pp. 355–368. doi:10.1007/978-94-011-5014-9_12

18. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies - revisited. NeuroImage. 2014;84: 971–985. doi:10.1016/j.neuroimage.2013.08.065

19. Daw ND. Trial-by-trial data analysis using computational models. In: Delgado MR, Phelps EA, Robbins TW, editors. Decision Making, Affect, and Learning: Attention and Performance XXIII. New York: Oxford University Press; 2011. pp. 3–38.

20. Daunizeau J, Adam V, Rigoux L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. PLoS Comput Biol. 2014;10: e1003441. doi:10.1371/journal.pcbi.1003441

21. Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. NeuroImage. 2007;34: 220–234. doi:10.1016/j.neuroimage.2006.08.035

22. Daunizeau J, den Ouden HEM, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ. Observing the observer (I): meta-bayesian models of learning and decision-making. PloS One. 2010;5: e15554. doi:10.1371/journal.pone.0015554

23. Piray P, Zeighami Y, Bahrami F, Eissa AM, Hewedi DH, Moustafa AA. Impulse control disorders in Parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation. J Neurosci. 2014;34: 7814–7824. doi:10.1523/JNEUROSCI.4063-13.2014

24. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 2007;104: 16311–16316. doi:10.1073/pnas.0706111104

25. Piray P. The role of dorsal striatal D2-like receptors in reversal learning: a reinforcement learning viewpoint. J Neurosci. 2011;31: 14049–14050. doi:10.1523/JNEUROSCI.3008-11.2011

26. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011;69: 1204–1215. doi:10.1016/j.neuron.2011.02.027

27. Dickinson A, Balleine BW. The role of learning in motivation. In: Gallistel R, editor. Stevens' Handbook of Experimental Psychology, Learning, Motivation, and Emotion. 3rd ed. 2002.

28. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8: 1704–1711. doi:10.1038/nn1560

29. Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol. 2011;7: e1002055. doi:10.1371/journal.pcbi.1002055

30. Piray P, Toni I, Cools R. Human Choice Strategy Varies with Anatomical Projections from Ventromedial Prefrontal Cortex to Medial Striatum. J Neurosci. 2016;36: 2857–2867. doi:10.1523/JNEUROSCI.2033-15.2016

31. Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron. 2013;80: 914–919. doi:10.1016/j.neuron.2013.08.009

32. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A. 2013;110: 20941–20946. doi:10.1073/pnas.1312011110

33. Doll BB, Bath KG, Daw ND, Frank MJ. Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. J Neurosci Off J Soc Neurosci. 2016;36: 1211–1222. doi:10.1523/JNEUROSCI.1901-15.2016

34. Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. eLife. 2016;5. doi:10.7554/eLife.11305

35. Casella G. An Introduction to Empirical Bayes Data Analysis. Am Stat. 1985;39: 83–87. doi:10.2307/2682801

36. Robbins H. An Empirical Bayes Approach to Statistics. The Regents of the University of California; 1956. Available: https://projecteuclid.org/euclid.bsmsp/1200501653

37. Yao Y, Raman SS, Schiek M, Leff A, Frässle S, Stephan KE. Variational Bayesian inversion for hierarchical unsupervised generative embedding (HUGE). NeuroImage. 2018;179: 604–619. doi:10.1016/j.neuroimage.2018.06.073

38. Raman S, Deserno L, Schlagenhauf F, Stephan KE. A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. J Neurosci Methods. 2016;269: 6–20. doi:10.1016/j.jneumeth.2016.04.022

39. Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, et al. Computational neuroimaging strategies for single patient predictions. NeuroImage. 2017;145: 180–199. doi:10.1016/j.neuroimage.2016.06.038

Appendix A

In this appendix, we give a formal treatment of the HBI framework. First, we give the probabilistic model underlying HBI. In A.2, our approach for making inference (the full proof is given in appendix B) and related assumptions are given. In A.3, the HBI algorithm is presented. In A.4, we show how the problem of statistical inference about group parameters is solved in HBI. In A.5, we show how HBI can be used for making inference about a new subject. Finally, in A.6, we highlight some practical points, such as the initialization of the parameters and their settings.

A.1 Probabilistic model

We begin by describing the probabilistic model of the HBI. Consider an observed dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_n$ is the dataset (e.g. choices) of nth subject and N indicates the number of subjects and a model-space including K candidate models, $M_1 \ldots M_K$. Moreover, suppose that the prior probability of each model in the population is given by $\mathbf{m} = \{m_1, \ldots, m_K\}$. For each dataset, $\mathbf{x}_n$, we assume that there is a latent variable $\mathbf{z}_n$ comprising a 1-of-K binary random vector, in which z_{kn} is one if $\mathbf{x}_n$ generated is by the kth model. Thus, the probability of the latent variable across all subjects, $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$, is assumed to have a multinomial distribution,

