Bayesian model selection and estimation: Simultaneous random effects for models and parameters

Schad, Daniel J.[1,*], Rapp, Michael A.[2], Dayan, Peter[3], Huys, Quentin J.M.[4]

1 Daniel J. Schad, Department of Cognitive Science, University of Potsdam, Potsdam, Germany; Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Berlin, Germany

2 Michael A. Rapp, Department of Cognitive Science, University of Potsdam, Potsdam, Germany; Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Berlin, Germany

3 Peter Dayan, Max Planck Institute Biological Cybernetics, Tübingen, Germany; Gatsby Computational Neuroscience Unit, University College London, London, UK

4 Quentin J.M. Huys, University College London, London, UK; Translational Neuromodeling Unit, ETH and University Zürich, Zürich, Switzerland; Department of Psychiatry and Psychotherapy, Hospital for Psychiatry and Psychosomatics, University of Zürich, Zürich, Switzerland

∗ E-mail: Corresponding danieljschad@gmail.com

Abstract

Bayesian model selection and estimation (BMSE) are powerful methods for determining the most likely among a set of competing hypotheses about the mechanisms and parameters that generated observed data. In group-studies, full inference is provided by hierarchical Bayesian models, which capture individual differences (random effects) as well as mechanisms/parameters common to all individuals (fixed effects). Previous hierarchical models have assumed random effects either for model parameters (Huys et al., 2011; Pinheiro & Bates, 2000) or for the model identity (Stephan et al., 2009). Here, we present a novel Variational Bayesian (VB) model which considers random effects for models and parameters simultaneously. As a first step, we evaluate a hierarchical method estimating random effects for parameters via expectation maximization (EM), while treating models as a fixed effect. Based on Monte Carlo simulations of reinforcement learning models of decision-making we show that the EM method efficiently recovers true effects from the data, and that it can be used to estimate GLMs at the level of individual-specific parameters. We derive model-evidences and error bars for group-level effects via importance sampling and demonstrate via simulations that this can be used for valid hypothesis tests on the model parameters. Second, we evaluate our new VB method to simultaneously consider random effects for models and parameters, and compare it to a sufficient statistics approach, where random effects for parameters and models are computed separately and combined for inference. Monte Carlo simulations show that both approaches provide successful estimation of model probabilities when uncertainty is low, but - as theoretically expected - reveal a higher correct probability mass of the new VB method under conditions of uncertainty. Compared to previous approaches (Huys et al., 2011; Stephan et al., 2009), the new VB method thus provides more precise inference in Bayesian model selection under uncertainty, and allows reducing biases in parameter estimation. Our new method suggests that we can and should understand the heterogeneity and homogeneity observed in group studies by investigating contributions of both, the underlying mechanisms and their parameters. We expect that this new random effects method will prove useful for a wide range of group studies in computational models of cognition, biology, and beyond.

Author Summary

Introduction

Model comparison and selection as well as parameter estimation are central to the scientific process as they allow testing different hypotheses about the causes of observed data [5]. Most scientific insights are based upon some kind of parameter estimation and/or model comparison, which reflect probabilistic statements about the beliefs in some hypothesis relative to other(s), given observed data. In classical (frequentist) approaches, parameter estimation and model comparison are based on the probability of observing data under a model and/or a parameter value relative to the probability under other models or parameter values. In the method of maximum likelihood estimation, the best estimate for a model parameter is the parameter value with the highest probability for generating the observed data, i.e., the highest likelihood. Similarly, the best statistic for model selection - according to the Neyman-Pearson lemma [6] - is the probability of observing the data under one model, divided by the probability under another model, which is known as the log-likelihood ratio. In classical frequentist approaches, the distributions of maximum likelihood estimates and of the log-likelihood ratio, under the null hypotheses of zero parameter values or no difference between models, are relatively easily computed for some models. Widely used examples include t- and F- statistics. In Bayesian approaches, the equivalents to maximum likelihood estimates and to the log-likelihood ratio are posterior parameter distributions and the log-evidence ratio, where the latter is commonly known as the Bayes Factor [7]. An important property of Bayesian model comparison is that it can deal with both nested and non-nested models. The frequentist approach, in contrast, can be viewed as a special case of the Bayesian approach, where for some nested models the null distributions can be easily computed.

Scientific investigations about the mechanisms and parameters generating observed data in a population of diverse individuals - i.e., a situation often encountered in the cognitive and biological sciences - make use of group studies, by obtaining repeated observations for a sample of individuals drawn from the general population [30] [1]. In such settings, hierarchical models provide a mean to consider individual differences, i.e., random effects, as well as parameters or mechanisms common to all individuals, i.e., fixed effects, within the same model. Such hierarchical models can be derived via classical (frequentist) [4], empirical Bayesian [8], or hierarchical Bayesian [3, 29] formulations. In empirical Bayes, the prior for individual-subject estimates is itself inferred from the observed data in the group [8].

Classical hierarchical models assume that individuals differ in the parameters generating observed data, while the computational mechanism underlying observed data is identical for all individuals [4,8]. The individual variance is thus captured by estimating random effects for parameters, while treating models as a fixed-effect. In this empirical Bayesian setting, group-level prior means estimate the average of the model parameter across the studied group of individuals, representing a common effect among all individuals. Prior variances estimate the variance of model parameters across individuals, capturing the variability of the model parameters in the studied group of individuals. Maximum a posterior parameters (MAP) per subject provide estimates for each individual's model parameters, thus capturing individual peculiarities. Such hierarchical models have been widely used in statistical and computational modeling of cognition, biology, and beyond [4, 8].

Recent work has proposed the complementary perspective that individuals can also differ with respect to the computational mechanisms generating observed data [3]. Such variability can be estimated via random effects for the model identity, where different posterior model probabilities are allowed for different individuals, and the distribution of these posterior model probabilities across a group is estimated via hierarchical Bayesian procedures. Stephan et al. [3] demonstrated that their hierarchical Bayesian approach to model comparison at the group level is superior to previous approaches based on classical and Bayesian approaches, and has since been widely and increasingly used in studies of (neuro-)biology and cognition [9]. As an important characteristic of this method, however, it does not consider a hierarchical model for random effects in parameters. That is, it does not estimate model parameters at the group-level, but instead relies on estimates which are previously computed, with parameter estimation usually performed for each

[1] Gibt es hier noch grundlegendere Literatur?

individual separately (e.g., [9, 27]).

In the present work, we aim to combine previous methods for random effects inference with respect to parameters [4, 8] versus models [3], and to consider the general case of estimating random effects for models and parameters in a hierarchical Bayesian model simultaneously. This new approach provides the possibility to account for the heterogeneity and homogeneity observed in group studies at both levels of the parameters and of the mechanisms generating observed data. We will consider parameter estimation and model comparison for analyses at the group level, without putting any constraints on the models compared. Thus, models can be non-linear, possibly dynamic, do not need to bear a hierarchical relationship to each other, i.e., they do not need to be nested, and they may explain data at any scale, including a continuous, dichotomous, mulitinomial, or mixed nature. Formally, our methods therefore represent generalized non-linear random effects models for parameters and models. In the present work, we apply our method to the domain of reinforcement learning models of decision-making [1, 8, 9]. However, the theoretical framework described in this work can be applied to any model, for example to cognitive models, e.g., of memory [11], semantic knowledge [10] or eye-movement control [2], to different source reconstruction methods for EEG / MEG, or dynamic causal models (DCMs) for fMRI or electrophysiological data [12, 13].

As a first step, we derive an algorithm for estimating classical hierarchical random effects models (for parameters) while treating models as a fixed-effect via Expectation Maximization (EM, [14]). We follow the work by Huys et al. [8] and again derive this algorithm in detail here to make the present paper self-contained. As a crucial next step, we outline the structure of our novel method assuming simultaneous random effects for models and parameters, and derive a variational Bayesian algorithm, which provides a set of intuitive and relatively simple update-equations constituting an efficient algorithm to obtain estimates for the posterior parameters. In the results section, we evaluate the random effects method for parameters via expectation maximization (EM), while treating models as a fixed-effect [8]. Based on Monte Carlo simulations of (generalized non-linear) reinforcement learning models of decision-making we show that the EM method efficiently recovers true effects from the data when the true model is the same for all individuals, and that it can be used to estimate general linear models (GLMs) at the level of individual-specific parameters. We derive model-evidences and error bars for fixed effects via importance sampling and demonstrate via simulations that this can be used to test hypotheses on the data. Second, we evaluate our new VB method to simultaneously consider random effects for models and parameters, and compare it to a sufficient statistics approach, where random effects for parameters [8] and models [3] are computed separately and combined for inference. Monte Carlo simulations show that both approaches provide successful estimation of model probabilities when uncertainty is low, but - as theoretically expected - reveal a higher correct probability mass of the new VB method under conditions of uncertainty. Compared to previous approaches [3, 8], the new VB method thus provides more precise inference in Bayesian model selection under uncertainty, and allows reducing biases in parameter estimation. Our new method suggests that we can and should understand the heterogeneity and homogeneity observed in group studies by investigating contributions of both, the underlying mechanisms and their parameters.

Methods

Random effects for parameters: An empirical Bayes' approach

We assume that the observed data $\mathfrak{Y} = \left\{ \{y\}_{t=1}^{T_n} \right\}_{n=1}^{N}$ from $1..N$ subjects with each $1..T_n$ obervations are i.i.d random samples from model M_k, which specifies a probability distribution for observing data $\mathfrak{Y}$ given individual subject parameters θ_n:

$$P\left(y_n \mid \theta_n\right) = M(\{y_{n,t}\}_{t=1}^{T_n}, \theta_n). \tag{1}$$

Under these very general assumptions, the model as well as its predictive probability distribution can take many forms, including different linear or non-linear functions of the parameters, as well as an arbitrary parametrized probability distribution for observed data of dichotomous, ordinal, continuous, mixed, or any other nature. For the set of individual parameters $\underline{\theta} = \{\theta_n\}_{n=1}^{N}$ we assume prior distributions with prior means μ_θ and prior variances σ_θ according to

$$P\left(\theta_n \mid \mu_\theta{,}\sigma_\theta\right) = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(\frac{\left(\theta_n - \mu_\theta\right)^2}{2\sigma_\theta^2}\right) \equiv Normal\left(\theta_n{;}\mu_\theta{,}\sigma_\theta\right). \tag{2}$$

Based on these assumptions, the probability of the data $\mathfrak{Y}$ and the individual parameters $\underline{\theta}$ given the model and the prior parameters is

$$P\left(\mathfrak{Y}{,}\underline{\theta} \mid \mu_\theta{,}\sigma_\theta\right) = P\left(\mathfrak{Y} \mid \underline{\theta}{,}\mu_\theta{,}\sigma_\theta\right) P\left(\underline{\theta} \mid \mu_\theta{,}\sigma_\theta\right) = \prod_{n=1}^{N} \prod_{t=1}^{T_n} P\left(y_{n,t} \mid \theta_n\right) P\left(\theta_n \mid \mu_\theta{,}\sigma_\theta\right). \tag{3}$$

Empirical Bayes

We use empirical Bayes' to infer estimates for the prior (i.e., the group-level parameters, including prior means and variances) and the posterior (i.e., posterior means and variances). We follow the approach by Huys et al. [8] and present the underlying derivations in more detail here. Specifically, we infer the prior from the group information by integrating over the vector of individual subject-parameters, $\underline{\theta}$,

$$P\left(\mathfrak{Y} \mid \mu_\theta{,}\sigma_\theta\right) \propto \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta} P\left(\mathfrak{Y}{,}\underline{\theta} \mid \mu_\theta{,}\sigma_\theta\right), \tag{4}$$

and prior parameters are set to their maximum likelihood estimates (MLE)

$$\hat{\mu}_\theta{,}\hat{\sigma}_\theta = \underset{\mu_\theta,\sigma_\theta}{\operatorname{argmax}} P(\mathfrak{Y} \mid \mu_\theta{,}\sigma_\theta) = \underset{\mu_\theta,\sigma_\theta}{\operatorname{argmax}} \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, P\left(\mathfrak{Y} \mid \underline{\theta}\right) P\left(\underline{\theta} \mid \mu_\theta{,}\sigma_\theta\right). \tag{5}$$

Expectation Maximization with Laplace approximation

To derive the MLE we use a reformulation of the model on log-scale (Eq. 6), introduce an approximating distribution $q(\underline{\theta})$ (Eq. 7), and make use of Jensen's inequality (Eq. 8):

$$\log p\left(\mathfrak{Y} \mid \mu_\theta, \sigma_\theta\right) = \log \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, p\left(\mathfrak{Y}, \underline{\theta} \mid \mu_\theta, \sigma_\theta\right) \tag{6}$$

$$= \log \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta})\, \frac{p\left(\mathfrak{Y}, \underline{\theta} \mid \mu_\theta, \sigma_\theta\right)}{q(\underline{\theta})} \tag{7}$$

$$\geq \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log \frac{p\left(\mathfrak{Y}, \underline{\theta} \mid \mu_\theta, \sigma_\theta\right)}{q(\underline{\theta})}. \tag{8}$$

To obtain MLE for the model parameters, we use Expectation Maximization (EM) [14], which involves iterations between an expectation step (E), to estimate the posterior distribution of latent subject-parameters (using a Laplace approximation), and a maximization step (M), to seak the MLE for the prior parameters.

E-step

For the E-step, we rewrite Equation (8) as

$$
\begin{aligned}
\log p\left(\mathfrak{Y} \mid \mu_\theta, \sigma_\theta\right) &\geq \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log \frac{P\left(\underline{\theta} \mid \mathfrak{Y},\mu_\theta,\sigma_\theta\right) P\left(\mathfrak{Y} \mid \mu_\theta,\sigma_\theta\right)}{q(\underline{\theta})} && (9)\\
&= \underbrace{\int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log \frac{P\left(\underline{\theta} \mid \mathfrak{Y},\mu_\theta,\sigma_\theta\right)}{q(\underline{\theta})}} \quad + \underbrace{\int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log P\left(\mathfrak{Y} \mid \mu_\theta,\sigma_\theta\right)} && (10)\\
&= \qquad\qquad \mathrm{KL}(P,q) \qquad\qquad + \qquad \log P\left(\mathfrak{Y} \mid \mu_\theta,\sigma_\theta\right). && (11)
\end{aligned}
$$

In Equation (11), the second term, $\log P\left(\mathfrak{Y} \mid \mu_\theta,\sigma_\theta\right)$, is exactly the log likelihood we want to determine. The first term in Equation (11), $\mathrm{KL}(P,q)$, is the Kullback-Leibler divergence between the distribution $P\left(\underline{\theta} \mid \mathfrak{Y},\mu_\theta,\sigma_\theta\right)$ and the function $q(\underline{\theta})$. Thus, our approximation (based on Eq. (11)) to the log likelihood is exact if the KL divergence between P and q is zero, i.e., if $\mathrm{KL}(P,q) = 0$, or if $q(\underline{\theta}) = P\left(\underline{\theta} \mid \mathfrak{Y},\mu_\theta,\sigma_\theta\right)$. We approximate $P\left(\underline{\theta} \mid \mathfrak{Y},\mu_\theta,\sigma_\theta\right)$ with a normal distribution and in each E-step (j) of the EM algorithm we update

$$
P\left(\theta_n \mid y_n,\mu_\theta^{(j)},\sigma_\theta^{(j)}\right) \sim q(\theta_n) = \mathit{Normal}\left(\theta_n; w_n,S_n\right). \qquad (12)
$$

We obtain the maximum a posteriori (MAP) estimate w_n via

$$
w_n \leftarrow \underset{\theta_n}{\operatorname{argmax}}\, P\left(y_n \mid \theta_n\right) P\left(\theta_n \mid \mu_\theta^{(j)},\sigma_\theta^{(j)}\right), \qquad (13)
$$

and estimate the standard deviation, S, of $q(\underline{\theta})$ as the inverse Hessian S_n^{-1} of the log likelihood:

$$
S_n^{-1} \leftarrow - \left.\frac{\partial^2 P(\mathfrak{Y}|\underline{\theta}) P\left(\underline{\theta}|\mu_\theta^{(j)},\sigma_\theta^{(j)}\right)}{\partial \theta_n^2}\right|_{\theta_n = w_n}. \qquad (14)
$$

Any method can now be used to estimate posterior parameters. We here make use of the *fminunc* function in Matlab, which implements a trust-region optimization procedure [31] employing derivatives of the log likelihood with respect to the parameters (for increased speed) to find MAP estimates for individual subject parameters w_n, and to obtain estimates of the Hessian at the MAP via numerical procedures.

M-step

Based on an estimate for the distribution of individual parameters $\underline{\theta}$, we next obtain updates for the estimates for the prior parameters. The M-step thereby aims at maximizing the log likelihood, $\log p(\mathfrak{Y} \mid \mu_\theta, \sigma_\theta)$, of the data $\mathfrak{Y}$ given the prior parameters μ_θ and σ_θ. To derive an update for the prior parameter estimates $\mu_\theta^{(j)}$ and $\sigma_\theta^{(j)}$ we compute the derivative of the log likelihood with respect to the prior parameters,

$$
\frac{\partial}{\partial(\mu_\theta, \sigma_\theta)} \log p(\mathfrak{Y} \mid \mu_\theta, \sigma_\theta). \qquad (15)
$$

We re-write Equation (8) as

$$
\begin{aligned}
\log p\left(\mathfrak{Y} \mid \mu_\theta, \sigma_\theta\right) &\geq \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log \frac{P\left(\mathfrak{Y} \mid \underline{\theta}\right) P\left(\underline{\theta} \mid \mu_\theta,\sigma_\theta\right)}{q(\underline{\theta})} && (16)\\
&= \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log\, P\left(\underline{\theta} \mid \mu_\theta,\sigma_\theta\right) + \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log\, P\left(\mathfrak{Y} \mid \underline{\theta}\right) \quad + \underbrace{\int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log \frac{1}{q(\underline{\theta})}}_{H(q)}, && (17)
\end{aligned}
$$

where $H(q)$ is the entropy[1] of q.

