Journal of Mathematical Psychology 71 (2016) 1–6

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Empirical priors for reinforcement learning models

Samuel J. Gershman [*]

Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA

HIGHLIGHTS

- Reinforcement learning models suffer from the difficulty of parameter estimation.
- Empirical priors improve predictive accuracy, reliability, identifiability, and detection of individual differences.
- These priors are fairly robust across model variants.

ARTICLE INFO

Article history:
Received 6 October 2015
Received in revised form
25 January 2016

Keywords:
Bayesian statistics
Q-learning
Parameter estimation
Model comparison

ABSTRACT

Computational models of reinforcement learning have played an important role in understanding learning and decision making behavior, as well as the neural mechanisms underlying these behaviors. However, fitting the parameters of these models can be challenging: the parameters are not identifiable, estimates are unreliable, and the fitted models may not have good predictive validity. Prior distributions on the parameters can help regularize estimates and to some extent deal with these challenges, but picking a good prior is itself challenging. This paper presents empirical priors for reinforcement learning models, showing that priors estimated from a relatively large dataset are more identifiable, more reliable, and have better predictive validity compared to model-fitting with uniform priors.

1. Introduction

Reinforcement learning (RL) models formalize the process through which stimulus-reward predictions are acquired and used to guide choice behavior (Sutton & Barto, 1998). These models have become important tools for developing a mechanistic understanding of RL in the brain, as well as its breakdown in psychiatric and neurological disorders (Maia & Frank, 2011). The successful application of RL models hinges on accurately estimating parameters, perform model comparison, and predict new data. Because these models are non-linear functions of their parameters, it is necessary to rely on optimization or Monte Carlo sampling (Daw, 2011). These methods are prone to errors which are computationally expensive to correct (e.g., one could run the optimizer with more initializations, or generate more Monte Carlo samples). There are also fundamental problems that more computation cannot address, such as estimation error due to small sample sizes and poor parameter identifiability.

When sample size is small and the data are noisy relative to the complexity of the model being fit, parameters can be "overfit"—i.e., the estimated parameters do not generalize to new datasets. Overfitting can be controlled by constraining the complexity of the model, or by placing prior probabilities on the parameters that control the "effective" complexity. Intuitively, if there are two parameters, and one parameter is constrained by the prior to take on a fixed value, then the model effectively has one parameter.

Priors can also aid identifiability. A model is identifiable if different parameter settings cannot produce equivalent likelihoods (Casella & Berger, 2002). Identifiability is not especially important if one's only goal is prediction or model comparison. However, if one wishes to interpret the parameter estimates (e.g., make an inference that a particular parameter lies within some range of values) or correlate them with other measurements (e.g., individual differences analyses), then identifiability is crucial. RL models suffer from non-identifiability; for example, equivalent likelihoods can be achieved by different combinations of learning rate and inverse temperature. One symptom of this non-identifiability is correlation between parameter estimates across participants—a commonly observed but poorly appreciated phenomenon.[1]

[*] Correspondence to: 52 Oxford St., Room 295.05, Cambridge, MA 02138, USA.
E-mail address: gershman@fas.harvard.edu.

[1] Fully Bayesian approaches, which estimate the posterior distribution (e.g., using Monte Carlo simulation) rather than a point estimate, can reveal non-identifiability by inspecting correlations between parameters in the joint posterior. The Laplace approximation, which we use below, produces a local Gaussian approximation of this joint distribution around the posterior mode.

http://dx.doi.org/10.1016/j.jmp.2016.01.006
0022-2496/

Different participants may have different fitted parameter values, but all these values may lie along an iso-likelihood contour in the parameter space. When changing one parameter can compensate for changes in another parameter so as to remain on the contour, then fitted parameter values will be correlated.[2]

The approach advocated in this paper is to use "empirical priors" estimated from a separate dataset. The basic idea is to use the distribution of parameter estimates to construct a parameterized prior that is transferable to other datasets. Below, we describe the steps involved, along with a quantitative evaluation. We ask four questions about empirical priors:

1. Do they improve predictive accuracy?
2. Do they improve reliability of parameter estimates?
3. Do they improve parameter identifiability?
4. Do they improve the measurement of individual differences?

