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
Available online 24 February 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.

© 2016 Elsevier Inc. All rights reserved.

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.

http://dx.doi.org/10.1016/j.jmp.2016.01.006
0022-2496/© 2016 Elsevier Inc. All rights reserved.

[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.

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:

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

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.

Download English Version:

https://daneshyari.com/en/article/326321

Download Persian Version:

https://daneshyari.com/article/326321

Daneshyari.com