Interestingly, in our approximation to the likelihood in Equation (17) terms 2 and 3

$$\int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log\ P\left(\mathfrak{Y} \mid \underline{\theta}\right) + \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log\ \frac{1}{q(\underline{\theta})}$$

do not depend on the prior parameters μ_θ and σ_θ. These terms are therefore irrelevant for computing the derivative with respect to the prior parameters. We can thus simplify Equation (17) and compute its derivative with respect to μ_θ and σ_θ as

$$\frac{\partial}{\partial(\mu_\theta, \sigma_\theta)} \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log p(\underline{\theta} \mid \mu_\theta{,}\sigma_\theta) \tag{18}$$

$$= \frac{\partial}{\partial(\mu_\theta, \sigma_\theta)} \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta} \prod_n q(\theta_n) \cdot \log \prod_n p(\theta_n \mid \mu_\theta{,}\sigma_\theta) \tag{19}$$

$$= \sum_n \int \mathrm{d}\theta_n\, q(\theta_n) \cdot \frac{\partial}{\partial(\mu_\theta, \sigma_\theta)} \log p(\theta_n \mid \mu_\theta{,}\sigma_\theta). \tag{20}$$

This yields results for the partial derivative with respect to μ_θ as

$$\frac{\partial}{\partial\mu_\theta} \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log p(\underline{\theta} \mid \mu_\theta{,}\sigma_\theta) = \sum_n \int \mathrm{d}\theta_n\, q(\theta_n) \cdot \frac{\mu_\theta - \theta_n}{\sigma_\theta^2} \tag{21}$$

To find the maximum likelihood of the prior mean, μ_θ, given the current estimates for the other parameters and the approximation from the E-step, we set the derivative in Equation (21) to zero:

$$\sum_n \int \mathrm{d}\theta_n\, q(\theta_n) \cdot \frac{\mu_\theta - \theta_n}{\sigma_\theta^2} = 0 \tag{22}$$

$$\sum_n \int \mathrm{d}\theta_n\, q(\theta_n) \cdot \theta_n = \mu_\theta \sum_n \int \mathrm{d}\theta_n\, q(\theta_n) \tag{23}$$

$$\mu_\theta = \frac{1}{N} \sum_n w_n. \tag{24}$$

This result shows that the prior means (i.e., the fixed effects) are simply averages of the posterior individual subject parameters. Any differences in certainty about these individual parameters are adjusted for by the shrinkage of individual posterior parameter means to the prior group mean. Importantly, we can use this result to update our estimate for μ_θ given current estimates for w_n.

Next, we compute the partial derivative of the log likelihood (Eq. 20) with respect to σ_θ. To simplify the calculations, we take the partial derivative with respect to $(\log \sigma_\theta)$ as

$$\frac{\partial}{\partial \log \sigma_\theta} \int_{-\infty}^{\infty} \mathrm{d}\underline{\theta}\, q(\underline{\theta}) \log p(\underline{\theta} \mid \mu_\theta{,}\sigma_\theta) = \sum_n \int \mathrm{d}\theta_n\, q(\theta_n) \cdot \left(-1 + \frac{(\theta_n - \mu_\theta)^2}{\sigma_\theta^2}\right) \tag{25}$$

$$= \quad -N \quad + \frac{1}{\sigma_\theta^2} \sum_n \int \mathrm{d}\theta_n\, q(\theta_n)(\theta_n - \mu_\theta)^2 \tag{26}$$

[1] The entropy of an ensemble x is defined as $H(x) \equiv \int \mathrm{d}x P(x) \log 1/P(x)$.

We set the derivative to zero to obtain an estimate for the prior variance as

$$0 = -N + \frac{1}{\sigma_\theta^2} \sum_n \int \mathrm{d}\theta_n \, q(\theta_n)(\theta_n - \mu_\theta)^2 \tag{27}$$

$$\sigma_\theta^2 = \frac{1}{N} \sum_n \; \mathbb{E}(\theta_n^2) - (\mathbb{E}(\theta_n))^2 \tag{28}$$

$$\sigma_\theta^2 = \frac{1}{N} \sum_n \left[(w_n)^2 + S_n\right] - (\mu_\theta)^2, \tag{29}$$

where $\mathbb{E}(\theta_n^2)$ is the second moment of the normal approximation to the posterior distribution of θ_n. Interestingly, Equations (27, 28) show that the prior variance is simply the squared deviation of individual parameters θ_n from the prior mean, averaged (a) over all possible posterior values of each individual parameter (i.e., the integral over θ_n) and (b) over all N subjects (i.e., the sum over n). This result provides a formula to update estimates for the prior variance conditional on estimates for the posterior of $\underline{\theta}$ and on the prior mean μ_θ.

These results provide a simple algorithm to obtain MLE for the prior parameters (i.e., fixed effects and random effects variances) and posterior estimates for the individual-subject parameters (i.e., conditional modes and conditional variances). Given some starting values for the prior parameters, a simple algorithm now allows to iterate until convergence between

1. **Expectation step**: computing an expectation for the normal approximation to the integral over $\underline{\theta}$ (based on MAP parameters and their certainty)

$$w_n \leftarrow \underset{\theta_n}{\operatorname{argmax}} \, P\left(y_n \mid \theta_n\right) P\left(\theta_n \mid \mu_\theta^{(j)}, \sigma_\theta^{(j)}\right) \tag{30}$$

$$S_n^{-1} \leftarrow - \left. \frac{\partial^2 P(\mathfrak{Y}|\underline{\theta}) P\left(\underline{\theta}|\mu_\theta^{(j)}, \sigma_\theta^{(j)}\right)}{\partial \theta_n^2} \right|_{\theta_n = w_n}, \tag{31}$$

and

2. **Maximization step**: maximizing the log likelihood with respect to the prior parameters μ_θ and σ_θ

$$\mu_\theta = \frac{1}{N} \sum_n w_n \tag{32}$$

$$\sigma_\theta^2 = \frac{1}{N} \sum_n \left[(w_n)^2 + S_n\right] - (\mu_\theta)^2. \tag{33}$$

Given the general assumptions on which this algorithm is based, it can be applied to estimate parameters for any model M for which an approximation to the posterior - with a maximum a posteriori (MAP) estimate and a Hessian at the MAP - can be computed. Interestingly, because the prior parameters depend on the data only through the individual parameters θ_n the simple resulting equations for updating the prior and the individual parameters (i.e., Equations 32, 33) are identical for any kind of computational or statistical model. While the Laplace approximation is exact for linear gaussian models, the posterior distribution of $\underline{\theta}$ may not always be ideally approximated by a normal distribution, and other approximations may be more accurate in some situations. In the results section, we evaluate this procedure for simple models of a dichotomous dependent variable, i.e., to decide between two choice options.

Model comparison and standard errors for the fixed effects

Ideal Bayesian model comparison relies on the posterior log probability $\log P(M \mid \mathfrak{Y})$ of each model M given the observed data $\mathfrak{Y}$. We assume a flat prior on the models, reflecting the assumption that all models are equally likely a priori. For model comparison we therefore instead use the model likelihood $\log P(\mathfrak{Y} \mid M)$ of the data given each of the models. This likelihood involves integrals over parameters at the prior group and at the individual level.

Obtaining the marginal group-level log likelihood (i.e., the model evidence) is not analytically feasible for the generalized non-linear models considered here. Therefore, we approximate the group-level log likelihood at the MLE via (importance) sampling J times (with $J = 1000$) from the empirical prior distribution $\theta^j \sim p(\theta \mid \widehat{\mu}^{ML}, \widehat{\sigma}^{ML})$:

$$\begin{aligned}\log p(\mathcal{Y} \mid \widehat{\mu}^{ML}, \widehat{\sigma}^{ML}) &= \sum_n \log \int d\theta p(y_n \mid \theta_n) p(\theta_n \mid \widehat{\mu}^{ML}, \widehat{\sigma}^{ML}) \qquad (34)\\ &\approx \sum_n \log \frac{1}{J} \sum_{j=1}^{J} p(y_n \mid (\theta_n)^j).\end{aligned}$$

To approximate the group-level integral, we make use of the Bayesian Information criterion [7]:

$$\begin{aligned}\log P(\mathfrak{Y} \mid M) &= \int d\mu_\theta \int d\sigma_\theta \; P(\mathfrak{Y} \mid \mu_\theta, \sigma_\theta) P(\mu_\theta, \sigma_\theta \mid M) \qquad (35)\\ &\approx -\frac{1}{2}\mathrm{BIC}_{\mathrm{int}} = \log P(\mathfrak{Y} \mid \hat{\mu}_\theta^{ML}, \hat{\sigma}_\theta^{ML}) - \frac{1}{2}\,|M| \log(|\mathfrak{Y}|).\end{aligned}$$

Another difficulty is to obtain standard errors for the group-level parameters. Here, we use importance sampling from the group-level (prior) parameters, by shifting the likelihood with respect to one parameter each to obtain shifted likelihoods [8]:

$$LL_{shift} = \log p(\mathcal{Y} \mid \widehat{\mu}^{ML} + e_l \delta \cdot \mathrm{shift}, \widehat{\sigma}^{ML}), \qquad (36)$$

where e_l is a vector of zeros of the same dimension as μ_θ with only entry l set to one, $\delta = 0.01$, and shift $\in \{-m : m\}$ with $m = 20$. The shifted likelihoods can be quickly computed by re-weighting the J samples drawn before:

$$\log p(\mathcal{Y} \mid \widehat{\mu}^{ML} + \delta e_l \cdot shift, \widehat{\sigma}^{ML}) \approx \sum_n \log \sum_{j=1}^{J} p(y_n \mid \theta^j) w_{njs}^l \qquad (37)$$

$$\widetilde{w}_{njs}^l = \frac{p(\theta_n^j \mid \widehat{\mu}^{ML} + \delta e_l \cdot shift, \widehat{\sigma}^{ML})}{p(\theta_n^j \mid \widehat{\mu}^{ML}, \widehat{\sigma}^{ML})} \qquad (38)$$

$$w_{njs}^l = \frac{\widetilde{w}_{njs}^l}{\sum_{j'} \widetilde{w}_{nj's}^l}. \qquad (39)$$

We then perform a Laplace approximation to these shifted likelihood samples LL_{shift}. We use a finite-difference approach ***

We further test the stability of this approximation to the (potentially) noisily sampled likelihood, by estimating the second moment at its maximum via quadratic regression, with design matrix $X = [X_{int}\, X_{linear}\, X_{quadratic}]$ and component-matrices $X_{int}^T = [1...1]$; $X_{linear}^T = \delta \cdot [-m : m]$; and $X_{quadratic}^T = \delta \cdot [-m : m]^2$:

$$b = (X^T X)^{-1} X^T \cdot LL_{shift} \quad . \tag{40}$$

$$\tag{41}$$

The approximate second moment around the maximum is then obtained from the quadratic regression coefficient $b_{quadratic}$ via:

$$\frac{\partial^2 p(\mathcal{Y} \mid \mu_\theta, \sigma_\theta)}{\partial^2 \theta_l}|_{\mu_\theta, \sigma_\theta = \widehat{\mu}^{ML}, \widehat{\sigma}^{ML}} \approx 2 \cdot b_{quadratic}, \tag{42}$$

which allows computing the standard errors as

$$SE = \sqrt{\frac{-1}{2 \cdot b_{quadratic}}} \qquad . \tag{43}$$

GLM on individual subject-parameters

Simultaneous random effects for models and parameters: A variational Bayesian approach

Hierarchical Bayesian model

The current hierarchical Bayesian model draws on the work by Stephan et al. [3] on a hierarchical Bayesian model with random effects for models, but extends this work to include simultaneous estimation of random effects for model parameters. We will include K models with probabilities $r = [r_1, ..., r_k]$ which are drawn from a Dirichlet distribution:

$$p(r \mid \alpha) = Dir(r, \alpha) = \frac{1}{Z(\alpha)} \prod_k r_k^{\alpha_k - 1} \tag{44}$$
$$Z(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}.$$

Here, the $\alpha = [\alpha_1, ..., \alpha_k]$ express how often each model occurs in the population, and $\alpha_k - 1$ can be interpreted as the effective number of subjects in which model k generated the observed data.

Drawing the probability for a specific model from the model probabilities r is implemented via the index-vector $m_n = \ [m_{n,1}, ..., m_{nk}\ ]$ with elements $m_{nk} \in \{0, 1\}$, where for each individual exactly one element is 1 and all other elements are zero, i.e., $\sum_k m_{nk} = 1$. Based on the model probabilities r, a multinomial probability distribution over m_n thus describes the probability that the data by subject n were generated by model k:

$$p(m_n \mid r) = \prod_k r_k^{m_{nk}}. \tag{45}$$

We can thus sample from this multinomial distribution to obtain a particular model m_{nk} for any given subject n.

The model parameters are denoted by $\Theta = \left\{ \{\theta_{nk}\}_{n=1}^{N} \right\}_{k=1}^{K}$. The probability of the observed data y given a model m_{nk} and model parameters $\underline{\theta}_k = \{\theta_{nk}\}_{n=1}^{N}$ is

$$p(y_n \mid \theta_n, m_n) = \prod_k p(y_n \mid \theta_{nk}, m_{nk})^{m_{nk}} = p(y_n \mid \theta_{nk*}) \tag{46}$$

with $(k^* : m_{nk} = 1)$.

For each model parameter $\underline{\theta}_k$ we use a Gaussian group prior, where μ_k indicates the mean of $\underline{\theta}_k$ across subjects and σ_k^2 indicates the variance of $\underline{\theta}_k$ across subjects:

$$p(\theta_{nk} \mid \mu_k, \sigma_k^2) = Normal(\theta_{nk}; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(\theta_{nk} - \mu_k)^2}{2\sigma_k^2}) \tag{47}$$

We use standard hyperpriors for the priors μ_k and σ_k^2: For the group-level mean μ_k, the hyper-prior is again a Gaussian distribution with mean μ^0 and standard deviation ν:

$$p(\mu_k \mid \mu^0, \nu) = Normal(\mu_k; \mu^0, \nu) = \frac{1}{\sqrt{2\pi}\nu} \exp(-\frac{(\mu_k - \mu^0)^2}{2\nu^2}) \tag{48}$$

The hyperprior for the group-level variance σ_k^2 is an inverse Gamma distribution with hyper parameters for the shape a_0 and for the scale b_0:

$$p(\sigma_k^2 \mid a_0, b_0) = IG(\sigma_k^2; a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \sigma_k^{-2(a_0+1)} e^{-b/\sigma_k^2}. \tag{49}$$

Given the overall structure of the hierarchical model in Figures 2 and 1, the joint probability of the parameters and the data y can be written as (we drop some subscripts for simplicity):

$$\begin{aligned}
&p(y, r, m, \theta, \mu, \sigma^2) = p(y \mid \theta, m) p(\theta \mid \mu, \sigma^2) p(\mu \mid \mu_0, \nu) p(\sigma^2 \mid a_0, b_0) p(m \mid r) p(r \mid \alpha_0) \qquad (50)\\
&= \left[\prod_k p(r_k \mid \alpha_{0k}) p(\mu_k \mid \mu_0, \nu) p(\sigma_k^2 \mid a_0, b_0) \left(\prod_n p(y_n \mid \theta_{nk}, m_{nk}) p(\theta_{nk} \mid \mu_k, \sigma_k^2) p(m_{nk} \mid r_k) \right) \right]\\
&= \frac{1}{Z(\alpha)} \left[\prod_k r_k^{\alpha_{0k}-1} N(\mu_k; \mu_0, \nu) IG(\sigma_k^2; a_0, b_0) \left(\prod_n [p(y_n \mid \theta_{nk}, m_{nk}) r_k]^{m_{nk}} N(\theta_{nk}; \mu_k, \sigma_k^2) \right) \right]\\
&= \frac{1}{Z(\alpha)} \left[\prod_k r_k^{\alpha_{0k}-1} \frac{1}{\sqrt{2\pi}\nu} e^{-\frac{(\mu_k - \mu^0)^2}{2\nu^2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \sigma_k^{-2(a_0+1)} e^{-b_0/\sigma_k^2} \left(\prod_n [p(y_n \mid \theta_{nk}, m_{nk}) r_k]^{m_{nk}} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(\theta_{nk} - \mu_k)^2}{2\sigma_k^2}} \right) \right]
\end{aligned}$$

The log joint probability is given by:

$$\begin{aligned}
&\log p(y, r, m, \theta, \mu, \sigma^2) \qquad (51)\\
&= -\log Z(\alpha) \left[\sum_k (\alpha_{0k} - 1) \log r_k + \log N(\mu_k; \mu_0, \nu) + \log IG(\sigma_k^2; a_0, b_0) + \right.\\
&\qquad \left. \left(\sum_n [\log p(y_n \mid \theta_{nk}, m_{nk}) + \log r_k] m_{nk} + \log N(\theta_{nk}; \mu_k, \sigma_k^2) \right) \right]\\
&= -\log Z(\alpha) \left[\sum_k (\alpha_{0k} - 1) \log r_k - \frac{1}{2} \log 2\pi - \log \nu_k - \frac{(\mu_k - \mu_0)^2}{2\nu_k^2} + a_0 \log b_0 - \log \Gamma(a_0) \right.\\
&\left. -2(a_0 + 1) \log \sigma_k - b_0/\sigma_k^2 + \left(\sum_n [\log p(y_n \mid \theta_{nk}, m_{nk}) + \log r_k] m_{nk} - \frac{1}{2} \log 2\pi - \log \sigma_k - \frac{(\theta_{nk} - \mu_k)^2}{2\sigma_k^2} \right) \right]
\end{aligned}$$

Variational Bayesian approach

To invert this hierarchical model, we use a variational Bayesian (VB) approach, in which we assume that an approximate posterior density q can be described by the following mean-field factorization:

$$q(r, m, \theta, \mu, \sigma^2) = q(r)q(m)q(\theta)q(\mu)q(\sigma^2) \tag{52}$$

The terms of the form $\log q(x)$ are the variational free energies:

$$\begin{aligned}
\log q(r) &\propto \int dm \int d\theta \int d\mu \int d\sigma^2\ q(m)q(\theta)q(\mu)q(\sigma^2) \log p(y, r, m, \theta, \mu, \sigma^2) \approx \log p(r \mid y) \\
\log q(m) &\propto \int dr \int d\theta \int d\mu \int d\sigma^2\ q(r)q(\theta)q(\mu)q(\sigma^2) \log p(y, r, m, \theta, \mu, \sigma^2) \approx \log p(m \mid y) \\
\log q(\theta) &\propto \int dr \int dm \int d\mu \int d\sigma^2\ q(r)q(m)q(\mu)q(\sigma^2) \log p(y, r, m, \theta, \mu, \sigma^2) \approx \log p(\theta \mid y) \\
\log q(\mu) &\propto \int dr \int dm \int d\theta \int d\sigma^2\ q(r)q(m)q(\theta)q(\sigma^2) \log p(y, r, m, \theta, \mu, \sigma^2) \approx \log p(\mu \mid y) \\
\log q(\sigma^2) &\propto \int dr \int dm \int d\theta \int d\mu\ q(r)q(m)q(\theta)q(\mu) \log p(y, r, m, \theta, \mu, \sigma^2) \approx \log p(\sigma^2 \mid y)
\end{aligned} \tag{53}$$

To obtain the approximate posteriors $q(x) = p(x \mid y)$ (with x $\in \{r, m, \theta, \mu, \sigma^2\}$), we first have to compute the integrals in Equation (53), where we make use of the log joint probability (see Eq. (51)) and omit terms that do not depend on the variable x in question. Then, we have to find the normalizing constants or partition functions to transform the resulting functions $q(x)$ into probability density functions.