To foreshadow our results, the answer to all four question is *yes*.

2. Methods

2.1. Participants

Dataset 1 (D1 hereafter) collects together 166 participants across 4 experiments reported in Gershman (2015). In that paper, model comparison suggested that participants behaved essentially the same across experiments, which licenses collapsing the experiments together. Dataset 2 (D2 hereafter) consists of new data from 40 participants doing the same task as the participants in D1 but with different reward probabilities (see below). In addition, we collected predictions of reward probability for the chosen option on every trial, using a continuous rating scale. Participants did both tasks on the web, via Amazon's Mechanical Turk service (they were thus drawn from the same population; participants were not excluded from doing both experiments). The experiment was approved by the Harvard Institutional Review Board and participants were paid for their participation.

2.2. Procedure

On each trial, participants were shown two colored buttons and told to choose the button that they believed would deliver the most reward. After clicking a button, participants received a binary (0, 1) reward with some probability. The probability for each button was fixed throughout a block of 25 trials. In D1, there were two types of blocks, presented in a randomized order: low reward rate blocks and high reward rate blocks. On low reward rate blocks, both options delivered reward with probabilities less than 0.5. On high reward rate blocks, both options delivered reward with probabilities greater than 0.5. These probabilities (which were never shown to participants) differed across experiments (see Gershman, 2015, for more details).

D2 followed the same procedure, but with different reward probabilities. Specifically, on each block one of the options always delivered reward with a probability less than 0.5, and the other option always delivered reward with a probability greater than 0.5. The 4 reward probability pairs were (0.4, 0.6), (0.3, 0.7), (0.2, 0.8) and (0.1, 0.9). Each reward probability pair was experienced for 25 trials (thus a total of 100 trials per subject). Condition order was randomized across participants. For the purposes of this paper, the differences between these conditions are not particularly important; performance depended on the difference in reward probability between the two options, but the model fits and parameter estimates did not differ appreciably across experiments or conditions.

2.3. Models

We fit 4 different models to participants' choice data:

- **M1**: **Single learning rate**. After choosing option $c_t \in \{1, 2\}$ on trial t and observing reward $r_t \in \{0, 1\}$, the value (reward estimate) of the option is updated according to:

$$V_{t+1}(c_t) = V_t(c_t) + \eta\delta_t, \tag{1}$$

where $\eta \in [0, 1]$ is the learning rate and $\delta_t = r_t - V_t(c_t)$ is the prediction error. The values were initialized to 0. This is a standard Q-learning model (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Sutton & Barto, 1998) with a single fixed learning rate. For this and subsequent models, all values are initialized to zero. A logistic sigmoid (softmax) transformation is used to convert values to choice probabilities:

$$P(c_t = 1) = \frac{1}{1 + e^{-\beta[V_t(1) - V_t(2)]}}, \tag{2}$$

where β is an "inverse temperature" parameter that governs the exploration–exploitation trade-off.

- **M2**: **Dual learning rates**. This model is identical to M1, except that it uses two different learning rates, η^+ for positive prediction errors ($\delta_t > 0$) and η^- for negative prediction errors ($\delta_t < 0$). This model has been explored by several authors (Daw, Kakade, & Dayan, 2002; Frank, Doll, Oas-Terpstra, & Moreno, 2009; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012).
- **M3**: **Single learning rate** + **stickiness**. This model is identical to M1, with the addition of a "stickiness" parameter ω that biases repetition of choices independent of reward history:

$$P(c_t = 1) = \frac{1}{1 + e^{-\beta[V'_t(1) - V'_t(2)]}}, \tag{3}$$

$$V'_t(c) = \begin{cases} V_t(c) + \omega & \text{if } c_{t-1} = c \\ V_t(c) & \text{if } c_{t-1} \neq c. \end{cases} \tag{4}$$

In words, the stickiness parameter adds a bonus onto the option value of the most recently chosen option. A number of studies have used this or similar parameterizations (e.g., Christakou et al., 2013; Gershman, Pesaran, & Daw, 2009).