$$p(\mathbf{Z}|\mathbf{m}) = \prod_k \prod_n m_k^{z_{kn}}. \tag{1}$$

Each model M_k in the model-space is supposed to compute the probability of a given dataset (e.g. a set of choices) given a set of parameters, $\mathbf{h}_{kn}$. For example, the reinforcement learning model computes the probability of choices using two parameters: a learning rate and a decision noise parameter. The number of models and their structures depend on specific scientific questions. Here, we take a general approach by making no specific assumption about the number of models, K. Thus, the kth model in the model-space, M_k, computes the probability of dataset $\mathbf{x}_n$ given the parameter vector $\mathbf{h}_{kn}$, which is denoted by $p(\mathbf{x}_n|\mathbf{h}_{kn}, M_k)$. Note that the number of parameters in model k, denoted by D_k, might be different across models. Since data for each subject is generated by one of the models, which is denoted in the binary vector $\mathbf{z}_n$, the probability of

the observed dataset given the model-space is

$$p(\mathbf{X}|\mathbf{H},\mathbf{Z}) = \prod_k \prod_n p(\mathbf{x}_n|\mathbf{h}_{kn}, M_k)^{z_{kn}}, \tag{2}$$

where $\mathbf{H}$ denotes all the parameters across all participants and models. The parameters of kth model are assumed to have a multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and precision matrix $\mathbf{T}_k$,

$$p(\mathbf{H}|\mathbf{Z},\boldsymbol{\mu},\mathbf{T}) = \prod_k \prod_n \mathcal{N}(\mathbf{h}_{kn}|\boldsymbol{\mu}_k, \mathbf{T}_k^{-1})^{z_{kn}}, \tag{3}$$

where $\mathbf{T}_k$ is a diagonal positive-definite matrix.

We also introduce a distribution over model frequencies, $\mathbf{m}$. Since this is a probability over probabilities (which sum to one), we use the Dirichlet distribution as the prior:

$$p(\mathbf{m}) = \mathrm{Dir}(\mathbf{m}|\boldsymbol{\alpha_0}) = C(\alpha_0) \prod_{k=1}^{K} m_k^{\alpha_0 - 1}, \tag{4}$$

where $C(\alpha_0)$ is the normalizing constant for the Dirichlet distribution.

We also take group parameters $\boldsymbol{\mu}$ and $\mathbf{T}$ as random variables, which allows us to evaluate their posterior distribution given data. We introduce conjugate priors for these variables, a Gaussian-Gamma prior in which the distribution over $\boldsymbol{\mu}_k$ depends on $\mathbf{T}_k$:

$$p(\boldsymbol{\mu}|\mathbf{T}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{a}_0, (b\mathbf{T}_k)^{-1})$$

$$p(\mathbf{T}) = \prod_{k=1}^{K} \prod_{i=1}^{D_k} \mathcal{G}(\tau_{ki}|v, s),$$

where $\mathcal{G}(.)$ denotes Gamma distribution. Here, τ_{ki} is the ith diagonal element of $\mathbf{T}_k$. Assuming that $\boldsymbol{\tau}_k$ is a vector containing τ_{ki}, by defining $\mathbf{T}_k = \mathrm{diag}(\boldsymbol{\tau}_k)$, in which diag(.) is an operator outputting a diagonal matrix with elements given by $\boldsymbol{\tau}_k$, we can write these two equations in a compact form:

$$p(\boldsymbol{\mu},\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{a}_0, \mathrm{diag}(b\boldsymbol{\tau}_k)^{-1})\mathcal{G}(\boldsymbol{\tau}_k|v, \mathbf{s}), \tag{5}$$

where we have defined:

$$\mathcal{G}(\boldsymbol{\tau}_k|v,\mathbf{s}) = \prod_{i=1}^{D_k} \mathcal{G}(\tau_{ki}|v,s),$$

in which v is a scalar and $\mathbf{s}$ is a vector with D_k elements all equal to s. The full probabilistic model is given by,

$$p(\mathbf{X},\mathbf{H},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\tau},\mathbf{m}) = p(\mathbf{X}|\mathbf{H},\mathbf{Z})p(\mathbf{H}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\tau})p(\mathbf{Z}|\mathbf{m})p(\boldsymbol{\mu}|\boldsymbol{\tau})p(\boldsymbol{\tau})p(\mathbf{m}). \quad (6)$$

A.2 Variational inference

The task of Bayesian inference is to compute the posterior probabilities of latent variables given data, $p(\mathbf{H},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\tau},\mathbf{m}|\mathbf{X})$. Since the inference is intractable for the probabilistic model outlined in section A.1, we employ variational inference to compute approximate posteriors. We take a so-called mean-field approach [15] by assuming that the posterior is partially factorized as follows:

$$q(\mathbf{H},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\tau},\mathbf{m}) = q(\mathbf{H},\mathbf{Z})q(\boldsymbol{\mu},\boldsymbol{\tau},\mathbf{m}). \quad (7)$$

Note that we force no factorization in the posterior between latent variables, $\mathbf{Z}$ and $\mathbf{H}$. Using a quadratic approximation of the conditional posterior, $q(\mathbf{H}|\mathbf{Z})$, we prove in Appendix B that these posteriors are given by,

$$q(\mathbf{H},\mathbf{Z}) = \prod_k \prod_n r_{kn}^{z_{kn}} \mathcal{N}(\mathbf{h}_{kn}|\boldsymbol{\theta}_{kn},\mathbf{A}_{kn}^{-1})^{z_{kn}} \quad (8)$$

$$q(\boldsymbol{\mu},\boldsymbol{\tau},\mathbf{m}) = \mathrm{Dir}(\mathbf{m}|\boldsymbol{\alpha}) \prod_k q(\boldsymbol{\mu}_k,\boldsymbol{\tau}_k) \quad (9)$$