The variational free energies for the individual posterior parameters θ are:

$$\begin{aligned}
\log q(\theta) &\propto \int dr \int dm \int d\mu \int d\sigma^2\ q(r)q(m)q(\mu)q(\sigma^2) \log p(y, r, m, \theta, \mu, \sigma^2) \\
&= \sum_k \sum_n \int q(m) m_{nk} dm \log p(y_n \mid \theta_{nk}, m_{nk}) + \int \int q(\mu)q(\sigma^2) \log N(\theta_{nk}; \mu_k, \sigma_k^2) d\mu d\sigma^2 \\
&= \sum_k \sum_n q(m_{nk} = 1) \log p(y_n \mid \theta_{nk}, m_{nk}) - \frac{(\theta_{nk} - \mathbb{E}[\mu_k])^2}{2\mathbb{E}_{\sigma_k^2}[\sigma_k^2]}.
\end{aligned} \tag{54}$$

Any optimization method (e.g., trust-region, gradient descent, or others) can be used to find the maximum a posteriori estimate for $\theta_{nk}^{*\mu}$. We perform a Laplace approximation to the posterior density and compute the posterior variance $\theta_{nk}^{*\sigma}$ from the Hessian, i.e., the 2^{nd} partial derivatives of $\log q(\theta_{nk})$ with respect to the model parameters θ_{nk}.

$$\theta_{nk}^{*\mu} = argmin_{\theta_{nk}} \log q(\theta_{nk}) = argmin_{\theta_{nk}} \left[q(m_{nk} = 1) \log p(y_n \mid \theta_{nk}, m_{nk}) - \frac{(\theta_{nk} - E[\mu_k])^2}{2E_{\sigma_k^2}[\sigma_k^2]} \right] \tag{55}$$

$$p(\theta \mid y) \approx q(\theta) = \prod_k \prod_n N(\theta_{nk}; \theta_{nk}^{*\mu}, \theta_{nk}^{*\sigma})$$

$$\theta_{nk}^{*\sigma} = -\left(\left. \frac{\partial^2 \log q(\theta_{nk})}{\partial^2 \theta_{nk}} \right|_{\theta_{nk} = \theta_{nk}^{*\mu}} \right)^{-\frac{1}{2}}$$

For the prior mean parameters (i.e., the fixed-effects), the variational free energies are as follows:

$$\begin{aligned}\log q(\mu) &\propto \int dr \int dm \int d\theta \int d\sigma^2 \; q(r)q(m)q(\theta)q(\sigma^2) \log p(y, r, m, \theta, \mu, \sigma^2) \qquad (56)\\ &= \sum_k \sum_n -\frac{(\mu_k - \mu_0)^2}{2\nu_k^2 N} - \frac{\int q(\theta)(\theta_{nk} - \mu_k)^2 d\theta}{2 \int q(\sigma^2)\sigma_k^2 d\sigma^2}\\ &= \sum_k -\frac{1}{2}\left(\frac{1}{\nu_k^2} + \frac{N}{\mathbb{E}[\sigma_k^2]}\right)\left(\mu_k - \frac{\left(\frac{\mu_0}{\nu_k^2} - \frac{\sum_n \mathbb{E}[\theta_{nk}]}{\mathbb{E}[\sigma_k^2]}\right)}{\left(\frac{1}{\nu_k^2} + \frac{N}{\mathbb{E}[\sigma_k^2]}\right)}\right)^2\end{aligned}$$

This can be recognized as the log of an (unnormalized) normal distribution: $-\frac{1}{2}\frac{1}{(\mu^{*\sigma})^2}(\mu - \mu^{*\mu})^2$ with parameters

$$\begin{aligned}&p(\mu \mid y) \approx q(\mu) = \prod_k N(\mu_k; \mu_k^{*\mu}, \mu_k^{*\sigma}) \qquad (57)\\ &\mu_k^{*\mu} = \frac{\frac{\mu_{0k}}{\nu_k^2} - \frac{\sum_n E[\theta_{nk}]}{E[\sigma_k^2]}}{\frac{1}{\nu_k^2} + \frac{N}{E[\sigma_k^2]}} = \frac{\frac{\mu_{0k}}{\nu_k^2} - \frac{\sum_n \theta_{nk}^*}{b_k^*/(a_k^* - 1)}}{\frac{1}{\nu_k^2} + \frac{N}{b_k^*/(a_k^* - 1)}}\\ &\frac{1}{(\mu^{*\sigma})^2}_k = \frac{1}{\nu_k^2} + \frac{N}{E[\sigma_k^2]}; \qquad \mu_k^{*\sigma} = \sqrt{\frac{1}{\frac{1}{\nu_k^2} + \frac{N}{E[\sigma_k^2]}}} = \sqrt{\frac{1}{\frac{1}{\nu_k^2} + \frac{N}{b_k^*/(a_k^* - 1)}}}\end{aligned}$$

Next, we compute the variational free energies for the prior variances σ^2:

$$\begin{aligned}&\log q(\sigma^2) \propto \int dr \int dm \int d\theta \int d\mu \; q(r)q(m)q(\theta)q(\mu) \log p(y, r, m, \theta, \mu, \sigma^2) \qquad (58)\\ &= \sum_k \left[-2(a_0 + 1)\log\sigma_k - b_0/\sigma_k^2 + \left(\sum_n -\log\sigma_k - \frac{\int\int q(\theta)q(\mu)(\theta_{nk} - \mu_k)^2 d\theta d\mu}{2\sigma_k^2}\right)\right]\\ &= \sum_k -\left(a_0 + \frac{N}{2} + 1\right)\log\sigma_k^2 - \left(b_0 + \frac{\sum_n \left[(\theta_{nk}^{*\mu})^2 + (\theta_{nk}^{*\sigma})^2 - 2\theta_{nk}^{*\mu}\mu_k^{*\mu} + (\mu_k^{*\mu})^2 + (\mu_k^{*\nu})^2\right]}{2}\right)/\sigma_k^2\end{aligned}$$

This can be recognized as the log of an inverse Gamma Distribution,

$$\log IG(\sigma^2; a^*, b^*) = \sum_k a_k^* \log b_k^* - \log\Gamma(a_k^*) - (a_k^* + 1)\log\sigma_k^2 - b_k^*/\sigma_k^2 \qquad (59)$$

where parameters a^* and b^* are the shape and the scale parameters of the approximate posterior distribution:

$$\begin{aligned}&p(\sigma^2 \mid y) \approx q(\sigma^2) = \prod_k IG(\sigma^2; a^*, b^*) = \prod_k \frac{(b_k^*)^{a_k^*}}{\Gamma(a_k^*)}(\sigma_k^2)^{-(a_k^*+1)} e^{-b_k^*/\sigma_k^2} \qquad (60)\\ &a_k^* = a_0 + \frac{N}{2}\\ &b_k^* = b_0 + \frac{\int\int q(\theta)q(\mu)\sum_n(\theta_{nk} - \mu_k)^2 d\theta d\mu}{2} = b_0 + \frac{\sum_n \left[(\theta_{nk}^{*\mu})^2 + (\theta_{nk}^{*\sigma})^2 - 2\theta_{nk}^{*\mu}\mu_k^{*\mu} + (\mu_k^{*\mu})^2 + (\mu_k^{*\nu})^2\right]}{2}\end{aligned}$$

As an interesting next step we derive the variational free energies for models m:

$$\begin{aligned}
\log q(m) &\propto \int dr \int d\theta \int d\mu \int d\sigma^2 \; q(r)q(\theta)q(\mu)q(\sigma^2)\log p(y, r, m, \theta, \mu, \sigma^2) \\
&= \sum_k \sum_n \left[\int q(\theta) \log p(y_n \mid \theta_{nk}, m_{nk}) d\theta + \int q(r) \log r_k dr \right] m_{nk} \\
&= \sum_k \sum_n [\log p(y_n \mid m_{nk}) + \Psi(\alpha_k) - \Psi(\alpha_S)] m_{nk},
\end{aligned} \tag{61}$$

with $\alpha_S = \sum_k \alpha_k$; Ψ is the digamma function. Note that solving Equation (61) involves computing the marginal likelihood $\log p(y_n \mid m_{nk}) = \int q(\theta) \log p(y_n \mid \theta_{nk}, m_{nk}) d\theta$, which is not analytically available for the generalized non-linear models considered here. We therefore approximate this marginal log likelihood via importance sampling J times from the approximate posterior $q(\theta)$ as: $h_j \sim q(\theta_{nk})$:

$$\log p(y_n \mid m_{nk}) = \int q(\theta) \log p(y_n \mid \theta_{nk}, m_{nk}) d\theta \approx \frac{1}{J} \sum_{j=1}^{J} \log p(y_n \mid h_j, m_{nk}). \tag{62}$$

Note that this sampling-approach demands processing time, and other approximations are available (e.g., individual-subject BICs or the free energy approximation) which save computing speed at the cost of accuracy of the approximation. When performing this approximation repeatedly (see the algorithm below), a useful trick is to perform sampling of the likelihood from a proposal distribution once, and then use importance sampling and shifting of samples (see above, Section ”Model comparison and standard errors for the fixed effects”) to estimate the marginal likelihoods, which is computationally far less expensive. If more accuracy is desired, once the algorithm converges a new proposal distribution can be sampled from the current estimates, followed with another round of shifting samples. To increase accuracy, this procedure can be repeated once, n times, or until a newly sampled likelihood does not differ from the previous iteration step based on shifted samples by some small degree.

Note that the approximated free energy $\log q(m)$ is unnormalized. To obtain the approximate normalized posterior, $q(m)$, we normalize the variational free energy so that $q(m_{nk} = 1)$ is the (normalized) posterior probability that model k generated the data from subject n:

$$\begin{aligned}
q_{nk}^{unnormalized} &= \exp\left(\log p(y_n \mid m_{nk}) + \Psi(\alpha_k) - \Psi(\alpha_S)\right) \\
q(m_{nk} = 1) &= \frac{q_{nk}^{unnormalized}}{\sum_k q_{nk}^{unnormalized}}.
\end{aligned} \tag{63}$$

Last, we compute the variational free energies for the model probabilities r as

$$\begin{aligned}
\log q(r) &\propto \int dm \int d\theta \int d\mu \int d\sigma^2 \; q(m)q(\theta)q(\mu)q(\sigma^2)\log p(y, r, m, \theta, \mu, \sigma^2) \\
&= \sum_k \left[(\alpha_{0k} - 1) \log r_k + \left(\sum_n \int q(m) m_{nk} dm \log r_k \right) \right] \\
&= \sum_k (\alpha_{0k} + \sum_n q(m_{nk} = 1) - 1) \log r_k
\end{aligned} \tag{64}$$

This is the log of an unnormalized Dirichlet density:
$\log Dir(r; \alpha) = \sum_k (\alpha_k - 1) \log r_k + ...$

with parameters

$$\alpha_k = \alpha_{0k} + \sum_n q(m_{nk} = 1), \tag{65}$$

where α_{0k} are the prior parameters for the models, which we assume to be $\alpha_0 = [1...1]$ for all models considered [3].

Note that the results for all posterior parameters accord to standard solutions for hierarchical Bayesian models using either Normal distributions (see Eq. 57 & 60; REF) or multinomial-Dirichlet distributions (see Eq. 63 & 65, [3]). As a new result (see Eq. 55), however, the present model assumes that to compute the posterior individual parameters $\theta_{n,k}$, the contribution of the likelihood is weighted by the posterior probability that the individual n is actually using the model k, $q(m_{n,k} = 1)$. If individual n most likely was using model k, that is, if the posterior model probability approaches one, then Equation (55) represents standard Bayesian estimation of posterior model parameters. If individual n, however, was likely using a different model, that is, if the posterior model probability is smaller than one, then posterior parameter estimation for this individual is increasingly dominated by the prior, whereas the likelihood is continuously down-weighted, and the posterior will eventually equal the prior for posterior model probabilities of zero. This formulation critically demonstrates that observed data is not informative for estimating parameters for a model if the data was (likely) not generated by the model k in question. Importantly, this extremely simple mathematical mechanism - combined with standard results for hierarchical Bayesian modeling [3] [REF] - enables our novel method to powerfully correct parameter estimates from biasing influences, and to optimize posterior model evidences, as we demonstrate in the Results section below.

Optimization Algorithm

The above results can be implemented as an optimization algorithm which updates estimates of α, $\mu^{*\mu}$, $\mu^{*\sigma}$, a^*, b^*, $\theta^{*\mu}$ and $\theta^{*\sigma}$ iteratively until convergence. By combining Equations 55, 57, 60, 63, and 65 we get the following pseudo-code of an algorithm that gives us the parameters of the conditional densities we seek. We first initialize the hyper-parameters as[2]:

$$\begin{aligned}
&\alpha = \alpha_0 \\
&q_{nk}^{unnormalized} = \exp\left(\Psi(\alpha_k) - \Psi\left(\sum_k \alpha_k\right)\right) \\
&q(m_{nk} = 1) = \frac{q_{nk}^{unnormalized}}{\sum_k q_{nk}^{unnormalized}} \\
&\mu^{*\mu} = \mu_0 \\
&\mu^{*\sigma} = \nu \\
&a^* = a_0 \\
&b^* = b_0
\end{aligned} \tag{66}$$

Until convergence:

1. Update estimates for the maximum a posteriori model parameters $\theta^{*\mu}_{nk}$ and their standard errors $\theta^{*\sigma}_{nk}$.

[2] Here, we use standard values of the hyper-priors, namely $\alpha_0 = [1...1]$, $\mu_0 = 0$, $\nu = ?check?$, $a_0 = ?check?$, and $b_0 = ?check?$.

$$\theta^{*\mu}_{nk} = argmin_{\theta_{nk}} \log q(\theta_{nk}) = argmin_{\theta_{nk}} \left[q(m_{nk}=1) \log p(y_n \mid \theta_{nk}, m_{nk}) - \frac{(\theta_{nk} - E[\mu_k])^2}{2E_{\sigma_k^2}[\sigma_k^2]} \right]$$

$$\theta^{*\sigma}_{nk} = -\left(\left. \frac{\partial^2 \log q(\theta_{nk})}{\partial^2 \theta_{nk}} \right|_{\theta_{nk}=\theta^{*\mu}_{nk}} \right)^{-\frac{1}{2}}$$

2. Update the group prior for the model parameters μ and σ. This involves updating the posterior Normal distribution for the group-mean:

$$\mu^{*\mu}_k = \frac{\left(\frac{\mu_{0k}}{\nu_k^2 N} - \frac{\theta^*_{nk}}{b^*_k/(a^*_k-1)} \right)}{\left(\frac{1}{\nu_k^2 N} + \frac{1}{b^*_k/(a^*_k-1)} \right)}$$

$$\mu^{*\sigma}_k = \sqrt{\frac{1}{\left(\frac{1}{\nu_k^2 N} + \frac{1}{b^*_k/(a^*_k-1)} \right)}}$$

3. It also involves updating the posterior Inverse Gamma distribution for the group-variance:

$$a^*_k = a_0 + \frac{N}{2}$$

$$b^*_k = b_0 + \frac{\sum_n \left[(\theta^{*\mu}_{nk})^2 + (\theta^{*\sigma}_{nk})^2 - 2\theta^{*\mu}_{nk}\mu^{*\mu}_k + (\mu^{*\mu}_k)^2 + (\mu^{*\nu}_k)^2 \right]}{2}$$

4. Next, we update the posterior model probabilities:

$$q^{unnormalized}_{nk} = \exp\left(\log p(y_n \mid m_{nk}) + \Psi(\alpha_k) - \Psi\left(\sum_k \alpha_k \right) \right)$$

$$q(m_{nk}=1) = \frac{q^{unnormalized}_{nk}}{\sum_k q^{unnormalized}_{nk}}$$

5. and finally update the parameters of the Dirichlet Distribution over models:

$$\alpha_k = \alpha_{0k} + \sum_n q(m_{nk}=1)$$

end.

Results

In the following simulations, we always treat parameters as random effects. Several previous studies have used standard random effects methods, where the model identity is treated as a fixed effect and random effects for parameters are estimated via the expectation maximization algorithm (EM) [8]. As a first step, we here evaluate the validity of this procedure for reinforcement learning models of decision-making. Second, we evaluate our novel variational Bayesian procedure estimating random effects for models and parameters simultaneously and compare it to a sufficient statistics approach, where random effects for parameters [8] and models [3] are computed separately and combined for inference. To evaluate and compare these methods, we rely on synthetic simulated data, and also on real observed data. The real data have been previously published and have been analysed in various ways, including an EM estimating random effects for parameters and group level model inference using $\mathrm{BIC}_{\mathrm{int}}$ [1].