- **M4**: **Dual learning rates** + **stickiness**. This model is a combination of models M2 and M3, with separate learning rates for positive and negative prediction errors, as well as a stickiness parameter.

2.4. Parameter estimation and model comparison

Parameters for model m and subject s (denoted θ_{ms}) were estimated by optimizing the maximum *a posteriori* (MAP) objective function—i.e., finding the posterior mode:

$$\hat{\theta}_{ms} = \underset{\theta_{ms}}{\operatorname{argmax}}\, p(D_s|\theta_{ms}, m)p(\theta_{ms}|m, \phi_m), \tag{5}$$

where $p(D_s|\theta_{ms}, m)$ is the likelihood of data D_s for subject s conditional on parameters θ_{ms} and model m, and $p(\theta_{ms}|m, \phi_m)$ is the prior probability of θ_{ms} conditional on model m and hyperparameters ϕ_m. We assume each parameter is bounded and use constrained optimization to find the MAP estimates.[3]

To compare models, we assumed that each model occurs with some frequency in the population (i.e., the assignment of models

[2] More complex identifiability issues, such as contours that do not change monotonically as a function of two parameters, will not be revealed by correlations. Furthermore, correlations can also reflect meaningful individual differences. In general, parameter correlations must be interpreted with caution.

[3] Software for performing optimization and other analyses reported in this paper is available at https://github.com/sjgershm/mfit. Reinforcement learning models and data are available at https://github.com/sjgershm/RL-models.

Table 1
Empirical prior distributions and hyperparameters. Gamma distribution is parameterized in terms of shape and scale. For gamma and beta distributions, a mode/standard deviation parameterization is also given.

	Inverse temperature	Learning rate	Stickiness
Bounds	[0, 50]	[0, 1]	[−5, 5]
M1	$\beta \sim$ Gamma(4.83, 0.73) Mode: 2.8 Standard deviation: 1.6	$\eta \sim$ Beta(0.007, 0.018) Mode: 0.5 Standard deviation: 0.4	
M2	$\beta \sim$ Gamma(5.09, 0.83) Mode: 3.39 Standard deviation: 1.87	$\eta^+ \sim$ Beta(0.009, 0.026) Mode: 0.5 Standard deviation: 0.41 $\eta^- \sim$ Beta(0.015, 0.023) Mode: 0.5 Standard deviation: 0.48	
M3	$\beta \sim$ Gamma(2.52, 1.34) Mode: 2.04 Standard deviation: 2.13	$\eta \sim$ Beta(0.195, 0.479) Mode: 0.61 Standard deviation: 0.35	$\omega \sim \mathcal{N}(0.12, 1.26)$
M4	$\beta \sim$ Gamma(4.82, 0.88) Mode: 3.36 Standard deviation: 1.93	$\eta^+ \sim$ Beta(0.01, 0.032) Mode: 0.51 Standard deviation: 0.42 $\eta^- \sim$ Beta(0.012, 0.021) Mode: 0.5 Standard deviation: 0.47	$\omega \sim \mathcal{N}(0.15, 1.42)$

to subjects is a random effect), and estimate this rate using a hierarchical Bayesian model (see Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009, for more details). This analysis proceeds in two steps. First, we approximate the marginal likelihood of each model by plugging the MAP estimates into the Laplace approximation (Bishop, 2006):

$$\begin{aligned}\ln p(D_s|m, \phi_m) &= \ln \int_{\theta_{ms}} p(D_s|\theta_{ms}, m) p(\theta_{ms}|m, \phi_m) \\ &\approx \ln p(D_s|\hat{\theta}_{ms}, m) + \ln p(\hat{\theta}_{ms}|m, \phi_m) \\ &\quad + \frac{K_m}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{H}|, \end{aligned} \tag{6}$$

where K_m is the number of parameters for model m and

$$\mathbf{H} = -\nabla\nabla \ln[p(D_s|\hat{\theta}_{ms}, m) p(\hat{\theta}_{ms}|m, \phi_m)] \tag{7}$$