$$q(\boldsymbol{\mu}_k,\boldsymbol{\tau}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{a}_k,\mathrm{diag}(\beta_k\boldsymbol{\tau}_k)^{-1})\mathcal{G}(\boldsymbol{\tau}_k|\nu_k,\boldsymbol{\sigma}_k), \quad (10)$$

where ν_k and β_k are scalars and $\boldsymbol{\sigma}_k$ is a vector with the same size as $\boldsymbol{\tau}_k$. In the next section, we provide the HBI algorithm, which iteratively updates the parameters of these distributions, r_{kn}, $\boldsymbol{\theta}_{kn}$, $\mathbf{A}_{kn}$, $\boldsymbol{\alpha}$, $\mathbf{a}_k$, ν_k, β_k, and $\boldsymbol{\sigma}_k$.

A.3 HBI algorithm

After initializing the individual parameter estimates, $\boldsymbol{\theta}_{kn}$ and $\mathbf{A}_{kn}$ and responsibilities r_{kn} for all subjects and models, as well as setting prior parameters $\mathbf{a}_0$, b, s, v and α_0 (see A.6 for a simple and intuitive way for initializing and setting prior parameters), the HBI algorithm performs these steps:

- **1.** Calculate the summary statistics:

$$\overline{N}_k = \sum_n r_{kn} \tag{11}$$

$$\overline{\boldsymbol{\theta}}_k = \frac{1}{\overline{N}_k} \sum_n r_{kn} \boldsymbol{\theta}_{kn} \tag{12}$$

$$\overline{\mathbf{V}}_k = \frac{1}{\overline{N}_k} \sum_n r_{kn} \Big(\boldsymbol{\theta}_{kn} \boldsymbol{\theta}_{kn}^{\mathsf{T}} - \overline{\boldsymbol{\theta}}_k \overline{\boldsymbol{\theta}}_k^{\mathsf{T}} + \mathbf{A}_{kn}^{-1} \Big). \tag{13}$$

- **2.** Update parameters of $q(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m})$ for all models:

$$\mathbf{a}_k = \frac{1}{\overline{N}_k + b} (\overline{N}_k \overline{\boldsymbol{\theta}}_k + b \mathbf{a}_0) \tag{14}$$

$$\beta_k = b + \overline{N}_k \tag{15}$$

$$\boldsymbol{\sigma}_k = \mathbf{s} + \frac{1}{2} \mathrm{diag} \Big(\overline{N}_k \overline{\mathbf{V}}_k + \frac{b \overline{N}_k}{b + \overline{N}_k} (\overline{\boldsymbol{\theta}}_k - \mathbf{a}_0)(\overline{\boldsymbol{\theta}}_k - \mathbf{a}_0)^{\mathsf{T}} \Big) \tag{16}$$

$$\nu_k = v + \frac{1}{2} \overline{N}_k \tag{17}$$

$$\alpha_k = \alpha_0 + \overline{N}_k. \tag{18}$$

- **3.** Update the individual posterior parameters $\boldsymbol{\theta}_{kn}$, $\mathbf{A}_{kn}$ and f_{kn}, by obtaining a quadratic approximation of the function, $\ell_{kn}(\mathbf{h})$, with respect to $\mathbf{h}$:

$$\ell_{kn}(\mathbf{h}) = p(\mathbf{x}_n | \mathbf{h}, M_k) \mathcal{N}(\mathbf{h} | \mathbb{E}[\boldsymbol{\mu}_k], \mathbb{E}[\mathbf{T}_k]^{-1}), \tag{19}$$

where $\mathbb{E}[\boldsymbol{\mu}_k] = \mathbf{a}_k$ and $\mathbb{E}[\mathbf{T}_k]^{-1} = \frac{1}{\nu_k} \mathrm{diag}(\boldsymbol{\sigma}_k)$. This approximation can be

written as

$$\ell_{kn}(\mathbf{h}) \simeq f_{kn} \exp(-\frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^{\mathsf{T}} \mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})). \tag{20}$$

Note that any quadratic approximation can be used here. For example, using a Laplace quadratic approximation (which is a very common approximation for analyzing behavioral and neural data [19-22]), $\boldsymbol{\theta}_{kn}$, $\mathbf{A}_{kn}$ and f_{kn} are given by the mode, Hessian of $\log \ell_{kn}$ and the maximum value of ℓ_{kn}, respectively:

$$\boldsymbol{\theta}_{kn} = \arg\max_{\mathbf{h}} \ell_{kn}(\mathbf{h})$$

$$\mathbf{A}_{kn} = -\nabla\nabla \log \ell_{kn}(\mathbf{h})|_{\boldsymbol{\theta}_{kn}}$$

$$f_{kn} = \ell_{kn}(\boldsymbol{\theta}_{kn}).$$

- **4.** Update responsibilities,

$$r_{kn} = \frac{\rho_{kn}}{\sum_{j=1}^{K} \rho_{jn}}, \tag{21}$$

where

$$\log \rho_{kn} = \log f_{kn} + \frac{1}{2} D_k \log 2\pi - \frac{1}{2} \log |\mathbf{A}_{kn}| + \lambda_k + \mathbb{E}[\log m_k] \tag{22}$$

$$\lambda_k = \frac{D_k}{2} \big(\psi(\nu_k) - \log \nu_k - \frac{1}{\beta_k}\big) \tag{23}$$

$$\mathbb{E}[\log m_k] = \psi(\alpha_k) - \psi(\sum_{k=1}^{K} \alpha_k),$$

in which $\psi(.)$ is the digamma function.