Models as fixed effects (Expectation Maximization)

Parameter estimation

To evaluate the validity of the expectation maximization (EM) algorithm for estimating random effects for parameters of reinforcement learning models of reward learning and decision-making, we used simulated data where the true decision-model is known. Specifically, for the first set of analyses, we simulated data from a simple reinforcement Q-learning model (i.e., the Rescorla-Wagner model, REF) with one state and two possible actions, with

$$Q_{t+1}(s_{s+1}, a_{t+1}) = Q_t(s_t, a_t) + \alpha \cdot \mathrm{RPE}_t, \quad (67)$$

and

$$\mathrm{RPE}_t = R_t - Q_t(s_t, a_t), \quad (68)$$

where $Q_t(s_t, a_t)$ represents the expected value Q of taking action a in state s at trial t, α is a free learning rate parameter, R represents the reward obtained after taking an action a, and the actions $a \in A = \{1, 2\}$ probabilistically lead to reward with equal and fixed probabilities $p_R = \{0.5, 0.5\}$. At choice, actions were selected based on the Q values according to the softmax function

$$P(a = \mathbf{a} \mid s_t, Q_t(s_t, a_t)) = \frac{e^{\beta Q_t(a, s_t)}}{\sum_{a'} e^{\beta Q_t(a', s_t)}} \quad (69)$$

with inverse noisiness or exploration parameter β. Model parameters were transformed to unbounded space via logistic, $\alpha = 1/(1 + \exp(-a))$, and exponential, $\beta = \exp(b)$, transformations for simulation and for fitting to impose constraints and to conform to normal distribution assumptions.

From a theoretical perspective, the precision of parameter estimates should increase with the number of data points entering the estimation. Here, we tested this expectation for our EM algorithm. We repeatedly simulated choice data from the reinforcement learning model described in the previous section and fitted the EM to the simulated data. For the simulations, we randomly sampled learning rate parameters for individual subjects from a group-level normal distribution with mean $\mu_a = -0.5$ (reflecting a prior mean for the learning rate in model space of $\mu_\alpha = 0.38$) and standard deviation of $\sigma_a = 0.5$ (indicating the true random-effects variance), and independently sampled β parameters for individual subjects from a normal distribution with mean $\mu_b = 1.5$ (reflecting a true prior mean in model space of $\mu_\beta = 4.48$) and standard deviation $\sigma_b = 0.25$ (indicating the true random-effects variance for the β parameter). Across simulation runs we varied the amount of simulated data, with the number of simulated trials per subject ranging from 60 to 1200 and with the number of simulated subjects ranging from 5 to 100. To test limit properties of our estimators for the means, variances, confidence intervals and precisions, we used repeated simulations for each combination of number of trials and number of subjects.

It is visible in Figure 3a+b that the group-level model parameters for the means (prior means) and variances (prior variances) are well recovered from the simulated data, as both, means and variances of group-level parameters closely corresponded to the true generating values. We observed some bias for estimates of the prior variance of the learning rate, which tended to be slightly deflated for small data sets. Furthermore, as theoretically expected the confidence intervals for the prior means continuously decreased with increasing numbers of data points, indicating that the estimates exhibit normative properties. More specifically, the precision of group-level mean parameter estimates (see Figure 3c+d) continually increased with increasing numbers of trials and subjects in a regular and expected manner.

As an interesting finding, precision of the group-level means increased more strongly with increasing number of subjects rather than with increasing number of trials. Even a very large number of trials yielded low precision for the prior means when the number of individuals was small, while to the contrary, a large number of individuals increased precision even when the number of trials per individual was small. This result interestingly illustrates the utility of testing large samples with many imdividuals to yield reliable findings in empirical group-studies, or may be seen to advocate the use of Bayesian methods to deal with the uncertainty involved in studies employing smaller sample sizes.

One difficulty in generalized non-linear hierarchical models is that confidence intervals for the group-level parameters are difficult to obtain. We here use the procedure proposed in Huys et al. [8] by using importance sampling to estimate the likelihood function for the group-level parameters, and performing a Laplace approximation to this sampled likelihood to obtain confidence intervals for the mean parameters at the group-level. Critically, however, it is unclear from this previous work (i) whether the Laplace approximation is adequate for the type of reinforcement learning models discussed here and (ii) whether our sampling scheme, which involves a finite difference approximation of the curvature, generates reliable estimates of the Hessian at the maximum likelihood estimate to obtain reliable and valid confidence intervals. Said differently, these two assumptions [8] propose that the sampled likelihood is approximately normally distributed, and that our sampling scheme generates estimates of the likelihood which are sufficiently reliable and do not strongly suffer from noise such that it is possible to derive valid and reliable parametric error bars via finite a difference approximation [8].

Here, we test these assumptions by investigating the surface of the log likelihood of the group-level mean model parameters (i.e., the priors), to check for approximate normality and for reliability. To this end, we chose representative exemplars from the simulations reported in Figure 4. We generated estimates for the log likelihood at a range of different points close to the maximum likelihood estimates via the described importance sampling scheme. If the likelihood is approximately normally distributed this would be visible in a quadratic function on the log-scale, which we use for plotting (see Figure 4). Moreover, reliable estimates would be visible in that individual estimates for the log-likelihood show little departure from a quadratic function fitted to these estimates.

The results from this analysis clearly support both assumptions in the procedure: It is visible in Figure 4 that for different numbers of subjects and for different numbers of trials the individual estimates of the likelihood (i.e., points in Figure 4) closely correspond to quadratic functions fitted through these estimates (i.e., lines in Figure 4). This close correspondence demonstrates that - in the studied parameter range - the Laplace approximation for the group-level likelihood is appropriate, and that individual importance samples are sufficiently reliable as a basis for obtaining error bars. Importantly, the simulations also exhibit the normative behaviour that the log-likelihood becomes increasingly peaked for larger amounts of data points, i.e., yielding a stronger curvature of the quadratic function for larger numbers of trials or subjects, reflecting the simple relation that precision for the parameter estimates increases with increasing amounts of data.

GLM on individual subject parameters

An important goal of computational modeling endeavours in group-studies is to investigate how individual differences, i.e., factors or covariates at the level of the individual, impact on estimates for specific model parameters. Examples include any designs including between-individual manipulations, like training condition, quasi-experimental variables like sex, age [**?**], or psychiatric disease status (REF), or measurements at the level of the individual, like measures of intelligence [1], personality, or neurophysiological measurements [32]. The present formulation of the hierarchical model allows implementing a general linear model to compare model parameters between groups of individuals or to regress model parameters on individual-based predictor variables (cf. Huys et al., 20??). Again, questions emerge as to whether the parameter estimates for this procedure are unbiased and whether the procedure for obtaining confidence intervals allow valid significance tests for hypotheses on how model parameters differ between individuals or between groups of individuals. In the following, we report simulations to evaluate this procedure for estimation and testing model parameters.

We simulated data from the simple reinforcement learning model introduced above (see Section *Parameter estimation*) under the assumption of a true positive influence of a between-subject covariate on the β parameter with a regression coefficient of +0.5 and a true negative influence of a between-subject covariate on the learning rate parameter with a regression coefficient of -0.5. We fitted the hierarchical model for the simulated data to evaluate parameter estimation and confidence intervals. Reported parameter estimates and confidence intervals are means across repeated simulations to test limit behavior. We found that estimates for the regression coefficients closely corresponded to the true generating values (see Figure 5a), with a minimal bias towards zero. This bias may result from extreme parameter values for some individuals, which are not sufficiently constrained by the limited data used for the present simulations. Again, confidence intervals were continuously reduced for larger amounts of (simulated) data. Estimates for the group-level means and variances remained intact.

We also evaluated the procedure for obtaining confidence intervals for the group-level regression coefficients combining importance sampling of the likelihood with a Laplace approximation. The results showed (see Figure 5b) that importance samples from the likelihood were highly stable and - on a log-scale - closely corresponded to a quadratic function, which supports the Laplace approximation for the present analysis.

Given the important role of hypothesis tests assessing between-individual or between-group effects in hierarchical models, we moreover estimated the ratio of type I errors, α, for our procedure. To this end we simulated data from the simple reinforcement learning model for $N_{sim} = 1,000$ times under the null hypothesis of no influence of the individual-difference covariate on the model parameters, fitted the model to each simulated data set, and computed standard errors and associated p-values for significance tests on the covariates. Across the 1,000 simulations, the observed rate of false positives concerning statistical tests for an influence of the covariate on the model parameters β and learning rate were .048 and .059, and thus closely corresponded to the nominal significance-level of $\alpha = .05$. More generally, it is visible in Figure 6 that expected and observed p-values were closely in aligned.

In summary, the present simulations demonstrate that the expectation maximization algorithm (EM), together with importance sampling and the Laplace approximation, provide an unbiased and valid approach for estimating hierarchical generalized nonlinear models with random effects for model parameters for reinforcement learning models of decision making, and for testing hypotheses on between-individual effects on the model parameters.

Model selection

For the standard EM framework treating models as a fixed effect we use Bayesian model comparison based on BIC_{int} [7]. To perform a test of model comparison we use models of higher complexity, namely the dual-control model for decision-making data by Daw et al. [9], which assumes two distinct model-free versus model-based systems for decision-making, and an additional component assuming a tendency to repeat choices from the previous trial. For the precise model formulation see Schad et al. [1]. For the present purpose, we simulated choices for $N_{subj} = 30$ subjects and $N_{trials} = 201$ trials in a two-step sequential decision task [9] using six different combinations of the three components (i) model-free RL, (ii) model-based RL, and (iii) choice repetition (for the realized combinations see Table 1). Group-level mean model parameters values were taken from Daw et al. [9], and were kept identical across different models. For each of the six simulated data sets, we fit each of the six models to this data, treating all parameters as random effects across subjects. The results (see Table 1) showed that for each of the six simulated data sets the true generating model was supported by the $\mathrm{BIC}_{\mathrm{int}}$, demonstrating that the procedure was able to extract the true generating model from the choice data. These results suggest that the $\mathrm{BIC}_{\mathrm{int}}$ allows for valid model comparison when all subjects in a given sample are using the same computational mechanism.

To summarize, the present simulations demonstrate that our approach to hierarchical generalized non-linear models via the Expectation Maximization (EM) algorithm, importance sampling, and Laplace approximation successfully recovers the true group-level parameters from the data, provides valid estimates for confidence intervals for the prior mean parameters, and allows to extract the true model structure from observed data via an approximation to Bayesian model evidence treating models as a fixed effect, i.e., $\mathrm{BIC}_{\mathrm{int}}$. These results demonstrate that hierarchical models with random effects for parameters but fixed-effects for the model are adequate for situations where the data from all individuals are generated from the same computational mechanism, i.e., if all subjects are using the same model, and provide a valid and valuable tool is such circumstances.

Models as random effect (Variational Bayes)

In the following, we study the situation where not only the model parameters, but also the computational mechanism generating observed behavior differs between individuals. We performed simulations where different individuals used different models of learning and decision-making, and then tried to identify the computational mechanism from the behavioral data. To this end, we use our novel full variational Bayesian method for estimating random effects for models and for parameters simultaneously. We compare this to a sufficient statistics approach, where random effects for parameters [8] and for models [3] are computed separately and combined for inference. Specifically, in the sufficient statistics approach we first used the hierarchical method described above (see *methods* section) to estimate random effects for parameters for each model for the whole sample of all individuals, and to obtain the marginal likelihoods for each model and each individual (i.e., the sufficient statistics). We then use the marginal likelihoods to estimate random effects for models using the algorithm by Stephan et al. [3], and we determine the probability mass for extracting the correct model from the behavioral data - averaged across individuals. As a last step, estimates for model parameters are adjusted for the individual posterior model probability using Bayesian parameter averaging [3]. In this procedure, group-level average parameters (prior means, μ_k) for each model k are estimated via a weighted combination of individual-subject estimates (conditional modes / maximum a posteriori estimates, $\theta^{*\mu}_{nk}$), with the weighting determined from the posterior model probabilities for each individual, $q(m_{nk})$:

$$\mu_k = \frac{1}{\sum_{n=1}^{N_s} q(m_{n,k})} \sum_{n=1}^{N_s} \theta_n \cdot q(m_{n,k}) \tag{70}$$

In the following sections, we first report simulations using the simple reinforcement learning model described in Section *Models as fixed effects* comparing simulated subjects with reward learning against simulated non-learners. Next, we extend this procedure to the more complex two-step model [9] to compare a set of three different mechanisms for decision-making based on model-based learning, model-free learning, and non-learning.

Posterior estimation under certainty: A simple reinforcement learning model

We expected that a particular advantage of our novel variational Bayesian algorithm (compared to the approach assuming fixed effects for the model) should be that it should be better able to provide adequate, i.e., unbiased, estimates of the group-level model parameters in the presence of heterogeneity in the generating mechanisms. Accordingly, such improved group-level estimates should provide better constraints for individual posterior parameter estimation, and thus improve the computation of individual posterior model probabilities. Group-level estimates or (empirically derived) priors are particularly important and effective in constraining posterior estimates at the individual level in the presence of uncertainty, which occurs e.g. when little data is available to constrain parameter estimates for each individual. To the contrary, prior or group-level information should play less of a role when sufficient data is available for each individual to constrain individual-level parameter estimates. This reasoning suggests that all methods should be well able to identify the correct model from observed data when uncertainty is low, e.g., in the case of many trials per subject and simple computational models with few parameters. To the contrary, efficient estimation of group-level prior parameters in our novel VB method should put it at an advantage when estimating posterior probabilities under conditions of uncertainty.

In the following, we tested this expectation using simulations from two simple models of decision-making: For one set of $N_s = 60$ subjects we assumed model-free learning in a reward-based decision-making paradigm and simulated choice data from the simple reinforcement learning model described in Section *Models as fixed effects*. Moreover, we assumed that a second set of $N_s = 30$ subjects did not learn from the reward-feedback provided (*Non-Learners*), and instead chose options randomly, with some tendency to repeat the action from the previous trial. In a first set of simulations, we simulated choice data for $N_t = 200$ trials per subject. We expected that this amount of data should provide good constraints for the model identity within individual subjects, as the two decision models exhibited an extremely simple model structure.

The results from these analyses (see Figure 7) showed that the average correct posterior probability mass was virtually one for the full Variational Bayes method and also for the sufficient statistics approach, demonstrating that the true model identity was extracted from the observed data with high certainty for all individuals. Moreover, estimation of model parameters (see Figure 7, lower panels) revealed that treating models as a fixed effect in a situation where the underlying computational mechanisms differ between subjects causes a divergence of estimated model parameters (see Figure 7, green crosses) from their true generating values (see Figure 7, black dots), reflecting biases in the parameter estimation. Specifically, in the present setting the estimate of the inverse noisiness parameter in the learning model was heavily biased downwards falsely suggesting a highly noisy decision-process, while in fact a group of (simulated) subjects exhibited deficits in learning - rather than in decision-making. At the same time, in the non-learning model the estimate for the parameter capturing mere choice repetition was biased upwards when treating models as a fixed effect, suggesting that learning - which was present in some subjects - was misinterpreted as contingency-free choice stickiness. These biases in parameter estimation could be overcome by either of the two approaches assuming random effects for the model identity (i.e., full VB and sufficient statistics), leading to accurate estimates of the group-level means. These results suggest that treating models as a fixed effect despite the presence of heterogeneity in the underlying computational mechanisms leads to biases in the estimation of model parameters based on classical approach estimating random effects for parameters. Moreover, they show that both methods treating models as a random effect could successfully recover the true model from the data and estimate model-specific parameters under conditions of certainty, i.e., when a lot of observed data is available for each individual and when the models are well constrained by these observations.

Posterior estimation under uncertainty

We next [3] tested how posterior model probabilities would change under conditions of uncertainty, i.e., if the observed data partially constrain the underlying mechanisms. We performed the same simulations as in the previous paragraph, but only used a small number of $N_t = 20$ trials per simulated subject. With such little data available to constrain individual subject-estimates, the prior group level should be important to efficiently pool information across subjects, revealing an advantage of the full variational Bayesian approach over the sufficient statistics procedure.

The results (see Figure 8) showed that - as expected - the average correct posterior probability mass was approximately at .86, i.e., rather low for the sufficient statistics approach, but could be improved to .90 using the full variational Bayesian algorithm. At the group-level, it was visible that the sufficient statistics approach showed inflated estimates for the number of subjects using the learning-model, but under-estimated the number of non-learners in the simulated data. Moreover, estimation of model parameters was again biased for the fixed-effects approach to the model identity, yielding inflated estimates for choice stickiness in the non-learning model, and deflated parameter estimates for the learning parameters. Given the sparse amount of data for each subject and the large resulting uncertainty, however, treating models as a random effect failed to correct estimation biases for the non-learning model, but somewhat improved parameter estimation for the learning model (see Figure 8).

[3] 2014May12-fit r ab ntr20g 2014May12 doprior0

[4] Simulation: 2014May22-fit 2step i 2014May22

The dual-control model

As a next step [4], we extended our simulations to a more complex model of decision-making [9], which has recently attracted much attention in the literature [1, 15–17]. The dual-control model [9] assumes distinct systems for reward-learning and decision-making, including (*i*) a model-free (or habitual) choice system, that learns values of choice options via a simple temporal difference learning algorithm, (*ii*) a model-based (or goal-directed) choice system, which plans ahead to anticipate future outcomes of current actions to compute choice value on the fly, and (*iii*) a separate contribution of choice stickiness, i.e., a tendency to repeat previous actions independent of reward. We performed simulations of variants of the dual-control model to generate choice behavior in a two-step sequential decision task, which has been designed to disentangle contributions of separate model-based versus model-free systems to choice behaviour [9]. In this task, in an initial (first-stage) state, selecting one of two options probabilistically causes a transition to one of two second-stage states, according to common or rare transitions. In the second-stage state, another choice causes transition to a final state with probabilistic delivery of reward, and with the reward probabilities randomly changing over the course of the task, which fosters ongoing learning across all $N_t = 201$ trials. Model-free versus model-based learning algorithms can be distinguished in this task: model-free learning does not consider the transition between states, whereas model-based control enacts a cognitive model of the task structure to guide choices.

We simulated data from three groups of subjects (each $N_s = 30$) performing this two-step task, and for each group assumed different underlying computational mechanisms: (a) the dual-control model including all three model-components of model-free choice, model-based choice, and choice stickiness; (b) a purely model-free learning model, which also included choice stickiness, with different learning parameters as compared to the full model; (c) non-learners with no reward learning, but a tendency to repeat previously selected options. We then fit random effects for models and parameters to this simulated data and compared model estimates to the true generating values. Note that the increased complexity of the computational models (containing up to 7 model parameters) together with the higher number of compared models should overall increase uncertainty in the estimation, despite the relatively high number of trials, and the improved performance of the full VB is therefore expected to extend to this modeling case.