is the Hessian matrix of second derivatives of the negative log posterior. The Laplace approximation assumes that the posterior is approximately Gaussian around the mode, which is a reasonable assumption when the amount of data is relatively large. The second step in our model comparison method is to use the marginal likelihood approximation as the likelihood in the hierarchical model described by Stephan et al. (2009). Each model is assumed to occur in the population with a latent frequency estimated using variational Bayesian inference. Once these frequencies have been estimated, we compute the *protected exceedance probability*, defined as the probability that a particular model is more frequent in the population than all the other models, averaged over the probability of the null hypothesis that all models are equally frequent. Since model comparison is not the focus of this paper, we refer the reader to Rigoux et al. (2014) and Stephan et al. (2009) for more details.

2.5. Empirical priors and cross-validation

Once subject-specific parameter estimates were obtained, we estimated empirical priors by maximum likelihood:

$$\hat{\phi}_m = \underset{\phi_m}{\operatorname{argmax}} \prod_s p(\hat{\theta}_{ms}|\phi_m), \tag{8}$$

where we have assumed that parameter estimates are independent and identically distributed across subjects. For convenience, we choose the prior to have a parametric form (typically in the exponential family), although this choice is not essential.

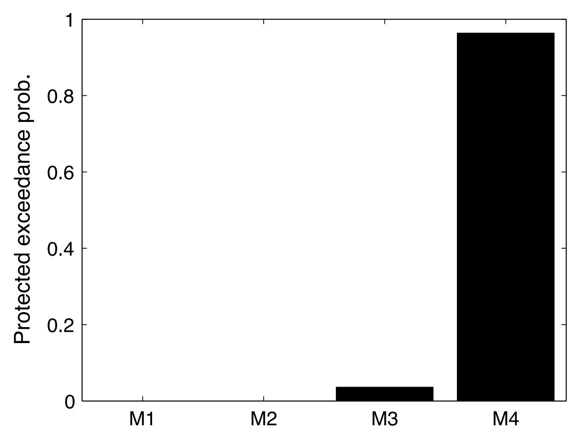

Fig. 1. Bayesian model comparison. The protected exceedance probability for each model. This is the probability that a particular model is more frequent in the population than all the other models, averaged over the probability of the null hypothesis that all models are equally frequent.

Empirical priors were estimated with D1 and then used for MAP estimation with D2. In order to evaluate the parameter estimates obtained using the empirical priors, we performed leave-one-block-out cross-validation with D2, whereby we fit the parameters on 3/4 blocks (using either empirical or uniform priors) and then computed the log likelihood of data on the held-out block. All cross-validation results are reported as averages of results across held-out blocks. Uniform priors were constructed to have a flat probability density function over the parameter range.

3. Results

3.1. Model comparison

We carried out Bayesian model comparison on D1 to determine the best model among those we considered. As shown in Fig. 1, M4 (dual learning rates + stickiness) had a decisively higher protected exceedance probability, indicating that it is with high probability more frequent in the population than the other models. Consequently, we focus on M4 for some of the analyses reported below.