- **5.** Terminate if stopping criteria are met, otherwise go to 1.

A.4 Statistical tests for group parameters

An important goal of computational modeling studies is to compute the distribution of parameters given data across the whole population. From a Bayesian viewpoint, this is given by the marginal posterior over the mean of group param-

eters, $\boldsymbol{\mu}_k$, which reads

$$\begin{aligned} p(\boldsymbol{\mu}_k|\mathbf{X}) &\simeq \int q(\boldsymbol{\mu}_k, \boldsymbol{\tau}_k) d\boldsymbol{\tau}_k \\ &= \int \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{a}_k, (\beta_k \boldsymbol{\tau}_k)^{-1}) \mathcal{G}(\boldsymbol{\tau}_k|\nu_k, \boldsymbol{\sigma}_k) d\boldsymbol{\tau}_k \\ &= \mathcal{S}t(\boldsymbol{\mu}_k|\mathbf{a}_k, \boldsymbol{\eta}_k, n_k), \end{aligned}$$

where $n_k = 2\nu_k = 2v + \overline{N}_k$ is the number of degrees of freedom of the Student distribution and $\boldsymbol{\eta}_k = \nu_k \beta_k \boldsymbol{\sigma}_k^{-1}$ is the inverse-scale parameter. Therefore, the random variable $\mathbf{t} = \boldsymbol{\eta}_k^{\frac{1}{2}}(\boldsymbol{\mu}_k - \mathbf{a}_k)$ takes a form of standard Student distribution with n_k degrees of freedom. By defining $s_{ki}^2 = \frac{2}{\beta_k}\sigma_{ki}$, in which s_{ki}^2 corresponds to empirical variance (c.f. equation (16)), we can write this result in an intuitive form,

$$p(\mu_{ki}|\mathbf{X}) = \mathcal{S}t\left(\frac{\mu_{ki} - a_{ki}}{s_{ki}/\sqrt{n_k}}|n_k\right). \tag{24}$$

Noting the similarity between $s_{ki}/\sqrt{n_k}$ and the standard error of the mean, we called $s_{ki}/\sqrt{n_k}$ the hierarchical error.

A.5 Predictive distribution for a new subject

In many situations, researchers are interested to fit a new dataset to a particular model and find corresponding parameters. In Bayesian statistics, this is called the predictive distribution and it is given by marginalizing over group parameters. Suppose that $\mathbf{x}^*$ and $\mathbf{h}_k^*$ denote the new dataset and its corresponding parameters for model k. The marginal distribution $p(\mathbf{x}^*, \mathbf{h}_k^*|z_k^* = 1, \mathbf{X})$ is the predictive distribution given the observed dataset $\mathbf{X}$ assuming that the new data is generated by the kth model. This distribution is given by:

$$\begin{aligned} p(\mathbf{x}^*, \mathbf{h}_k^*|z_k^* = 1, \mathbf{X}) &= \int p(\mathbf{x}^*|\mathbf{h}_k^*, M_k) p(\mathbf{h}_k^*|\boldsymbol{\mu}_k, \boldsymbol{\tau}_k, z_k^* = 1) p(\boldsymbol{\mu}_k, \boldsymbol{\tau}_k|\mathbf{X}) d\boldsymbol{\mu}_k d\boldsymbol{\tau}_k \\ &= p(\mathbf{x}^*|\mathbf{h}_k^*, M_k) \mathcal{S}t(\mathbf{h}_k^*|\mathbf{a}_k, (1+\beta_k)^{-1}\boldsymbol{\eta}_k, n_k), \end{aligned}$$

where $\boldsymbol{\eta}_k$ and n_k have been defined in the previous section. This distribution can also be written in terms of standard Student distribution with n_k degrees of freedom. Furthermore, if we assume that $b = 2v$, which is a reasonable

assumption (see the next section), this distribution is given by

$$p(\mathbf{x}^*, \mathbf{h}_k^* | z_k^* = 1, \mathbf{X}) = p(\mathbf{x}^* | \mathbf{h}_k^*, M_k) \mathcal{S}t(\mathrm{diag}(\mathbf{s}_k)^{-1}(\mathbf{h}_k^* - \mathbf{a}_k) | n_k),$$

where $\mathbf{s}_k$ is a vector of corresponding empirical deviance parameters, defined in the previous section. Using this joint distribution, one can use sampling methods to obtain the posterior over parameters, $p(\mathbf{h}_k^* | z_{kn} = 1, \mathbf{X}, \mathbf{x}^*)$, or to obtain the maximum-a-posteriori parameters, $\boldsymbol{\theta}_k^*$, given by

$$\boldsymbol{\theta}_k^* = \arg\max_{\mathbf{h}} p(\mathbf{x}^* | \mathbf{h}, M_k) \mathcal{S}t(\mathrm{diag}(\mathbf{s}_k)^{-1}(\mathbf{h} - \mathbf{a}_k) | n_k). \tag{25}$$

Note that for many degrees of freedom due to large values of $\overline{N}_k$, the Student distribution tends to a Gaussian with mean $\mathbf{a}_k$ and deviance matrix $\mathrm{diag}(\mathbf{s}_k)$. However, small values of $\overline{N}_k$ lead to a small number of degrees of freedom and heavier tailed distributions than Gaussians, which are more robust against outliers.