Model parameters for the dual-control model include inverse noisiness parameters at first and second task stages ($b1$, $b2$), learning rates at first and second stages ($a1$, $a2$), eligibility parameter (l), moderating stage-skipping updates of first-stage values by second-stage prediction errors, choice stickiness parameters at first and second stage ($r1$ or r, $r2$), and a mixing-weight parameter, w, which determines the relative influence of model-free versus model-based systems on action choice.

As a main result (see Figure 9), the benefit of the full VB method to extract the true underlying computational mechanism from observed behavioral data with higher accuracy as compared to the sufficient statistics approach generalized to this more complex situation, and was even more pronounced compared to the simple RL models treated above: While the average correct probability mass ranged between .75 and .90 for the sufficient statistics approach, the full VB method reached average correct probability masses of .95 and above for the same data. Likewise, results for model comparison at the group-level showed that the sufficient statistics-approach yielded inflated estimates for the number of subjects using the full dual-control model (see Figure 9b), but under-estimated the number of non-learning subjects. Moreover, a fixed-effects approach to the model identity again yielded strong biases in the estimation of model parameters, which were most pronounced for the dual-control model, and for the non-learners model. For example, the fixed-effects approach under-estimated the critical parameter w, thus suggesting a lower degree of model-based relative to model-free choice control. Only some of the biases could be reduced by treating models as a random effect, with improved estimation for the sufficient statistics approach [3, 8] and for the full variational Bayesian algorithm (see Figure 9). However, as above and to little surprise, for either method some of the uncertainty in the present analysis could not be overcome, such that the limited behavioral data did not fully constrain all computational model parameters.

Further increasing uncertainty

In an additional step [5], we aimed at further increasing uncertainty about the mechanisms and parameters that have generated observed data to compare inference methods. To this end, we repeated the simulations reported in the previous section (*The dual-control model*) using only 101 trials in the two-step task for each individual subject. Again, we expected that the sufficient statistics approach should suffer more strongly from the uncertainty, while the hierarchical structure of the full VB method may increase robustness. The results (see Figure 10) supported this expectation. Indeed, the sufficient statistics approach strongly suffered from the reduced amount of data available for each subject, yielding average correct probability masses between .55 (for the full dual-control model) and .80 (for the non-learning model). To the contrary, the correct probability mass from the full VB method remained its high accuracy of nearly .95 for all models (see Figure 10, left panel). The sufficient statistics approach also failed to correctly estimate the number of subjects using each model, with a bias of inflated estimates for the hybrid model (MF + MB) and deflated estimates for the model-free learner (MF) and for the non-lear models. These biases were reduced by the full VB method (see Figure 10b). At the same time, estimation of model parameters strongly suffered from the reduced amount of available data, yielding stronger biases in parameter estimation. In all methods residual bias remained given the weak observational basis. However, at least some of these biases could be overcome via simultaneous random effects for models and parameters (e.g., see parameters in the non-learner model).

Real experimental data

We next [6] applied the novel random-effects method to real experimental data from a previously published study [1], where choice data from 27 subjects performing the two-step task was analyzed fitting the dual-control model via random effects for parameters but fixed effects for the models. For this data, Bayesian model comparison based on $\mathrm{BIC}_{\mathrm{int}}$ provided support for the full dual-control model with model-based and model-free choice strategies. Here, we fitted the three models from the previous two sections (i.e., the dual-control model, a model-free learning model, and a non-learning model) via simultaneous random effects for models and parameters to the empirical data. Consistent with our previous analyses [1], the results based on the full VB method (see Figure 11) showed that the full dual-control model was most common among the three tested models, with an estimated 15/27 subjects using the dual-control model, whereas estimates for the model-free learning model (7/27 subjects) and the non-learning model (8/27 subjects) were considerably lower. Likewise, the sufficient statistics approach yielded similar results, except that it identified somewhat more subjects using the dual-control models, and less non-learners.

In the estimation of model parameters, using simultaneous random effects for models and parameters strongly changed the estimated model parameters for the non-learners and for the model-free model as compared to the fixed-effects approach to models. Specifically, for the model-free model (MF) the learning rates were higher for the random effects approaches compared to the fixed-effects approach, and for the non-learners, the repetition rates were decreased.

One [7] difficulty with evaluating computational methods on real data is that the true models and parameters generating the observed data are unknown. Therefore, and to more thoroughly evaluate the models with simultaneous random effects for models and parameters, we used the posterior estimates for the model identities and the model parameters derived using the full VB on the real data to simulate data on the two-step task, and then again fitted the hierarchical models to this simulated data to confirm that the methods are successful for a range of parameters directly obtained from a real empirical data set. The results (see Figure 12) again showed an advantage in Bayesian model selection of the full VB algorithm over the sufficient statistics approach, revealed by a high average correct probability mass with values (for the dual-control and the model-free choice models) ranging between $P_{average,c} = .85$ and $P_{average,c} = .95$ for the full VB as compared to values of between $P_{average,c} = .75$ and $P_{average,c} = .90$

[5] Simulation: 2014May22-fit 2step k 101a 2014May22

[6] basicData Pilot3 WSTraw 2014May27-fit 2014May27

[7] 2014May29-fit 2step Pilot 2014May29

for the sufficient statistics approach. The results for the group-level, estimating the number of subjects using each model, were again rather biased for the sufficient statistics approach, over-estimating model-free (MF) learners. This bias could again be improved based on the full VB method. In the estimation of model parameters, fixed-effects for the models again caused biases in the estimation of model parameters, which could be reduced using simultaneous random effects for models and parameters.

Note that throughout the reported simulations, we used sampling to invert the computational models, i.e., to obtain the marginal likelihoods (see Eq. 62). We preferred this computation-intensive method over simpler approaches based on subject-based BIC in order to increase accuracy at the individual-subject level. We also performed a few simulations, inverting models via subject-based BIC scores. While this clearly reduced processing time, posterior accuracy was markedly reduced, supporting the use of the (importance) sampling procedure.

As a last check, we tested how the modelling approaches assuming random effects for the model would perform when all subjects' data are actually generated by the same model. We again used the simple reinforcement learning model used above to simulate subjects' data. In the model estimation, we moreover allowed a non-learner model to explain the data. We used 20 trials only to have a situation with sufficient uncertainty. The results presented in Figure 13 show that, despite the uncertainty due to the small number of trials, both random effects methods were well able to determine the correct model from the data for most subjects, suggesting that the methods may not be too sensitive towards violations of their underlying assumption that multiple models are present in the population. However, results may vary when competing models, which are not true for the observed data, are more confounded with the true underlying model.

Discussion

In the present work, we have evaluated a classical hierarchical models (based on the Expectation Maximization algorithm, EM, [14]) treating computational models as a fixed effect [8] and introduced a novel approach for model estimation and selection in group-studies. In the novel approach, we have combined previous approaches treating either parameters [4,8] or models [3] as random effects and developed a variational Bayesian method for estimating random effects for models and parameters simultaneously. For the classical hierarchical approach, we demonstrate via simulations that it provides valid estimates of model parameters, confidence intervals, and measures for fixed-effects Bayesian model comparison when all individuals use the same computational mechanism. However, we also demonstrate via simulations that it performs sub-optimally when individuals differ with respect to the underlying mechanisms generating observed data. We provide provisional evidence suggesting that these shortcomings can be overcome by our novel method as it allows to more precisely quantify the belief that a given model is more likely than any other in a group of individuals, yielding a higher power to determine the correct model and the correct model parameters for each individual subject and the number of subjects using a particular model, relative to an approach which estimates random effects for parameters versus models separately and combines them for inference. Critically, our variational Bayesian approach rests on treating the model parameters θ_n and the model switches m_n as random variables, within a full hierarchical model for multi-subject data (see Fig. 1), and thus accommodates random effects at the between-subject level by considering contributions of the mechanisms and the parameters causing observed data.

We demonstrate via simulations that our approach to hierarchical generalized non-linear inference with models as a fixed effect is successful when the observed data by all individuals are generated by the same underlying computational mechanism. However, further simulations showed that when different individuals' data is generated by different computational mechanisms, then treating the model as a fixed effect under conditions of uncertainty yields biases in parameter estimation and impoverishes the posterior model probabilities. Such biases and inaccuracies can be reduced by estimating random effects for models and for parameters simultaneously, thereby considering two sources contributing to individual differences, namely the mechanisms and the parameters generating observed data. Our simulations showed that under conditions of certainty, i.e., when enough data was available for each individual subject to constrain the mechanisms and parameters that have generated the observed data (i.e., simple RL model with 200 trials), all methods were successful in selecting the correct models from the data and in correcting biases in parameter estimation. However, when posterior uncertainty was present due to insufficient data to constrain models and parameters within individual subjects (i.e., simple RL model with 20 trials and all models of the two-step task), then simultaneous estimation of random effects for models and parameters consistently increased the chance to infer the correct model for each individual and the correct number of individuals using each model. We found that parameter estimation exhibited biases in fixed-effects estimation, but that the a sufficient statistics approach and a full VB method could sometime correct estimates of model parameters from observed data. This increase in validity and power is expected for the novel full VB method based on the pooling of information across the group of all individuals to inform individual estimates not only of parameters or models, but rather to pool information related to both sources of variation simultaneously to inform their covariation. Specifically, this approach allows using corrections for the parameter estimates to inform model inversion, and thus to avoid over-fitting a model's parameters to an individual's data, which were likely not generated by this model.

Hierarchical (generalized) linear models, treating models as a fixed effect, are increasingly used as a tool for statistical testing in the cognitive and neuro-sciences [9,18–21]. A steadily increasing number of studies also makes use of these hierarchical methods in the context of non-linear models of cognitive or neuronal functioning, i.e., using hierarchical generalized non-linear models, which treat models as fixed-effect - e.g., in the field of reward-based learning and decision-making, with a focus on reinforcement learning models [9,17,23] - or in models of neuroimaging data [24], and the application of this methodology is increasingly spreading to become the standard in the field of decision-making and cognitive modeling more generally. Expectation Maximization (EM, [14]) together with parametric approximations (i.e., Laplace approximation) to solve integrals over individual-level model parameters provides an efficient algorithm to estimate such models [25].

Here, we evaluated the EM approach to hierarchical generalized non-linear random effects models in the context of simple reinforcement learning and choice models and demonstrate that it is efficient and accurate in this domain. In particular, the simulations showed that the EM exhibits unbiased estimates of the true generating model parameters at the group-level and that the precision for group-level estimates scales with the number of individuals and trials as theoretically expected. More specifically, we probed the form of the likelihood function for group-level parameter estimates around its maximum and found it to closely approximate a normal distribution, supporting the use of the Laplace approximation to the group-level likelihood to compute approximations to the model evidence (i.e., $\mathrm{BIC}_{\mathrm{int}}$) and to compute confidence intervals for the group-level means. We further tested the method via simulations by investigating whether the supported statistical tests provide valid and reliable results. To this end, we repeatedly ($N_{sim} = 1,000$) simulated a hierarchical model for a simple RL model of decision-making under a null distribution and fitted the model testing influences of a between-subject covariate on the model parameters via z-tests. These simulations showed that the statistical test reached significance ($p < .05$) at nominal rates of .048 and .059 (for the inverse noise and learning rate parameters), and that the distribution of empirical p-values overall closely corresponded to theoretical expectations. Provisionally, we also showed that a fixed-effects metric for Bayesian model comparison, the integrated Bayesian Information Criterion (i.e., $\mathrm{BIC}_{\mathrm{int}}$; [7]), is able to extract the true underlying model from six data sets, which were each generated by a different computational model. We note that the method is applicable to any model that generates a probability distribution over some observed data, and can accommodate models with any (e.g., non-linear) structure and any (e.g., dichotomous, continuous, mixed multivariate) nature of the observed variables. We note in this context that the algorithm involves repeated estimation of posterior model parameters for each individual, creating a need for fast fitting procedures. Here, availability of gradients, i.e., the partial derivatives of the log likelihood with respect to the model parameters, can be of great help. Taken together, the present simulations demonstrate that the EM algorithm provides a valid and efficient tool for studying computational models of observed group-data if for all individuals the observed data was generated by the same underlying computational mechanism. [8]

Previous work by Stephan et al. [3], however, demonstrated that not only parameters, but also the mechanisms generating observed data in group studies differ between individuals, and that accounting for such differences by treating computational models as a random effect via a hierarchical Bayesian model yields a more powerful method for model comparison compared to classical and previous Bayesian approaches. Their random-effects approach to the model identity has since been further developed [26] and has become a standard in fields such as dynamic causal models of fMRI and EEG data [27] and cognitive modeling of reward learning and decision-making [9, 28].

In the present work, we extend the approach by Stephan et al. [3] and introduce a novel hierarchical Bayesian method for model estimation and model selection in group studies that has several advantages compared to previous methods. This hierarchical Bayesian method estimates random effects for models and for parameters simultaneously, and thus considers individual differences and commonalities in both the mechanisms and their parameters generating observed data. We demonstrated using simulations of a simple reinforcement learning model and the dual-control model for the two-step task [9] that our novel method provides more precise inference in Bayesian model comparison compared to previous approaches, and that it is particularly beneficial in the presence of moderate degrees of posterior uncertainty.

Specifically, we compare our novel hierarchical Bayesian model to the previous approach by Stephan et al. [3]. In this previous method, model evidences

$$p(y_n \mid m_{nk}) = \int p(y_n \mid \theta_{nk}) p(\theta_{nk} \mid m_{nk}) d\theta_{nk}$$

are computed prior to hierarchical inference and are then fed into a hierarchical Bayesian model estimating random effects for the models [3]. Thus, the individual model evidences (or marginal probabilities) serve as sufficient statistics, where all the information about the data derived from the model is computed within the individual subject, and then transferred to a random effects analysis at the second (group) level.

The novel hierarchical Bayesian model that we introduce builds on and extends this prior work [3, 8] by combining the estimation of random effects for parameters and for models within a single hierarchical framework. Combining estimation for these two sources of variation within the same hierarchical model has the advantage that the covariance among models and parameters can be explicitly accounted for. We evaluated the full hierarchical model by comparing it to a sufficient statistics approach, where model parameters are estimated via the standard hierarchical model for parameters (EM), and computational models are inverted based on these estimates to obtain marginal model likelihoods, i.e., model evidences. These model evidences then served as sufficient statistics to inform estimation of random effects for the model identity using the algorithm developed by Stephan et al. [3].

Based on the elaborate hierarchical structure of our novel Bayesian model, we expected to find advantages in Bayesian model comparison in situations of posterior uncertainty. Monte Carlo simulations showed that both approaches - the new VB method and the sufficient statistics approach - provide successful estimation of model probabilities when uncertainty is low (see section *Posterior estimation under certainty*), but - as theoretically expected - reveal a higher correct probability mass and a more precise estimation of the number of subjects using each model for the full hierarchical Bayesian model when posterior probabilities for models and parameters are uncertain (see section *Posterior estimation under uncertainty* and all simulations involving the two-step task).

Specifically, we found across repeated simulations that the sufficient statistics approach over-estimated

[8] Despite these encouraging results, we note that it is known for generalized linear mixed-effects models of clustered binary data that the Laplace approximation can be less accurate in some specific circumstances [29]. Although we did not observe such inaccuracies in the current simulations they may occur for other types of models or other parameter ranges, and therefore caution is needed not only for linear [29], but also for non-linear models discussed here. However, the present simulations suggest that in some common situations investigated here the assumptions adopted in the hierarchical approach are reasonable, allowing application of these methods to understand the computational mechanisms generating observed data.

the number of subjects using more complex learning models, but under-estimated the number of non-learners. A potential reason for this bias could be that in the present simulations the more complex learning models were also able to produce the simpler non-learning patterns in a specific range of their parameters (e.g., by assuming very low learning rates or high choice noisiness), essentially yielding a nested model structure. Fitting a hierarchical model with fixed-effects for the models to such data, we found biases in the estimation of learning parameters such that in the complex models choice behavior was estimated to be less efficient than was true for the generating parameters (i.e., reduced learning rates and increased choice noisiness). This bias results from the inclusion of non-learning subjects in the estimation procedure. As one consequence of this result, group-level prior parameters for the complex model attributed relatively high probability to behavior that was somewhat more similar to the non-learning models, and was thus able to explain behavior from individuals which were actually generated by a random choice strategy. These biases can thus cause the imprecision of the sufficient statistics approach when estimating the number of subjects using each model. However, we showed that these biases can be corrected based on our novel hierarchical method, as this method considers biases in parameter estimation when inverting the computational models.

Compared to previous approaches to modeling for group studies [3,8], our VB method thus provides precise inference in Bayesian model selection under uncertainty, and it allows reducing biases in parameter estimation. The simulation results suggest that we should be able to understand the heterogeneity and homogeneity observed in learning and alcohol dependence by investigating contributions of both, the underlying computational mechanisms and the learning parameters.

The particular mathematical mechanism underlying the advantage of the novel hierarchical Bayesian model is that random effects for models and for parameters are combined to reciprocally inform their estimation. Specifically, in accord with standard Bayesian assumptions, estimation of individual posterior subject parameters in our novel model is based on a linear combination of the prior information (from the group-level) as well as an influence from the likelihood for that individual. Critically, however, as a novel result and contrary to standard approaches, the contribution of the likelihood for posterior parameter estimation is weighted by the posterior probability, $q(m_{nk} = 1)$, that this individual's data was generated by the computational model k in question (see Eq. 54). This property of our algorithm allows it to prevent over-fitting of models to individuals' data when the model likely did not generate the observed data, but instead relies on group-level prior information under such circumstances. This correction of parameter estimates, in turn, corrects the estimation of posterior model probabilities, avoiding partitions of parameter space in the model inversion that are - based on the group-level estimates - rather unlikely. This latter mechanism should indeed represent a main driving force for the improvement in Bayesian model selection that is visible in the simulations.