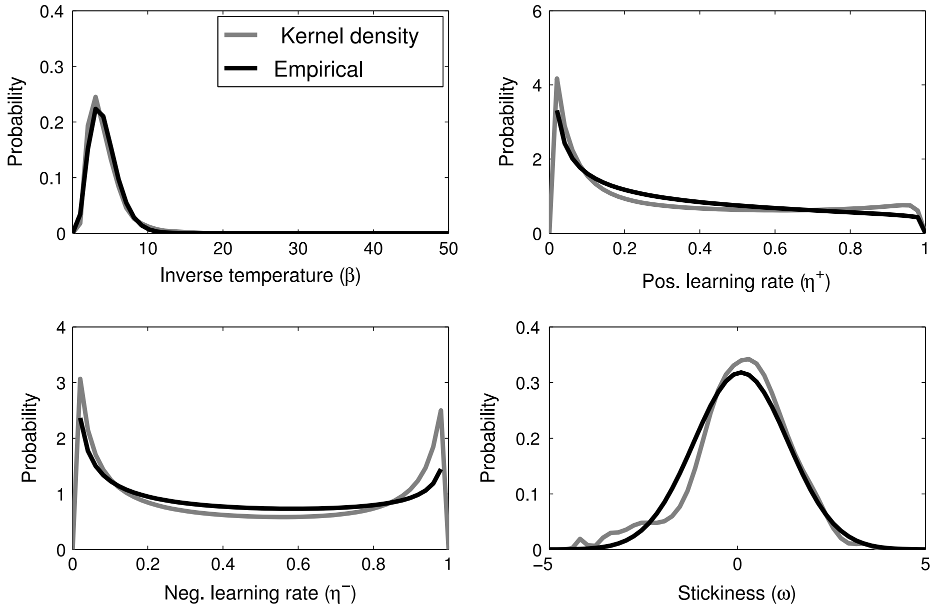

Fig. 2. Empirical priors for M4. Each panel shows a kernel density plot (using a Gaussian kernel with the optimal plugin bandwidth) of the parameter estimates for the group and an empirical prior obtained by fitting a parametric distribution to the parameter estimates using maximum likelihood.

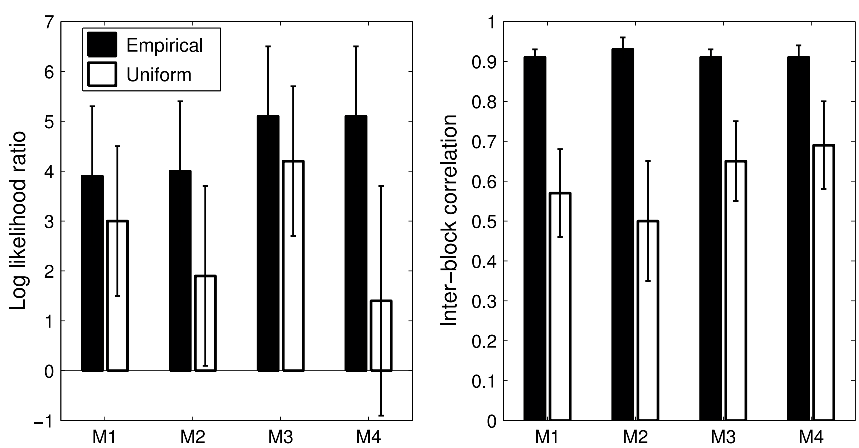

Fig. 3. Cross-validation results. (Left) Log ratio between the model posterior predictive likelihood and the likelihood under a random policy. (Right) Pearson correlation (estimated using Bayesian inference) between parameter estimates across different sets of training blocks. Error bars represent 95% credible intervals (across participants) around the posterior mean, after averaging over folds.

3.2. Empirical priors

The empirical priors fit to D1 for all 4 models are shown in Table 1. In order to provide some flexibility for future applications of these priors, parameters for the gamma and beta distributions are shown in the mode and standard deviation parameterization; this allows a modeler to (for example) reuse the mean but increase the standard deviation in order to obtain a vaguer prior. The empirical priors for the best-fitting model M4 are shown in Fig. 2. One noteworthy aspect of the hyperparameter estimates is that they typically do not vary substantially across models, indicating that the priors will be fairly robust across different variants of the models studied here.

We now turn to a quantitative evaluation of these empirical priors on a separate dataset (D2).

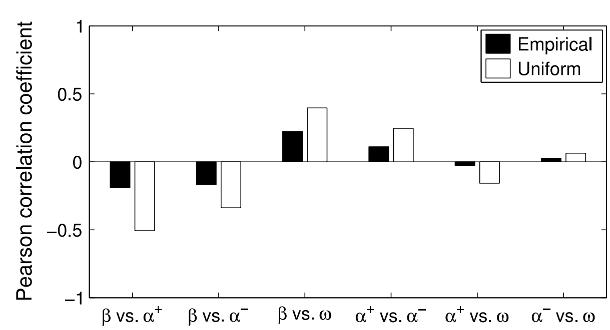

Fig. 4. Parameter estimate correlations. Pearson correlations (estimated using Bayesian inference) between the estimates for different pairs parameters. Correlations are computed across participants based on the entire dataset.