A.6 Parameters, initialization and convergence criteria

As the mean-field variational inference is an iterative framework, it also depends on the initialization of the parameters. In this section, we provide priors that do not bias the final solution and also provide some intuitive criteria for the initialization.

We initialize the parameters $\boldsymbol{\theta}_{kn}$ and $\mathbf{A}_{kn}$ by fitting all models separately to all participants (with some initial Gaussian prior), i.e., assuming as if $z_{kn} = 1$. These values are then used to calculate summary statistics according to equations (11-13).

Furthermore, we need to define prior parameters. The free parameter α_0 indicates prior frequency of each model. We take uninformative priors on frequency of models, which is given by $\alpha_0 = 1$ for all models. The prior mean, $\mathbf{a}_{0k}$, is assumed to be zero. Given equation (15), we see that b can be interpreted as the effective number of prior samples associated with models. Also, given equation (17), v could be interpreted as the half of the effective number of prior samples associated with models. Assuming that the priors account for one sample, which is a common assumption in Bayesian statistics, we take $b = 1$ and $v = \frac{1}{2}$. Finally, since s has always an additive effect on $\boldsymbol{\sigma}_k$ according to equation (16), we assume

a small positive value for s, allowing that $\boldsymbol{\sigma}_k$ to be driven dominantly by data. In all our analyses, we assumed $s = 0.01$.

Finally, the algorithm presented in A.3 requires stopping criteria. In our analyses, we terminated the algorithm if the change in normalized value of parameters between two consecutive iterations, $j-1$ and j, defined as

$$\hat{d} = \sqrt{\frac{1}{K}\sum_k \frac{1}{D_k}\sum_i (\hat{\theta}_{ki}^j - \hat{\theta}_{ki}^{j-1})^2},$$

was smaller than 0.01. Here, $\hat{\theta}_{ki}^j$ is defined according to summary statistics of parameters on the jth iteration:

$$\hat{\theta}_{ki}^j = \bar{\theta}_{ki}/\overline{V}_{ki}^{\frac{1}{2}},$$

where θ_{ki} and $\overline{V}_{ki}$ are the ith element of $\bar{\boldsymbol{\theta}}_k$ and $\overline{\mathbf{V}}_k$ defined in (12-13), respectively. In our analyses, we also set 50 as the maximum number of iterations, although almost always the algorithm stopped before hitting this number.

A.7 Exceedance probability

Using the posterior over $\mathbf{m}$, one can also derive the so-called exceedance probability and protected exceedance probability, as defined in previous works [13,18]. We reproduce the equations here for completeness.

The exceedance probability of kth model, ϕ_k, is defined as the probability that model M_k is more likely than any other model in the model-space and it is given by

$$\phi_k = \text{Prob}(m_k > m_j|\boldsymbol{\alpha}), \quad \forall j \neq k. \tag{26}$$

Computing protected exceedance probabilities, as defined in [18], also requires to run the HBI under the (prior) null hypothesis, H_0, that there is no difference between models (i.e. $\alpha_0 \to \infty$). The alternative hypothesis, H_1, is the original case, in which $\alpha_0 = 1$. If we define L and L_0 as the log-likelihood (actually the variational lower bound as its approximation) of all data given the model-space under H_1 and H_0, respectively, then the protected exceedance probability of kth model, $\tilde{\phi}_k$, is defined as:

$$\tilde{\phi}_k = \phi_k(1 - P_0) + \frac{1}{K}P_0, \tag{27}$$

where

$$P_0 = \frac{1}{1 + \exp(L - L_0)}. \tag{28}$$

Appendix B

In this appendix, we provide the proof of the results given in sections A.2 and A.3. The proof is given in three parts by obtaining 1) the functional form of $q(\mathbf{H}, \mathbf{Z})$; 2) the posterior $q(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m})$ and corresponding update equations; and 3) the update equations for the posterior over latent variables, $q(\mathbf{H}, \mathbf{Z})$.

B.1 The functional form of the posterior over H and Z

Let us first consider the derivation of the functional form for the factor $q(\mathbf{H}, \mathbf{Z})$. According to standard results in variational inference [15,16], the log of this factor is given by:

$$\log q(\mathbf{H}, \mathbf{Z}) = \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m}}[\log p(\mathbf{X}, \mathbf{H}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m})] + \text{constant}^{\backslash \mathbf{H}, \backslash \mathbf{Z}},$$

where the constant term denotes all the terms independent of the corresponding variables. Note that the expectation is taken with respect to the current estimates of $q(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m})$. By using equation (6) and absorbing all the terms which are independent of $\mathbf{H}$ and $\mathbf{Z}$ into the additive constant, we have:

$$\log q(\mathbf{H}, \mathbf{Z}) = \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\tau}}[\log p(\mathbf{X}, \mathbf{H}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\tau})] + \mathbb{E}_{\mathbf{m}}[\log p(\mathbf{Z}|\mathbf{m})] + \text{constant}^{\backslash \mathbf{H}, \backslash \mathbf{Z}}.$$