Another advantage of our novel hierarchical Bayesian model is that it allows to use hyper-priors, e.g., prior information from previous experiments or plausibility considerations, to constrain estimates at the group-level. Many previous studies employing random effects for models (e.g., [9,27,28]) have used standard Bayesian analysis at the individual subject-level, applying prior information repeatedly for each individual for estimating model parameters. One potential difficulty with such an approach is that the precise choice for a prior can have a strong impact on the estimation of individual subject parameters, even if the data support a different pattern of parameter values. The hierarchical structure in the novel method presented here, to the contrary, applies prior information to constrain estimates at the group-level, thus enhancing the power of the analysis to shift away from false previous believes, by pooling the data across all individuals to collectively counteract prior expectations.

A rather technical issue involved in our novel variational Bayesian procedure relates to our approach to model inversion to compute individual model evidences, which is an integral part of Bayesian model comparison. Our hierarchical model involves repeatedly computing model evidences for individual subjects by inverting the model to compute the marginal likelihood. This model inversion involves integrals over model parameters, which - for many interesting and non-trivial computational models - cannot be performed analytically, but demand approximations. Different approximations for the integrals involved in model inversion bring about different computational costs, and some more precise methods can be computationally quite expensive.

With respect to random effects for the model identity, Stephan et al. [3] use pre-computed model evidences as input to their variational Bayesian algorithm, delegating the need to choose an approximation method to previous analysis steps. This approach thus makes use of a sufficient statistic for model evidence, which is computed for each subject and for each model independently of the other individuals' and models' probabilities. It then takes this sufficient statistic to the second level to compute posterior model probabilities via a hierarchical estimation scheme. This approach shares similarities with repeated measures ANOVAs or repeated measures multiple regression [34], which use individual-subject based estimates for effect sizes as a sufficient statistic, and with SPM, the leading statistical software package for analyzing Neuroimaging data, which takes effect sizes β from first-level GLM analyses to the second level for computation of group-statistics.

Our novel variational Bayesian algorithm for simultaneous random effects for models and parameters does not rely on such a sufficient statistic, but rather estimates the full hierarchical Bayesian model, and recomputes model evidences as they change with changing parameter estimates through the iterating curse of fitting. It is exactly this simultaneous estimation of random effects for models and parameters, which provides the advantages of the current approach over the previous sufficient statistics approach. However, recomputing model evidences with changes in parameter estimates also provides a computational challenge if detailed approximations are used, and can strongly increase the time needed to fit models. Thus, the choice of approximations to the model evidence becomes important.

A common way to compute the model evidence is to use the Laplace approximation for distributions over model parameters, and thus to compare models based on the Bayesian Information Criterion (BIC, [7]). This approach has the advantage that the BIC is relatively easy to obtain and demands little computational resources. Its efficiency, however, comes at a cost as the independent normal approximation to the posterior may not be appropriate, and may fail to adequately consider posterior uncertainty and posterior covariances. Indeed, we found in some preliminary simulations (not shown) that using the BIC to compute model evidences at the individual-subject level failed to recover the computational models in some circumstances, e.g., in the simulations based on parameters derived from empirical data (see Section *Real experimental data*). We therefore chose to compute model evidences via sampling from the posterior parameter distributions. One property of this approach is that sampling can be very demanding in terms of computational resources, and when done repeatedly, can considerably prolong the fitting process. As an alternative, to speed up this process, we extended our approach to use importance sampling, such that previously computed samples can be updated via re-weighting in each individual step of the optimization algorithm, which heavily reduces computing demands. Occasional re-sampling from the current posterior estimates can be used to avoid inaccuracies due to this importance sampling scheme, and taken together, provides a workable solution to reduce processing demands while keeping accuracy high.

A currently open question concerning the novel hierarchical model is how it behaves as a function of the number of individuals using a particular model, and whether a minimal number of subjects are needed per model to support and justify the complex hierarchical structure. Indeed, it is unclear at present whether fitting many possible models to few individuals may result in overfitting, and in a posterior overconfidence of the variational Bayesian method. However, in the present work, we assessed the average correct posterior probability mass across individuals. Overconfidence should reduce the correct probability mass, since it would mean to allocate much probability mass to a false model. While this may occur for individual subjects with the present analyses, the average probability mass showed a clear increase compared to the sufficient statistics approach. However, future work may develop methods to balance complexity versus accuracy for the random effects procedure, to determine the optimal number of models included in the hierarchical framework.

An alternative approach to model comparison in hierarchical models has been to include a random effects parameter into the model which codes for the model identity (REF ?? - *I have only encountered people telling me this. Are there published studies employing this procedure?*). That is, this approach combines predictions from two (or more) models according to a weighting parameter to generate a single prediction for observed data, and the weighting parameter can then be estimated as a random effect for observed data from a group of individuals. It is a common procedure for the prior mean of the weighting parameter to test whether it gives more weighting to one model rather than to another model included in the analysis, and this statistical test is sometimes used for random effects inference on the model space. Note, however, that this approach fundamentally differs in its assumptions from the approach of treating models as a random effect in hierarchical Bayesian model comparison (see [3] and the present VB). Specifically, using a random effects parameter to code the model identity does not assume that either one model or another model generated data from an individual subject. Instead, it assumes that the data of each subject was generated by a weighted mixture of the models in question, thus assuming that all subjects used all models, but did so to differing degrees. Statistical tests on the weighting parameter in this context then allows drawing conclusions about which of the simultaneous mechanisms most strongly impacts current behavior.

To the contrary, our present analyses of random effects for models considers the case where alternative hypotheses exist about the mechanisms that have generated observed data, with the aim to obtain the posterior probability that a specific mechanism (or a class of mechanisms, see below) has generated the observed data, rather than another. In this sense, random effects inference on the model can be considered a case of model selection, rather than an estimate of relative model contribution. Interestingly, specific instantiations of such alternative mechanisms could involve mixtures of several candidate mechanisms, and this is indeed a case that we have considered in our analyses of the two-step task (see Section *Models as random effects*), where distinct model-based and model-free decision-systems are thought to differentially control decision-making, with a weighting parameter ω determining the degree to which each controls behavioral choice. In order to decide between alternative hypotheses about the mechanisms generating observed data, however, methods are needed that quantify the believe in a certain hypothesis about the mechanisms and parameters generating observed data being true rather than a set of other hypotheses. Based on previous work by Stephan et al. [3], the present work aims to extend and combine this approach with simultaneous consideration of random effects for parameters.

In accord with previous arguments [3], our theoretical and simulation results suggest that which method is best for a specific data set depends on the scientific question asked. Sometimes, a fixed-effects approach to the model may be more appropriate. In the cognitive neurosciences, for example, some cognitive and neuronal processes, like basic phenomena in psychophysics or basic physiological mechanisms, may be the same for all individuals, and in such situations fixed-effects approaches to the models will be appropriate. Whenever different individuals can exhibit different mechanisms, however, - which is most likely to be the case for more complex and higher-level cognitive functions like decision-making or reasoning, - the random effects methods presented in this paper are needed and provide the more appropriate method. For example, it is likely that in some mental diseases like addiction, patients and healthy control subjects show heterogeneity with regard to the cognitive-affective mechanisms for information processing, or that individuals at different ages or from different cultures show different cognitive or neuronal mechanisms, or different developmental trajectories governed by different computational or personality-related mechanisms. Likewise, many self-trained behaviors common to all individuals, like eye-movements, exhibit a large degree of qualitative individual characteristics, which may likely reflect individual differences in the strategies chosen to solve certain tasks, like e.g., reading or decision-making, and seem to reflect individual differences in the evolved mechanisms.

A potential further and interesting application of our novel hierarchical Bayesian method is to enter the same computational model more than once in the model estimation. As a consequence, two instances of the same model should differentiate and create different clusters of model parameters, essentially yielding a clustering algorithm in the space of model-parameters. This interestingly extends the kind of clustering discussed in the present paper - i.e., to cluster individuals based on the computational mechanisms that has generated their observed data - and complements it with a more traditional view on clustering based on model parameters. Although the implications and feasibility of such an approach need to be tested in future work, this reasoning suggests that our present algorithm may not only provide a novel and powerful method for Bayesian model selection and estimation for group studies, but may also represent a generalization of clustering algorithms to the case where clusters are defined on either the models or the parameters generating observed data.

In summary, we demonstrated that in a widely employed hierarchical modeling approach, estimating hierarchical generalized non-linear models for the case of reinforcement learning models of decision-making (cf. [8]) exhibits normative properties when the observed data by all individuals from a group are generated by the same underlying computational mechanism, but exhibits biases and inaccuracies when the generating mechanism differs between individuals. To overcome these shortcomings, our novel variational Bayesian model combines previous methods - estimating either parameters [4,8] or models as random effects in a hierarchical model [3] - to estimate random effects for models and parameters simultaneously. We demonstrated via simulations that our novel VB method exhibits increased accuracy in Bayesian model selection and parameter estimation under uncertainty compared to previous approaches [3,8], as theoretically expected. Our new variational Bayesian method suggests that we can and should understand the homogeneity and heterogeneity observed in group studies by considering both, the underlying mechanisms and their parameters. We expect that this new hierarchical method will prove useful for a wide range of computational modeling approaches in group studies on cognition, biology, and beyond.

Software note

The software used in the present simulations is freely available to the community as part of the open-source package TAPAS (www.translationalneuromodeling.org/tapas/) and at the Open Science Foundation (https://osf.io/...).

Acknowledgments

This research was supported by Deutsche Forschungsgemeinschaft (DFG FOR 1617 Alcohol and Learning), Grant RA1047/2-1.

References

1. Schad DJ, Jünger E, Garbusow M, Sebold M, Bernhardt N, et al. (2014) Processing speed and working memory shift the balance from model-free to model- based reinforcement learning. in press .

2. Schad DJ, Engbert R (2012) The zoom lens of attention: Simulating shuffled versus normal text reading using the swift model. Visual Cognition 20: 391.

3. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46: 1004-17.

4. Pinheiro J, Bates DM (2000) Mixed-effects models in S and S-PLUS. Springer.

5. Pitt MA, Myung IJ, Zhang S (2002) Toward a method of selecting among computational models of cognition. Psychol Rev 109: 472-91.

6. Neyman J, Pearson ES (1933) The testing of statistical hypotheses in relation to probabilities a priori. In: Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge Univ Press, volume 29, pp. 492–510.

7. Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90: 773-795.

8. Huys QJ, Cools R, Golzer M, Friedel E, Heinz A, et al. (2011) Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Computational Biology 7: e1002028.

9. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69: 1204-1215.

10. Griffiths TL, Steyvers M, Tenenbaum JB (2007) Topics in semantic representation. Psychological Review 114: 211.

11. Oberauer K, Kliegl R (2006) A formal model of capacity limits in working memory. Journal of Memory and Language 55: 601–626.

12. Friston KJ, Harrison L, Penny W (2003) Dynamic causal modeling. Neuroimage 19: 1273-302.

13. Stephan KE, Kasper L, Harrison LM, Daunizeau J, den Ouden HE, et al. (2008) Nonlinear dynamic causal models for fmri. Neuroimage 42: 649–662.

14. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via em algorithm. J Roy Stat Soc B Met 39: 1-38.

15. Wunderlich K, Smittenaar P, Dolan RJ (2012) Dopamine enhances model-based over model-free choice behavior. Neuron 75: 418-24.

16. Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ (2013) Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron 80: 914-9.

17. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. P Natl Acad Sci USA 110: 20941-6.

18. Kliegl R (2007) Toward a perceptual-span theory of distributed processing in reading: A reply to rayner, pollatsek, drieghe, slattery, and reichle (2007). Journal of Experimental Psychology: General 136: 530.

19. Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59: 390-412.

20. Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68: 255-278.

21. Schad DJ, Nuthmann A, Engbert R (2012) Your mind wanders weakly, your mind wanders deeply: Objective measures reveal mindless reading at different levels 125: 179-194.

22. Lee SW, Shimojo S, O'Doherty JP (2014) Neural computations underlying arbitration between model-based and model-free learning. Neuron 81: 687-99.

23. Economides M, Guitart-Masip M, Kurth-Nelson Z, Dolan RJ (2014) Anterior cingulate cortex instigates adaptive switches in choice by integrating immediate and delayed components of value in ventromedial prefrontal cortex. The Journal of Neuroscience 34: 3340–3349.

24. Penny WD (2014) Population level models of dynamical systems. In: Statistical challenges in Neuroscience.

25. MacKay DJ (2003) Information theory, inference and learning algorithms. Cambridge University Press.

26. Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies—revisited. Neuroimage 84: 971–985.

27. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, et al. (2013) Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron 80: 519-30.

28. Dezfouli A, Balleine BW (2013) Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLoS Comput Biol 9: e1003364.

29. Fong Y, Rue H, Wakefield J (2010) Bayesian inference for generalized linear mixed models. Biostatistics 11: 397-412.

30. Crowder MJ, Hand DJ (1990) Analysis of repeated measures, volume 41. CRC Press.

31. Coleman TF, Li Y (1996) An interior trust region approach for nonlinear minimization subject to bounds. SIAM Journal on optimization 6: 418–445.

32. Cavanagh JF, Eisenberg I, Guitart-Masip M, Huys Q, Frank MJ (2013) Frontal theta overrides pavlovian learning biases. The Journal of Neuroscience 33: 8541–8548.

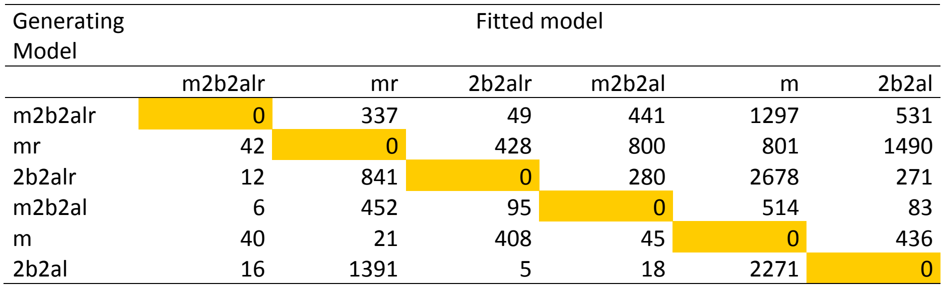
33. Rescorla RA, Wagner AR, et al. (1972) A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory 2: 64–99.

34. Lorch RF, Myers JL (1990) Regression analyses of repeated measures data in cognitive research. Journal of Experimental Psychology: Learning, Memory, and Cognition 16: 149.

Figure Legends

Tables

Table 1. Results from Bayesian model comparison showing relative $\mathrm{BIC}_{\mathrm{int}}$ values for six data sets simulated from six different computational models of decision making in a setting treating models as fixed-effects.



Generating Model	Fitted model					
	m2b2alr	mr	2b2alr	m2b2al	m	2b2al
m2b2alr	0	337	49	441	1297	531
mr	42	0	428	800	801	1490
2b2alr	12	841	0	280	2678	271
m2b2al	6	452	95	0	514	83
m	40	21	408	45	0	436
2b2al	16	1391	5	18	2271	0

Note. m = model-based choice, 2b2al = model-free choice, r = choice repetition. Relative $\mathrm{BIC}_{\mathrm{int}}$ values are computed as the $\mathrm{BIC}_{\mathrm{int}}$ minus the smallest $\mathrm{BIC}_{\mathrm{int}}$ from all tested models per simulated data set.

Supporting Information Legends

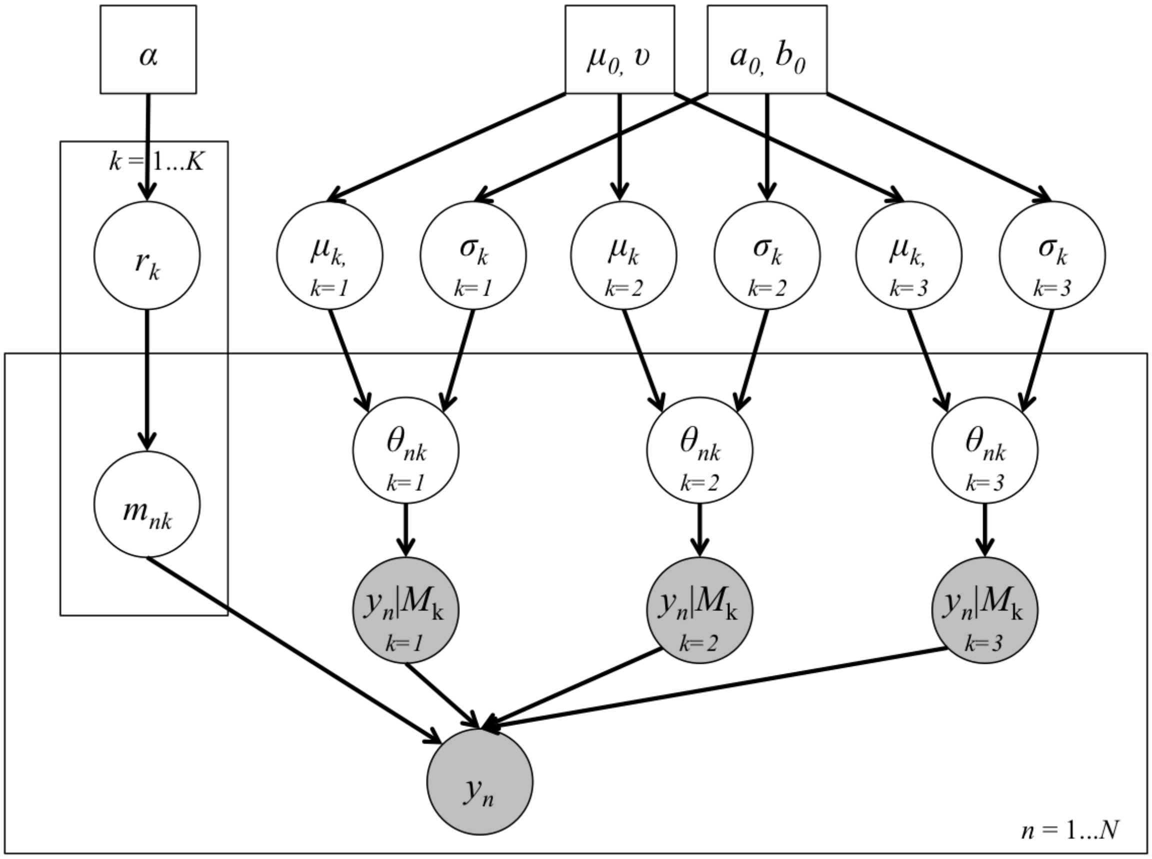


Figure 1. Bayesian dependency graphs for the random effects generative model for multi-subject data. Rectangles denote deterministic parameters and shaded circles represent observed values. $\alpha =$ parameters of the Dirichlet distribution (number of model "occurrences"); $r =$ parameters of the multinomial distribution (probabilities of the models); $m =$ model labels; $\theta =$ individual subject parameters; $\mu =$ prior group mean; $\sigma^2 =$ prior group variance; $\mu_0, \nu =$ hyper-priors for the μ parameter; $a_0, b_0 =$ hyper-priors for the σ^2 parameter; $y =$ observed data; $y \mid m =$ probability of the data given model k; $k =$ model index; $K =$ number of models; $n =$ subject index; $N =$ number of subjects.