3.3. Quantitative evaluation

Cross-validation results on D2 are shown in Fig. 3(Left). The performance metric plotted here is the log ratio of each model's posterior predictive likelihood (i.e., the likelihood assigned to held-out trials conditional on the estimated parameters) and the likelihood under a random policy (i.e., choosing each option with equal probability). Thus, a log likelihood ratio of 0 indicates that a particular model predicts no better than chance. The empirical prior resulted in a higher log likelihood ratio compared to using

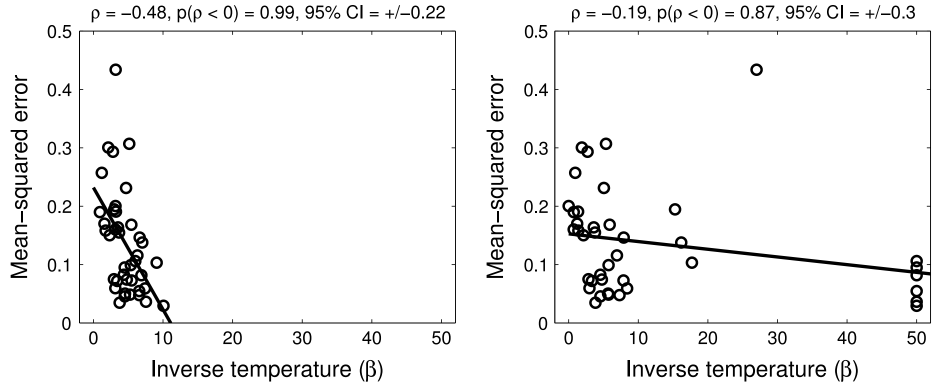

Fig. 5. Individual differences analysis. Relationship between inverse temperature and reward prediction accuracy, defined as the mean-squared error between predicted and observed reward, with a least-squares line superimposed. (Left) Empirical priors. (Right) Uniform priors. The analysis above each graph reports the results of a Bayesian correlation test: the posterior mean Pearson correlation coefficient, the posterior probability that the Pearson correlation coefficient is less than 0, and the 95% credible interval.

uniform priors (posterior probability of a difference greater than 0: 0.98, Bayesian t-test).[4] Moreover, uniform priors were not on average significantly different from chance ($p = 0.13$). These results confirm that empirical priors improve cross-validated predictive accuracy.

Parameter reliability tells a similar story (Fig. 3(Right)): parameter estimates are more strongly correlated across cross-validation folds for empirical priors compared to uniform priors (posterior probability of a difference greater than 0: 0.99, Bayesian t-test). Note that high reliability can be trivially achieved by using very strong priors, such that estimates for every subject are identical; however, this would deleteriously affect predictive accuracy (see also Scheibehenne & Pachur, 2015). The fact that we see both high accuracy and high reliability indicates that the empirical priors are achieving a good balance between regularization and data fit.

As mentioned in the Introduction, correlations between parameters is symptomatic of non-identifiability: When parameters trade off against each other to achieve equivalent likelihoods, these different parameter settings are indistinguishable and hence the parameter estimates are not interpretable. We found that the absolute values of parameter correlations are strongly reduced when using empirical priors (Fig. 4), with a posterior probability of 0.78 that the mean difference in correlations was greater than 0. This arises because the empirical priors suppress regions of the parameter space, thereby constraining the set of parameter configurations with high posterior probability. These constraints improve identifiability.