Substituting the two conditional distribution on the right-hand side using equations (1-3), we have:

$$\log q(\mathbf{H}, \mathbf{Z}) = \sum_k \sum_n z_{kn}(\log I_{kn} + \mathbb{E}_{\mathbf{m}}[\log m_k]) + \text{constant}^{\backslash \mathbf{H}, \backslash \mathbf{Z}},$$

where,

$$\log I_{kn} = \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\tau}}[\log p(\mathbf{x}_n|\mathbf{h}_{kn}, M_k)\mathcal{N}(\mathbf{h}_{kn}|\boldsymbol{\mu}_k, \mathbf{T}_k^{-1})].$$

Note that we have defined $\mathbf{T}_k = \text{diag}(\boldsymbol{\tau}_k)$. We assume that there is a quadratic approximation of I_{kn} with respect to $\mathbf{h}_{kn}$,

$$I_{kn} \propto \exp(-\frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^\top \mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})),$$

which gives,

$$\log q(\mathbf{H}, \mathbf{Z}) = \sum_k \sum_n z_{kn}(-\frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^\top \mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn}) + \mathbb{E}_{\mathbf{m}}[\log m_k]) + \text{constant}^{\backslash \mathbf{H}, \backslash \mathbf{Z}}.$$

Since $\log q(\mathbf{H}|\mathbf{Z}) = \log q(\mathbf{H}, \mathbf{Z}) - \log q(\mathbf{Z})$, we can read off terms involving $\mathbf{H}$ in $\log q(\mathbf{H}, \mathbf{Z})$ to obtain $\log q(\mathbf{H}|\mathbf{Z})$:

$$\log q(\mathbf{H}|\mathbf{Z}) = \sum_k \sum_n z_{kn}(-\frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^\top \mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})) + \text{constant}^{\backslash \mathbf{H}}.$$

Requiring that this distribution should be normalized, we obtain:

$$q(\mathbf{H}|\mathbf{Z}) = \prod_k \prod_n \mathcal{N}(\mathbf{h}_{kn}|\boldsymbol{\theta}_{kn}, \mathbf{A}_{kn}^{-1})^{z_{kn}}. \tag{29}$$

Subtracting $\log q(\mathbf{H}|\mathbf{Z})$ from $\log q(\mathbf{H}, \mathbf{Z})$ cancels out the quadratic component and yields $\log q(\mathbf{Z})$, which is a linear function with respect to z_{kn}. Therefore, we have:

$$q(\mathbf{Z}) = \prod_k \prod_n r_{kn}^{z_{kn}}. \tag{30}$$

The functional form of $q(\mathbf{H}, \mathbf{Z})$ is then given by,

$$q(\mathbf{H}, \mathbf{Z}) = \prod_k \prod_n r_{kn}^{z_{kn}} \mathcal{N}(\mathbf{h}_{kn}|\boldsymbol{\theta}_{kn}, \mathbf{A}_{kn}^{-1})^{z_{kn}}.$$

Here our goal was to obtain the functional form of the posterior over latent variables. We will obtain values of r_{kn}, $\boldsymbol{\theta}_{kn}$ and $\mathbf{A}_{kn}$ in section B.3.

B.2 The posterior over $\boldsymbol{\mu}$, $\boldsymbol{\tau}$ and $\mathbf{m}$

We continue with obtaining the functional form and update equations for the other variational factor $q(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m})$. The posterior of $\mathbf{m}$ is independent from the posterior over $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ because the log-posterior decomposes into the terms that

only depend on $\mathbf{m}$ and terms that only depend on $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$:

$$\begin{aligned}\log q(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m}) &= \mathbb{E}_{\mathbf{H},\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{H}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m})] + \text{constant}^{\backslash \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m}} \\ &= \mathbb{E}_{\mathbf{H},\mathbf{Z}}[\log p(\mathbf{H}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\tau}) + \log p(\boldsymbol{\mu}, \boldsymbol{\tau})] + \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z}|\mathbf{m}) + p(\mathbf{m})] \\ &+ \text{constant}^{\backslash \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{m}},\end{aligned}$$

where we have used equation (6). This implies that the variational posterior $q(\boldsymbol{\mu}, \mathbf{T}, \mathbf{m})$ factorizes to give $q(\boldsymbol{\mu}, \mathbf{T})q(\mathbf{m})$. Thus, the posterior over $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ is given by:

$$\log q(\boldsymbol{\mu}, \boldsymbol{\tau}) = \mathbb{E}_{\mathbf{H},\mathbf{Z}}[\log p(\mathbf{H}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\tau})] + \log p(\boldsymbol{\mu}, \boldsymbol{\tau}) + \text{constant}^{\backslash \boldsymbol{\mu}, \boldsymbol{\tau}},$$

in which we absorbed any terms independent of $\boldsymbol{\mu}$, $\boldsymbol{\tau}$ into the additive constant. Substituting for the distributions on the right-hand side, we have:

$$\begin{aligned}\log q(\boldsymbol{\mu}) = &\sum_k \sum_n \mathbb{E}_{\mathbf{H},\mathbf{Z},\boldsymbol{\tau}}[z_{kn} \log \mathcal{N}(\mathbf{h}_{kn}|\boldsymbol{\mu}_k, \mathbf{T}_k^{-1})] + \\ &\sum_k \log \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{a}_0, (b\mathbf{T}_k)^{-1}) + \sum_k \log \mathcal{G}(\boldsymbol{\tau}_k|v, \mathbf{s}) + \text{constant}^{\backslash \boldsymbol{\mu}, \boldsymbol{\tau}}.\end{aligned}$$