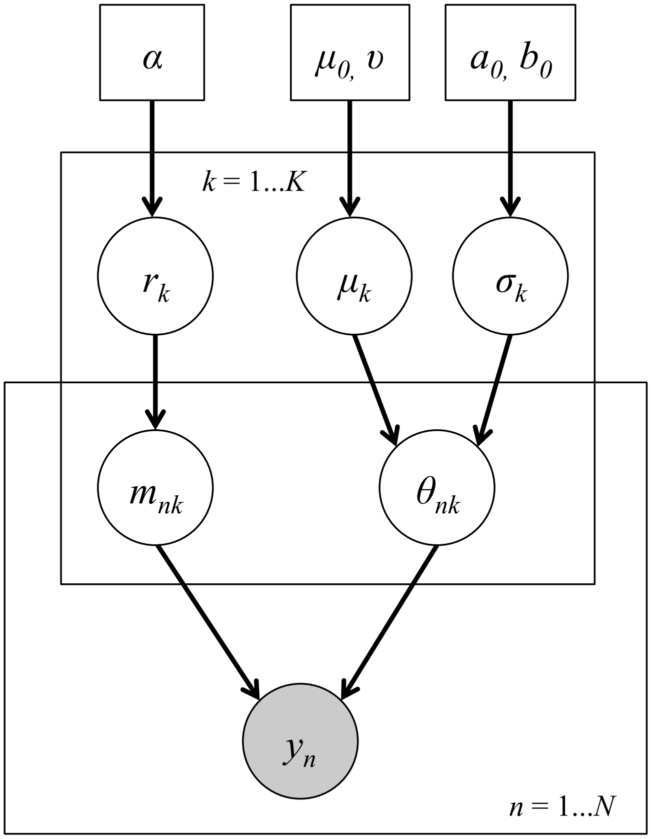


Figure 2. Reduced Bayesian dependency graphs for our random effects generative model for multi-subject data. Rectangles denote deterministic parameters and shaded circles represent observed values. α = parameters of the Dirichlet distribution (number of model "occurrences"); α_0 = hyper-priors for the Dirichlet distribution; r = parameters of the multinomial distribution (probabilities of the models); m = model labels; θ = individual subject parameters; μ = prior group mean; σ^2 = prior group variance; μ_0, ν = hyper-priors for the μ parameter; a_0, b_0 = hyper-priors for the σ^2 parameter; y = observed data; k = model index; K = number of models; n = subject index; N = number of subjects.

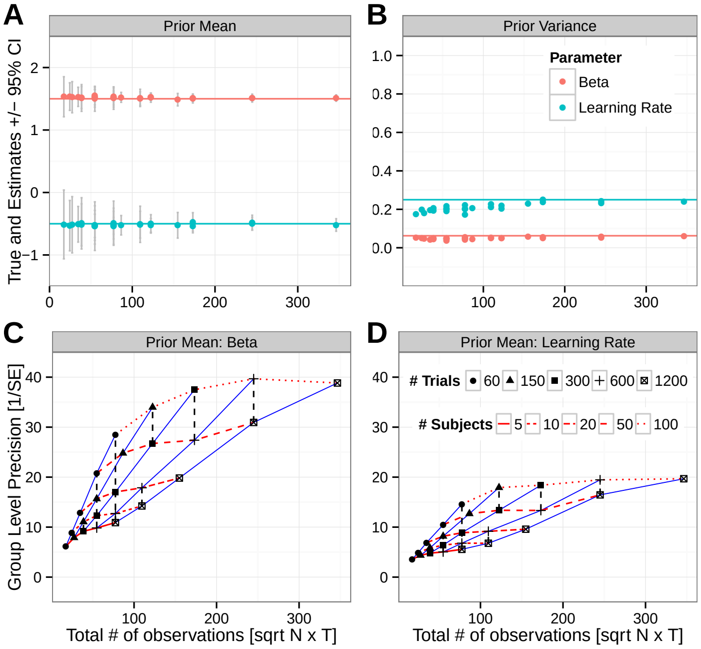


Figure 3. Random effects for parameters, fixed-effects for models. Results from Monte Carlo simulations based on a known decision process testing the random effects method for estimating group-level parameters (Huys et al. 2012) for learning rate (green) and choice noisiness (red, $\beta = beta$) of a simple reinforcement learning model as a function of the total number of observations (square root of the number of subjects times the number of trials). Estimated parameters (points) with 95 % confidence intervals (CI) are means over repeated simulations. Panel A: The random effects model specifies average model parameters across the whole group of subjects, reflecting the prior means. Panel B: Random effects model specifies random variance of model parameters across subjects, i.e., random effects variance or prior variances. Throughout, estimates closely correspond to the true generating parameter values (solid lines), suggesting it is possible to use the Expectation Maximization algorithm (Dempster et al. 1977) to obtain unbiased estimates of model parameters at the group level when all subjects use the same model. Panels C+D: Precision (i.e., 1/standard error) of the prior means as a function of the number of trials (indicated by different shapes) and the number of subjects (indicated by different line types) for the β parameter (Panel C) and for the learning rate parameter (Panels D).

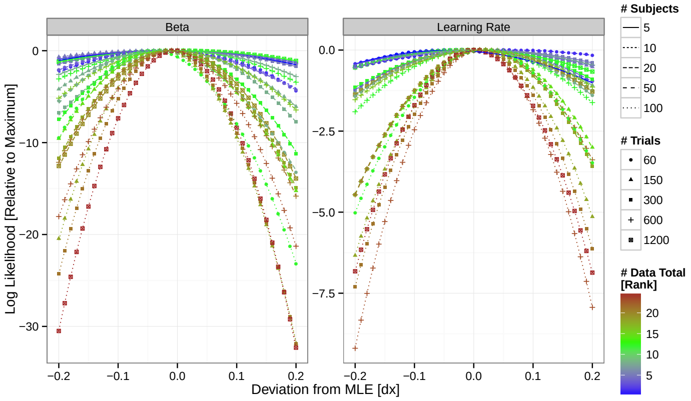


Figure 4. Relative log-likelihood for the group-level means. Relative log-likelihood is shown for the β parameter (left panel, *beta*) and for the learning rate (right panel) as a function of the deviation (dx) from the maximum likelihood estimate (MLE) and of the number of subjects and the number of trials. The log-likelihood is estimated via importance sampling for different values of the group-level mean parameters (points), and fitted via quadratic functions (lines) for visualization. Sampled log-likelihood values closely correspond to the quadratic function, which (*i*) indicates approximately normally distributed likelihoods, validating the use of the Laplace approximation, and (*ii*) indicates reliable sampling from the likelihood, validating the importance sampling approach.

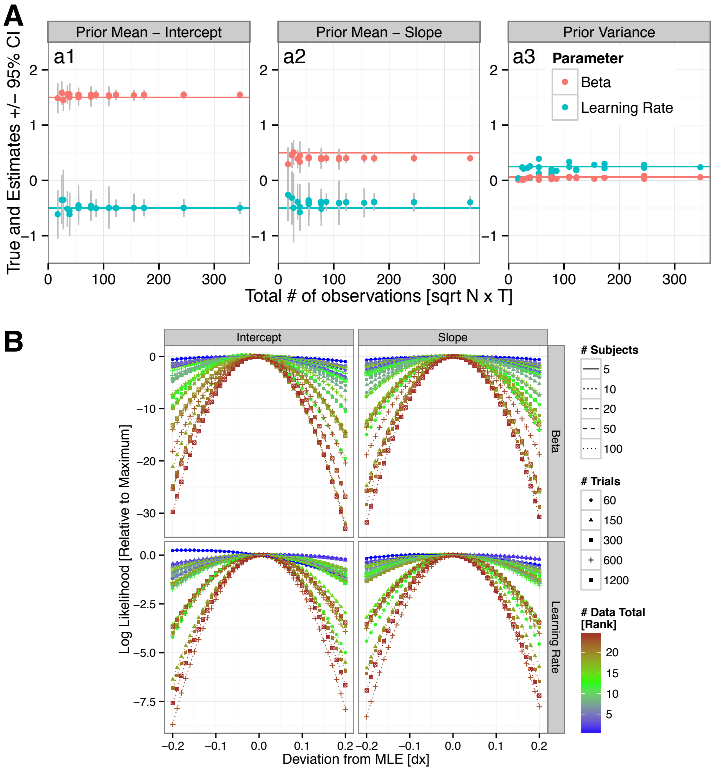


Figure 5. Evaluating GLM on individual subject parameters. Results from evaluating the random effects method for estimating between-subject GLMs on the model parameters learning rate (α) and the inverse choice noisiness (β, beta) of a simple reinforcement learning model as a function of the total number of simulated observations. Panel A: True generating parameters (lines) and estimated parameters (points) for the group-level means (left, Prior Mean - Intercept), for the regression coefficients from the GLMs (middle, Prior Mean - Slope), and for the group-level variance (right, Prior Variance). Estimated parameters (points) with 95 % confidence intervals (CI) are means over repeated simulations. Panel B: Relative log-likelihoods for the group-level means (Intercept, right panels) and the regression coefficients (Slope, left panels) for the β parameter (upper panels, *Beta*) and for the learning rate (lower panels) as a function of the deviation (dx) from the maximum likelihood estimate (MLE). The log-likelihood is estimated via importance sampling (points), and fitted via quadratic functions (lines). Sampled log-likelihood values closely correspond to the quadratic function, validating the present procedure for the GLM at the level of model-parameters.

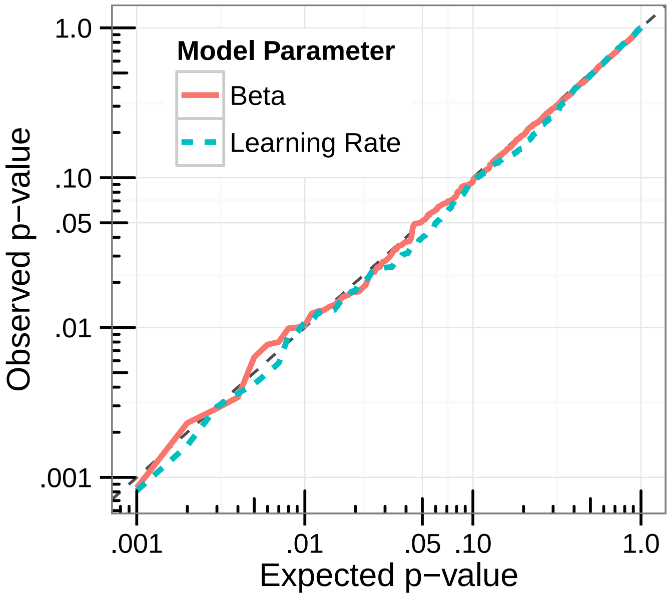


Figure 6. P-P plot for subject GLM. P-P plot showing the proportion of type I errors for GLM coefficients for the beta and learning rate parameters, estimated on 1,000 independent tests performed under a null hypothesis. Simulations are based on $N_{subj} = 30$ subjects and $N_{trials} = 60$ trials. Both axes are presented in logarithmic scale.

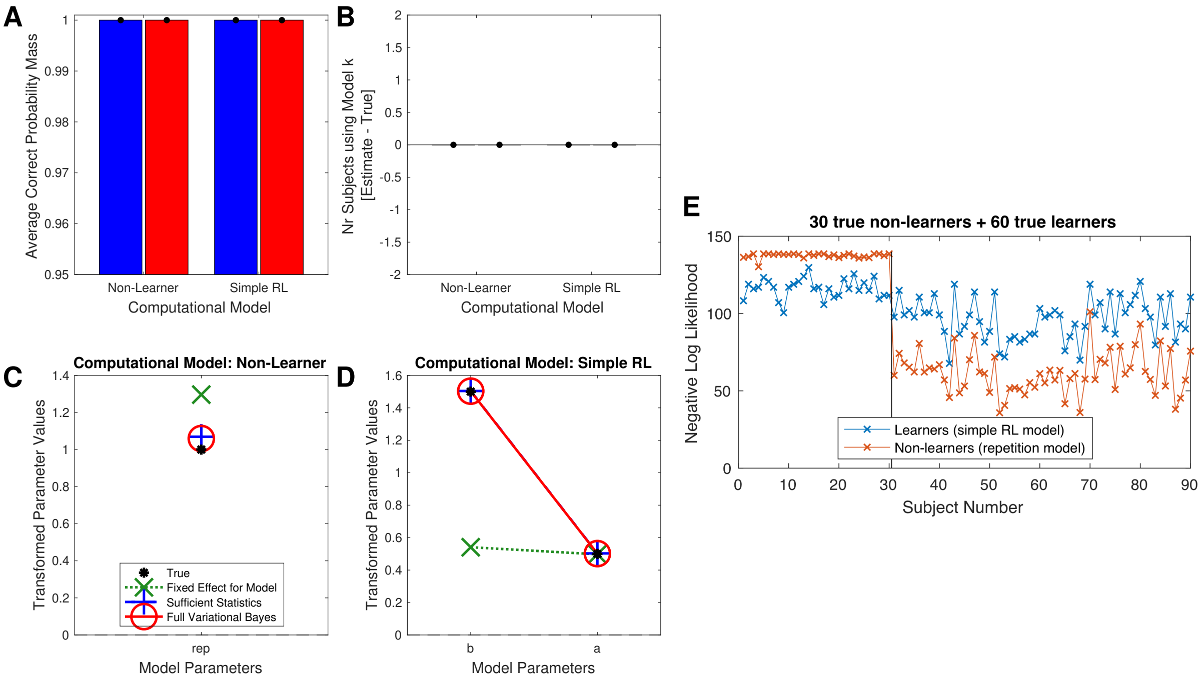


Figure 7. Comparing methods for Bayesian model estimation and comparison under conditions of certainty. Results from Monte Carlo simulations comparing methods for Bayesian model comparison (panels A + B) and parameter estimation (panels C + D) under conditions of high certainty with many trials per individual ($N_t = 200$) which well constrain the simple models used. We simulated two groups of subjects performing a simple reward-learning task using different generating mechanisms: (i) a learning model allowing for model-free learning of expected values of available actions ($N_s = 60$) and (ii) a model assuming no learning, but a mere tendency to repeat previous choices (non-learners, $N_s = 30$). **Panels A+B:** Results show that the full variational Bayesian (red bars) and also the *Sufficient Statistics* approach (blue bars; combining Huys et al. 2012, and Stephan et al. 2009) extract the true generating model with high certainty. **Panels C+D:** True (black points) and estimated model parameters at the group level (fixed-effects) are displayed for the non-learning model (panel C; repetition parameter, rep) and for the learning model (panel D, inverse noise parameter b and learning rate a). Results show that a fixed-effects approach to the model identity (model as fixed effect; green marks; Huys et al. 2012) can exhibit biases in parameter estimation when the true model differs between subjects. Precise estimates are obtained using Bayesian subject averaging (*Sufficient Statistics* approach; blue marks) or via simultaneous random effects for models and parameters in the full hierarchical Bayesian model (full Variational Bayes; red marks). **Panels E:** Results show negative log likelihood for the non-learner model (orange) and for the simple RL model (blue) for each individual subject. The first 30 subjects' data is generated from a non-learning model, and data from subjects 31 to 90 are generated by a simple RL model. The results show that simple maximum likelihood estimation per subject is sufficient to obtain the correct model in each individual subject.

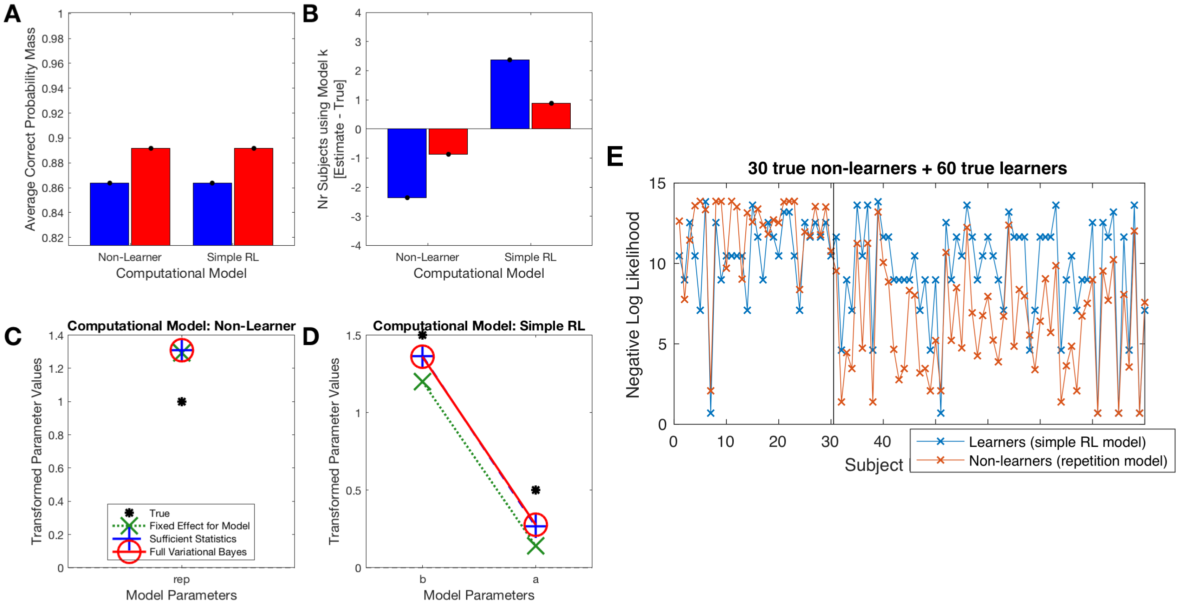


Figure 8. Comparing methods for Bayesian model estimation and comparison under uncertainty. Results from Monte Carlo simulations comparing methods for Bayesian model comparison (panels A + B) and parameter estimation (panels C + D) under conditions of high uncertainty with few trials per individual ($N_t = 20$), which poorly constrains models. All other details of the simulations are identical with those reported in Figure 7. **Panels A+B:** Results show that the *Sufficient Statistics* approach (blue bars; combining Huys et al. 2012, and Stephan et al. 2009) has difficulties extracting the correct model from the observed data, and that the full variational Bayesian approach (red bars) improves this accuracy, both at the level of individual posterior model probabilities (panel A) as well as for group-level estimates for the number of individuals using each model in question (panel B). **Panels C+D:** True (black points) and estimated model parameters at the group level (fixed-effects) are displayed for the non-learning model (panel C; repetition parameter, *rep*) and for the learning model (panel D; inverse noise parameter b and learning rate a). Results show that a fixed-effects approach to the model identity (green marks; Huys et al. 2012) can exhibit biases in parameter estimation when the true model differs between individuals. Given the sparse amount of data for each individual, attempts to correct these biases via random effects for models and parameters fail for the non-learning model (panel C), but somewhat improve estimation for the learning model, with the best estimates obtained by the variational Bayesian algorithm (red marks; panel D). **Panels E:** Results show negative log likelihood for the non-learner model (orange) and for the simple RL model (blue) for each individual subject. The first 30 subjects' data is generated from a non-learning model, and data from subjects 31 to 90 are generated by a simple RL model. The results show that simple maximum likelihood estimation per subject with 20 trials is much more noisy than with 200 trials, and that the wrong model wins in some of the subjects.