Finally, we turn to individual differences. One concern about using strong priors is that they regularize parameter estimates for different subjects closer together, and therefore eliminate variability which might be correlated with other individual difference measures. Note, however, that regularization also eliminates noise due to poor parameter estimates. Since this will reduce the variance of parameter estimates across subjects, it has the potential to *increase* correlations. We demonstrate this using D2, where we collected continuous ratings of reward probability on each trial. We reasoned that the inverse temperature, which can be viewed as a rough proxy for decision confidence, would be correlated with reward prediction accuracy (as measured by the mean-squared error between predicted and observed reward). Intuitively, participants with higher decision confidence, provided they are reasonably well-calibrated with their actual accuracy, will tend to exhibit higher accuracy (lower mean-squared error). Indeed, this is the case when using empirical priors, but the relationship is much weaker when using uniform priors (Fig. 5). While the empirical prior estimates are more clustered, they also show a tighter relationship with reward prediction accuracy.[5]

4. Conclusion

We have argued that empirical priors offer several distinct advantages over uniform priors when fitting RL models: They improve predictive accuracy, reliability, identifiability, and measurement of individual differences. The empirical priors estimated here can be potentially applied to a wide range of models and tasks that share similar parameterizations. We noted that the priors are robust across parameterizations, suggesting that they are fairly transferable.

This paper has focused on RL models, but empirical priors could benefit other areas of cognitive science. For example, Smith (2006) has pointed out that a large class of categorization models suffers from "colliding parameters": the similarity and choice components effectively cancel each other out (but see Navarro, 2007). This is, in essence, an identifiability issue, and therefore can potentially be remedied by using empirical priors. Many other areas involve the quantification of individual differences in terms of computational models, particularly in cognitive neuroscience and psychiatry. To the extent that empirical priors aid the analysis of individual differences, these domains will similarly benefit.

An alternative to empirical priors is to use hierarchical modeling (Gelman & Hill, 2006). Rather than fitting a prior to one dataset and then applying it to another dataset, one could estimate the prior on the same dataset. The advantage of this approach is that the priors are potentially better tuned to the dataset at hand. The disadvantage is that the risk of overfitting is greater, in the sense that both parameters and hyperparameters are being fit to the data, in contrast to the use of empirical priors, where only the parameters are fit after the priors are obtained (see Scheibehenne & Pachur, 2015, for further discussion of pitfalls with hierarchical modeling). Another disadvantage is that hierarchical modeling is somewhat more computationally involved, although new software developments for generic Bayesian inference (e.g., Stan, JAGS, BUGS) are making this task easier. A comprehensive comparison of empirical and hierarchical priors is an important task for future work.

The approach adopted in this paper, while motivated by Bayesian concepts (priors, posteriors, etc.), is not fully Bayesian:

[4] For this and all following statistical tests, we use the R package BayesianFirstAid to compute parameter estimates and posterior probabilities.

[5] Six participants were fit with inverse temperatures at the parameter upper bound (50), but the results do not change materially when these participants are removed.

model-fitting is based on point estimation rather than computing the full posterior. While this will be unsatisfying for the hardcore Bayesian, the goal of this paper was not to defend point estimation but rather to show how the use of empirical priors can improve widely used model-fitting techniques. The usefulness of informative priors has been amply demonstrated in cognitive science (Vanpaemel, 2011; Vanpaemel & Lee, 2012), and there are a variety of ways that such priors can be constructed and employed. Many RL researchers are comfortable with point estimation, and this paper was designed to be surgical in its modification of current practices. Nonetheless, empirical priors are perfectly compatible with a fully Bayesian approach.

Acknowledgments

I am grateful to Nathaniel Daw and Josh Buckholtz for helpful discussions. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury.
Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of Cognitive Neuroscience*, *25*, 1807–1823.
Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *Decision making, affect, and learning: attention and performance XXIII. Vol. 23* (p. 1).
Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*, 603–616.
Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.
Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*, 1062–1068.
Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*, 16311–16316.
Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*, 1–8.
Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The Journal of Neuroscience*, *29*, 13524–13531.
Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*, 154–162.
Navarro, D. J. (2007). Similarity, distance, and categorization: A discussion of Smith's (2006) warning about "colliding parameters". *Psychonomic Bulletin & Review*, *14*, 823–833.
Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, *32*, 551–562.
Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage*, *84*, 971–985.
Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, *22*, 391–407.
Smith, J. D. (2006). When parameters collide: A warning about categorization models. *Psychonomic Bulletin & Review*, *13*, 743–751.
Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*, 1004–1017.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press.
Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*, 106–117.
Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.