Note that we have defined $\mathbf{T}_k = \text{diag}(\boldsymbol{\tau}_k)$. Using equation (B.1) and by absorbing terms independent of $\boldsymbol{\mu}_k$ and $\boldsymbol{\tau}_k$ into the additive constant, we have:

$$\begin{aligned}\log q(\boldsymbol{\mu}_k, \boldsymbol{\tau}_k) = &\sum_n \frac{1}{2} r_{kn} \log |\mathbf{T}_k| - \sum_n \frac{1}{2} (\boldsymbol{\mu}_k - \boldsymbol{\theta}_{kn})^{\mathsf{T}} r_{kn} \mathbf{T}_k (\boldsymbol{\mu}_k - \boldsymbol{\theta}_{kn}) + \\ &- \sum_n \frac{1}{2} r_{kn} \text{Tr}(\mathbf{A}_{\text{kn}}^{-1} \mathbf{T}_{\text{k}}) + \frac{1}{2} \log |\mathbf{T}_k| + \\ &- \frac{1}{2} (\boldsymbol{\mu}_k - \mathbf{a}_0)^{\mathsf{T}} b \mathbf{T}_k (\boldsymbol{\mu}_k - \mathbf{a}_0) + \\ &+ \sum_{i=1}^{D_k} \log G(v, s) + (v-1) \log \tau_{ki} - s\tau_{ki} + \text{constant}^{\backslash \boldsymbol{\mu}_k, \boldsymbol{\tau}_k}.\end{aligned}$$

As the right-hand side is quadratic with respect to $\boldsymbol{\mu}_k$, the posterior over $\boldsymbol{\mu}_k$ also takes the form of a Gaussian with a variance depending on $\boldsymbol{\tau}_k$:

$$q(\boldsymbol{\mu}_k|\boldsymbol{\tau}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{a}_k, (\beta_k \mathbf{T}_k)^{-1}),$$

where

$$\mathbf{a}_k = \frac{1}{\overline{N}_k + b} (\sum_n r_{kn} \boldsymbol{\theta}_{kn} + b\mathbf{a}_0)$$

$$\beta_k = b + \overline{N}_k,$$

and N_k is given by:

$$\overline{N}_k = \sum_n r_{kn}.$$

By subtracting $\log q(\boldsymbol{\mu}_k|\boldsymbol{\tau}_k)$ from $\log q(\boldsymbol{\mu}_k, \boldsymbol{\tau}_k)$, we obtain the posterior over $\boldsymbol{\tau}_k$:

$$q(\boldsymbol{\tau}_k) = \mathcal{G}(\boldsymbol{\tau}_k|\nu_k, \boldsymbol{\sigma}_k),$$

where

$$\begin{aligned}\boldsymbol{\sigma}_k =& \frac{1}{2}\sum_n \mathrm{diag}(r_{kn}[(\boldsymbol{\theta}_{kn} - \overline{\boldsymbol{\theta}}_k)(\boldsymbol{\theta}_{kn} - \overline{\boldsymbol{\theta}}_k)^\mathsf{T} + \mathbf{A}_{kn}^{-1}]) \\ &+ \frac{1}{2}\frac{b\overline{N}_k}{b + \overline{N}_k}\mathrm{diag}((\overline{\boldsymbol{\theta}}_k - \mathbf{a}_0)(\overline{\boldsymbol{\theta}}_k - \mathbf{a}_0)^\mathsf{T}) + \mathbf{s}\end{aligned}$$

$$\nu_k = v + \frac{1}{2}\overline{N}_k,$$

and $\overline{\boldsymbol{\theta}}_k$ is given by:

$$\overline{\boldsymbol{\theta}}_k = \frac{1}{\overline{N}_k}\sum_n r_{kn}\boldsymbol{\theta}_{kn}.$$

Finally, we consider the factor $q(\mathbf{m})$:

$$\log q(\mathbf{m}) = \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z}|\mathbf{m})] + \log p(\mathbf{m}) + \mathrm{constant}^{\backslash \mathbf{m}}.$$

Substituting for the two distributions on the right-hand side, we have

$$\log q(\mathbf{m}) = \sum_k \sum_n r_{kn} \log m_k + \log C(\alpha_0) + \sum_k (\alpha_0 - 1)\log m_k + \mathrm{constant}^{\backslash \mathbf{m}}.$$

Therefore $q(\mathbf{m})$ takes the form of Dirichlet distribution:

$$q(\mathbf{m}) = \mathrm{Dir}(\mathbf{m}|\boldsymbol{\alpha}),$$

where $\boldsymbol{\alpha}$ has components α_k given by,

$$\alpha_k = \alpha_0 + \overline{N}_k.$$

B.3 The posterior over H and Z

We have already seen in section B.1 that $q(\mathbf{H}, \mathbf{Z})$ could be written,

$$\log q(\mathbf{H}, \mathbf{Z}) = \sum_k \sum_n z_{kn}(\log I_{kn} + \mathbb{E}_{\mathbf{m}}[\log m_k]) + \text{constant}^{\backslash \mathbf{H}, \backslash \mathbf{Z}},$$

where,

$$\log I_{kn} = \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\tau}}[\log p(\mathbf{x}_n | \mathbf{h}_{kn}, M_k) \mathcal{N}(\mathbf{h}_{kn} | \boldsymbol{\mu}_k, \mathbf{T}_k^{-1})].$$