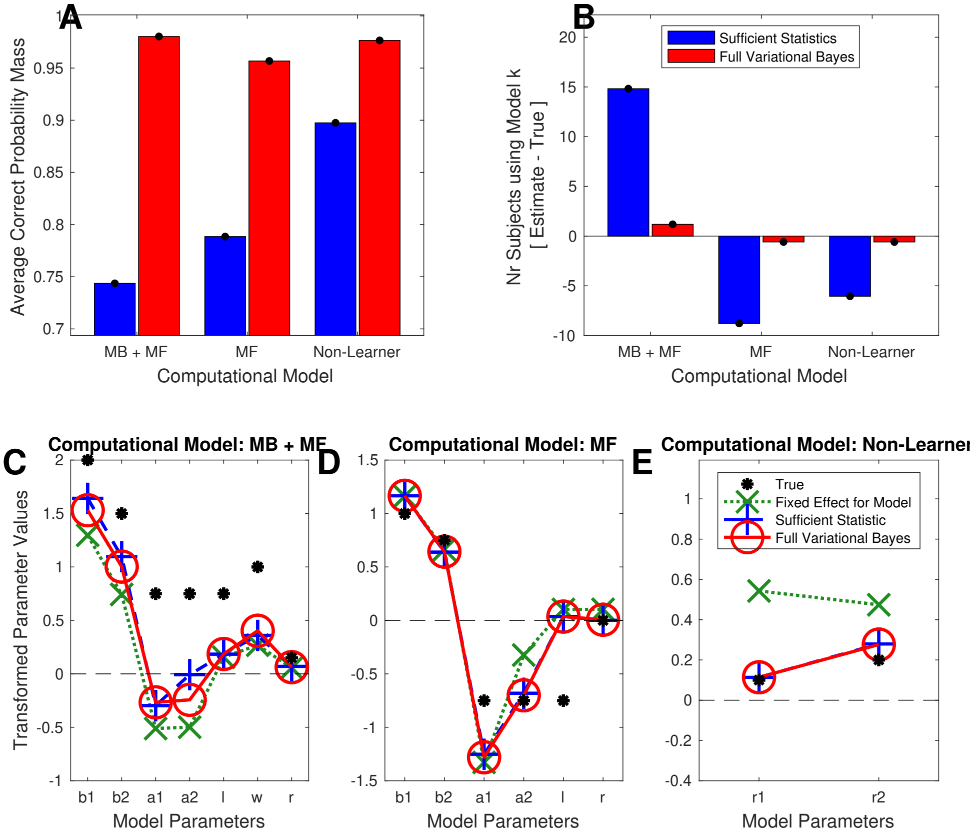


Figure 9. Comparing methods for Bayesian model estimation and comparison for the two-step task. Results from Monte Carlo simulations comparing methods for Bayesian model comparison (panels A + B) and learning parameter estimation (panels C-E). We simulated three groups of subjects (each $N_s = 30$) performing the two-step task using different computational mechanisms: (*i*) the dual-control model allowing for both model-based (MB) and model-free (MF) choice in combination with choice stickiness (”MB + MF”); (*ii*) a purely model-free learning model which also includes choice stickiness (”MF”); and (*iii*) a model assuming no learning (i.e., a mere tendency to repeat previous choices; ”Non-Learner”). **Panels A+B:** Results show that the full variational Bayesian method (red bars) extracts the true generating model with high certainty, while the *Sufficient Statistics* approach (blue bars; combining Huys et al. 2012, and Stephan et al. 2009) is less certain about the true model identities (panel A) and is less precise in the estimated number of subjects using each model (panel B). **Panels C-E:** True (black points) and estimated learning parameters at the group level are displayed for the full dual-control model (model-based and model-free, MB + MF; panel C), the model-free learner (MF, panel D), and the non-learning model (panel E). Results show that a fixed-effects approach to the model identity (fixed effect for models; green marks; Huys et al. 2012) can exhibit biases in parameter estimation when the true model differs between subjects. Biases are reduced using random effects for models and parameters (*Sufficient Statistics*, blue marks), with best estimates obtained by the Variational Bayes (red marks).

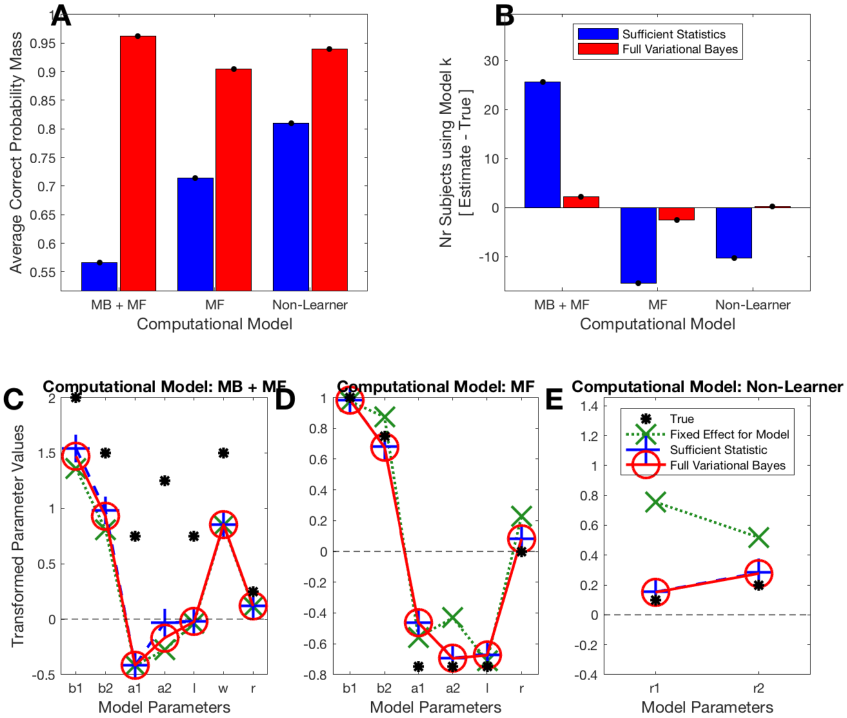


Figure 10. Increasing uncertainty in the two-step task. Results from Monte Carlo simulations comparing methods for Bayesian model comparison (panels A + B) and learning parameter estimation (panels C-E). We used the same simulations as reported in Figure 9, but used a smaller number of $N_t = 101$ trials per subject. **Panels A+B:** Results show that the full Variational Bayes (red bars) still extracts the true generating model with high certainty (panel A) and provides good estimates for the number of subjects using each model (panel B), while the "Sufficient Statistics" approach (blue bars; combining Huys et al. 2012, and Stephan et al. 2009) is even less certain about the true model identities (panel A) and overestimates the number of subjects using the full dual control model (panel B, MB + MF). **Panels C-E:** True (black points) and estimated learning parameters at the group level are displayed for the full dual-control model (model-based and model-free, MB + MF; panel C), the model-free learner (MF, panel D), and the non-learning model (panel E). Results show that a fixed-effects approach to the model identity (fixed effect for models; green marks; Huys et al. 2012) can exhibit biases in parameter estimation when the true model differs between subjects. Biases are strong in this condition with high uncertainty, and can only sometimes be overcome using random effects for models and parameters (*Sufficient Statistics*, blue marks), with the best estimates obtained by the Variational Bayes (red marks).

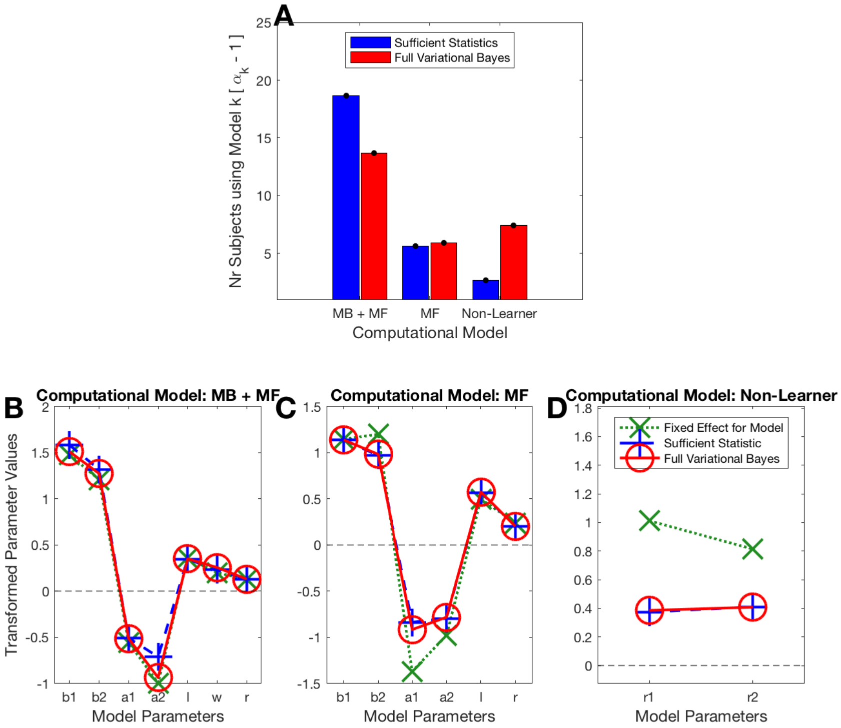


Figure 11. Real experimental data on the two-step task. Results from applying the methods for Bayesian model comparison (panel A) and learning parameter estimation (panels B-D) to real experimental data from the two-step task (Schad et al., 2014). We assumed the same computational models as reported in Figure 9. **Panel A:** Results show that the full dual-control model (MB + MF) is most likely for the observed data, which is consistent with previous findings (Schad et al., 2014; Daw et al., 2011), with little differences between the sufficient statistics approach (blue bars) and the full VB algorithm (red bars). **Panels B-D:** Estimated learning parameters at the group level are displayed for the full dual-control model (model-based and model-free, MB + MF; panel B), the model-free learner (MF, panel C), and the non-learning model (panel D). Results show that a fixed-effects approach to the model identity (fixed effect for models; green marks; Huys et al. 2012) reveals different estimates for some of the model parameters. Specifically, for the dual-control model the learning parameters (a1 and a2) are somewhat down-biased, and for the non-learner model the repetition parameters are biased upwards for the fixed-effects approach compared to the *Sufficient Statistics* (blue marks) and the full hierarchical Bayesian (red marks) approaches.

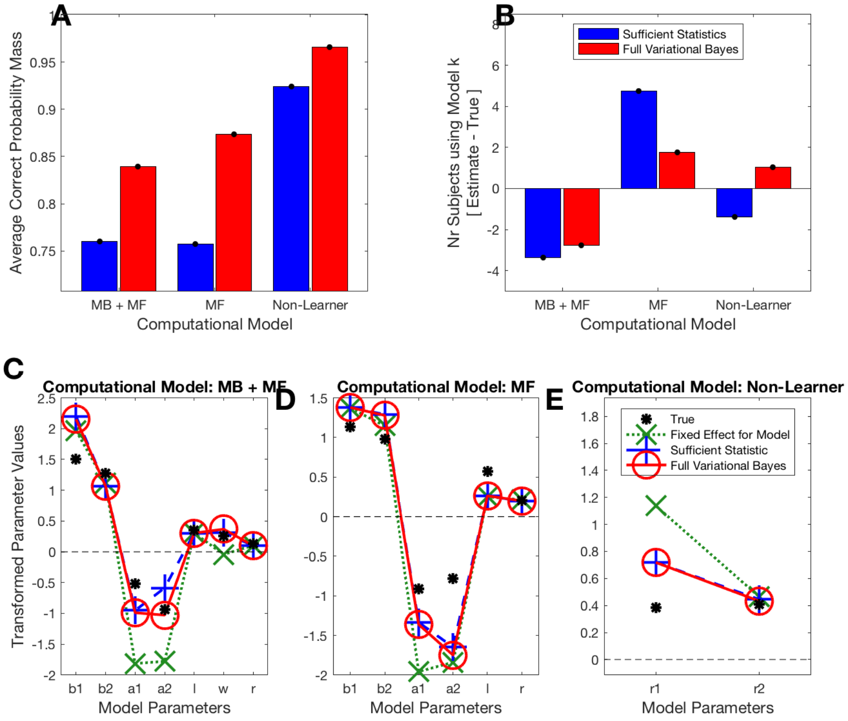


Figure 12. Simulating based on results from real experimental data. Results from Monte Carlo simulations comparing methods for Bayesian model comparison (panels A + B) and learning parameter estimation (panels C-D). We used the same simulation methods as reported in Figure 9, but derived the number of subjects using each model and the model parameters such as estimated from the real empirical data set with results reported in Figure 11. **Panels A+B:** Results again show that the full variational Bayesian approach (red bars) extracts the true generating model with high certainty, while the *Sufficient Statistics* approach (blue bars; combining Huys et al. 2012, and Stephan et al. 2009) is uncertain about the true model identities. **Panels C-E:** True (black points) and estimated learning parameters at the group level are displayed for the full dual-control model (model-based and model-free, MB + MF; panel C), the model-free learner (MF, panel D), and the non-learning model (panel E). Results show that a fixed-effects approach to the model identity (fixed effect for models; green marks; Huys et al. 2012) exhibits biases in parameter estimation when the true model differs between subjects. Biases are reduced using simultaneous random effects for models and parameters with similar results obtained for the sufficient statistics (blue marks) and full Variational Bayes (red marks) approaches.

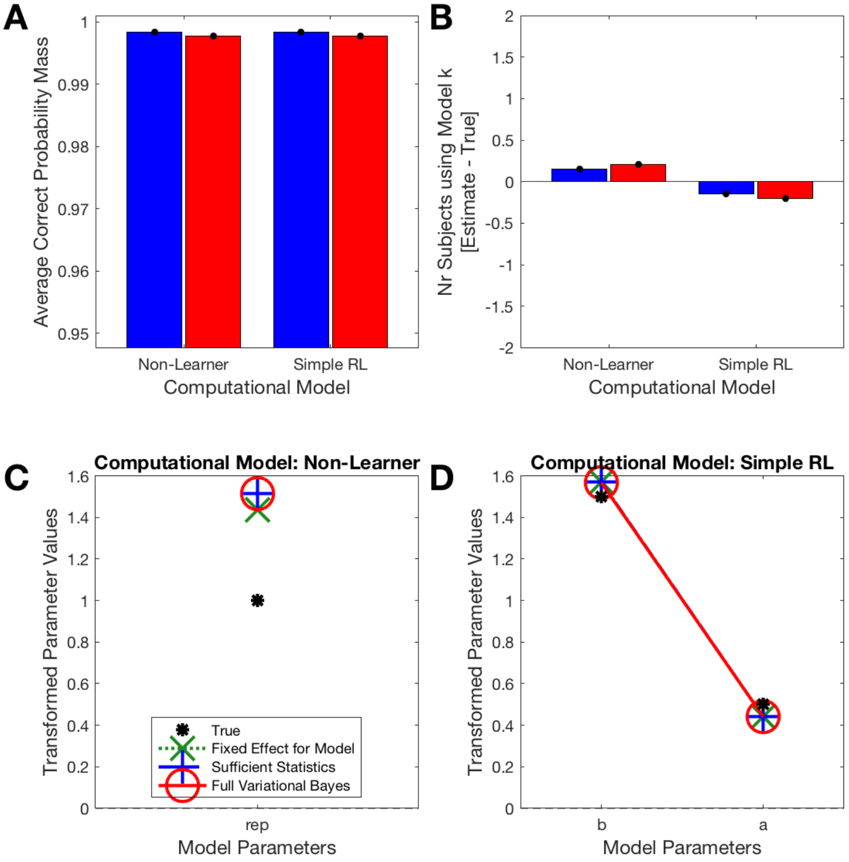


Figure 13. Simulating all subjects' data from the same simple RL model. Results from Monte Carlo simulations comparing methods for Bayesian model comparison (panels A + B) and learning parameter estimation (panels C-D). We used the same simple RL model for simulating data from all subjects. Next, we fitted these data assuming random effects in the models, where each subjects' data could be fitted by the simple RL model or by a non-learner (simple repetition) model.**Panels A+B:** The results show that despite the small number of 20 trials, both methods are successful in identifying the correct model from the data. Both, the full variational Bayesian approach (red bars) and the *Sufficient Statistics* approach (blue bars; combining Huys et al. 2012, and Stephan et al. 2009) extract the true generating model with high certainty. **Panels C-E:** True (black points) and estimated learning parameters at the group level are displayed for the simple RL learner model (panel C) and for the non-learning model (panel D). Results show show similar parameter estimation for all methods.