Since we have already obtained $q(\boldsymbol{\mu}, \boldsymbol{\tau})$, we can now compute I_{kn}:

$$\begin{aligned}\log I_{kn} = &+ \log p(\mathbf{x}_n | \mathbf{h}_{kn}, M_k) - \frac{1}{2} D_k \log 2\pi + \frac{1}{2} \mathbb{E}[\log |\mathbf{T}_k|] \\ &- \frac{1}{2} \mathbb{E}_{\mathbf{T}, \boldsymbol{\mu}}[(\mathbf{h}_{kn} - \boldsymbol{\mu}_k)^\mathsf{T} \mathbf{T}_k (\mathbf{h}_{kn} - \boldsymbol{\mu}_k)].\end{aligned}$$

By using equation (9), $\log I_{kn}$ is given by:

$$\begin{aligned}\log I_{kn} = &\log p(\mathbf{x}_n | \mathbf{h}_{kn}, M_k) - \frac{1}{2} D_k \log 2\pi + \frac{1}{2} \mathbb{E}[\log |\mathbf{T}_k|] \\ &- \frac{1}{2} (\mathbf{h}_{kn} - \mathbf{a}_k)^\mathsf{T} \mathbb{E}[\mathbf{T}_k] (\mathbf{h}_{kn} - \mathbf{a}_k) - \frac{1}{2} \mathbb{E}[\mathrm{Tr}(\mathbf{T}_\mathrm{k} (\beta_\mathrm{k} \mathbf{T}_\mathrm{k})^{-1})],\end{aligned}$$

which can be written in the form,

$$\log I_{kn} = \log p(\mathbf{x}_n | \mathbf{h}_{kn}, M_k) \mathcal{N}(\mathbf{h}_{kn} | \mathbf{a}_k, \mathbb{E}[\mathbf{T}_k]^{-1}) + \lambda_k,$$

where λ_k is independent of $\mathbf{h}_{kn}$ and is given by

$$\lambda_k = \frac{1}{2} \mathbb{E}[\log |\mathbf{T}_k|] - \frac{1}{2} \log |\mathbb{E}[\mathbf{T}_k]| - \frac{1}{2} \frac{D_k}{\beta_k},$$

Substituting the moments of $\mathbf{T}_k = \mathrm{diag}(\boldsymbol{\tau}_k)$ with their values under $q(\boldsymbol{\tau}_k)$,

$$\log \mathbb{E}[\boldsymbol{\tau}_k] = D_k \log \nu_k - \sum_i \log \sigma_{ki},$$

$$\mathbb{E}[\log \boldsymbol{\tau}_k] = D_k \psi(\nu_k) - \sum_i \log \sigma_{ki},$$

gives equation (23).

Now, we make a quadratic approximation of $p(\mathbf{x}_n | \mathbf{h}_{kn}, M_k) \mathcal{N}(\mathbf{h}_{kn} | \mathbf{a}_k, \mathbb{E}[\mathbf{T}_k]^{-1})$ with respect to $\mathbf{h}_{kn}$ (for example using Laplace approximation or any other

method):

$$p(\mathbf{x}_n|\mathbf{h}_{kn}, M_k)\mathcal{N}(\mathbf{h}_{kn}|\mathbf{a}_k, \mathbb{E}[\mathbf{T}_k]^{-1}) \simeq \\ f_{kn}\exp(-\frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^\mathsf{T}\mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})).$$

Substituting this approximation into $\log I_{kn}$, we obtain

$$\log I_{kn} = \log f_{kn} - \frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^\mathsf{T}\mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn}) + \lambda_k.$$

Therefore, we have:

$$\log q(\mathbf{H}, \mathbf{Z}) = \sum_k \sum_n z_{kn}\Big(+ \log f_{kn} - \frac{1}{2}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn})^\mathsf{T}\mathbf{A}_{kn}(\mathbf{h}_{kn} - \boldsymbol{\theta}_{kn}) \\ + \lambda_k + \mathbb{E}[\log m_k]\Big) + \text{constant}^{\backslash\mathbf{H},\backslash\mathbf{Z}}.$$

Subtracting this equation from $\log q(\mathbf{H}|Z)$ given by the log of equation (29), we have:

$$\log q(\mathbf{Z}) = \sum_k \sum_n z_{kn} \log \rho_{kn} + \text{constant}^{\backslash\mathbf{H},\backslash\mathbf{Z}},$$

where

$$\log \rho_{kn} = \log f_{kn} + \frac{1}{2}D_k \log 2\pi - \frac{1}{2}\log|\mathbf{A}_{kn}| + \lambda_k + \mathbb{E}[\log m_k].$$

Requiring that $q(\mathbf{Z})$ be normalized, we obtain equation (30), where

$$r_{kn} = \frac{\rho_{kn}}{\sum_{j=1}^{K}\rho_{jn}},$$

which completes the proof.