Assignment 1

By Kenneth C. Enevoldsen

Github: https://github.com/KennethEnevoldsen/Advanced-Cognitive-Modelling/tree/master/assignment_2

Explain the delta learning rule with reference to the Rescorla-Wagner model of classical conditioning. Explain the softmax function with reference to the Luce choice rule as a model of choice. Explain how the delta rule and the softmax function work together in the Q-learning algorithm as a model of sequential choice on bandit tasks. Present the parameters of the Q-learning algorithm in terms of their psychological interpretations. Evaluate the Q-learning model by comparison to the choice kernel model, as it is presented in the Wilson and Collins paper, using the a priori methods presented in the paper and in class.

The delta learning rule, first invented by Rescorla & Wagner (1972) is a method for updating internal states Q^k of an agent according to observation r and is today frequently used in neural. It is defined as:

$$Q_t^k = Q_{t-1}^k + \alpha(r_{t-1} - Q_{t-1}^k)$$

Where t is the time step[1], α is the learning rate and $(r - Q^k)$ is the prediction error, sometimes denoted using the delta notation, hence the name. The delta rule thus updated it internal state Q_t^k according to its degree of prediction error and its rate of learning. In the case of this assignment the internal state denotes an agents belief about the profit of bandit in the sequential choice bandit task, with higher Q indicating higher expected utility of the given choice and r indicating the reward. Given the updated internal states the agent make a choice using the softmax choice rule:

$$p_t^k = \frac{exp(\beta Q_t^k)}{\sum_{i=1}^k exp(\beta Q_t^i)}$$

Where p_t^k denotes the probability of choosing k and β denotes the behavioural temperature i.e. the degree of exploration. The softmax serve a couple of purposes, first it makes it so that

$$\sum_{i=1}^{k} P_t^i = 1$$

Following the second law of probability (Pleskac, 2015), note however that this can be achieved simply using

$$p_t^k = \frac{Q_t^k}{\sum_{i=1}^k Q_t^k}$$

Following the Luce choice axiom (Pleskac, 2015). However this limits the Q_t^k to be positive, which in the given scenario is an unreasonable assumption. Similarly the softmax included the behavioural temperate which allow one to model the degree of exploration.

[1] Note that t can be disregarded without consequence

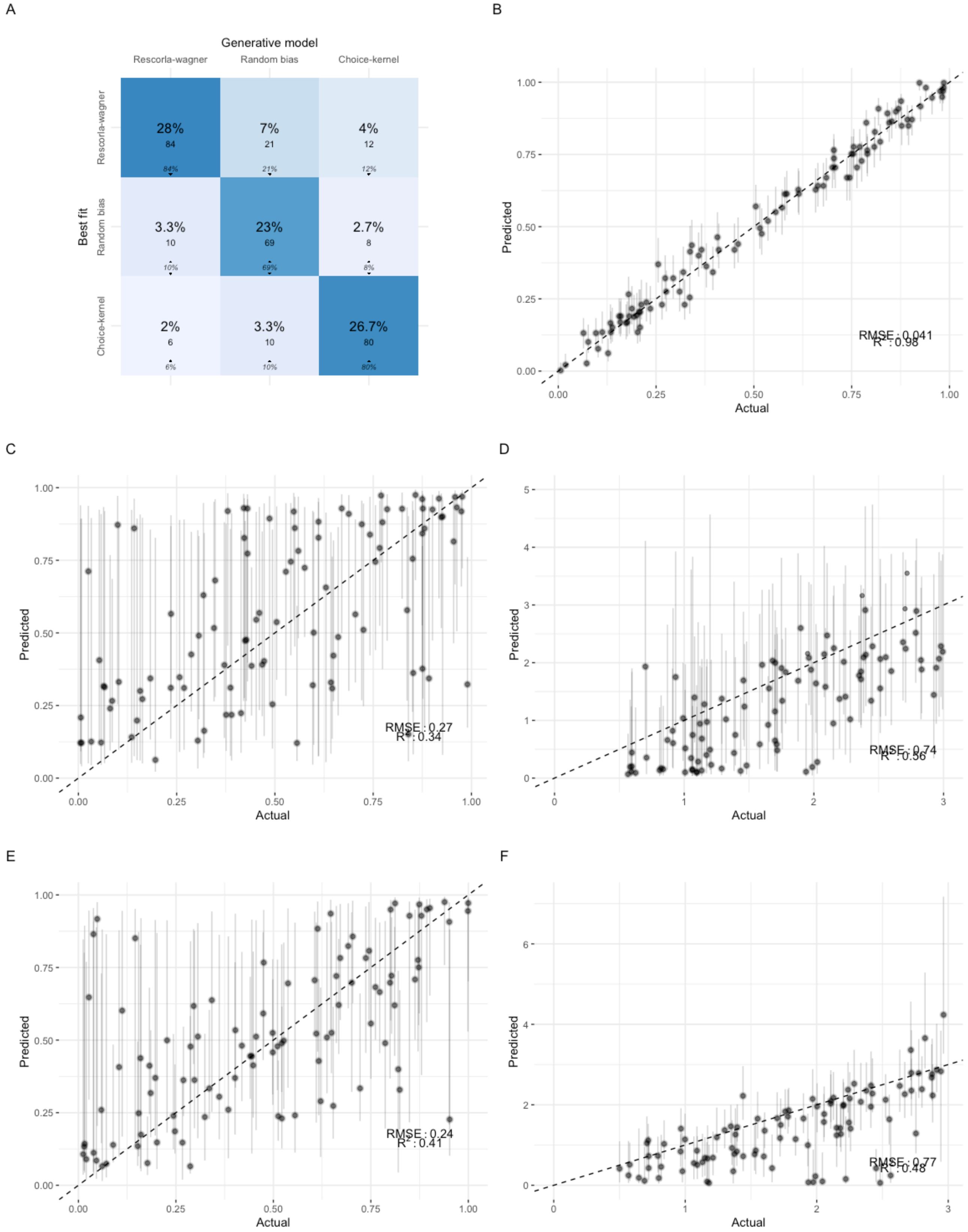

Figure 1: A) A confusion matrix showing the model recovery, darker colours shades indicate higher values. B-F) The parameter recovery of the agents plotting the predicting measured using MAP against the actual. The stippled line indicate the best case scenario and the error bar at each point indicate the 0.89 CI. The parameters is as follows; the random bias agent's parameter θ (B), the Rescorla-Wagner agent α and β (C-D, respectively), and the choice kernels α and β (E-F, respectively).

Thus the softmax function and the delta learning rule work together in the Q-learning model to model choice, in this case in the bandit task, where the delta rule update prediction according to the prediction error and learning rate. The softmax normalises the prediction into probabilities, with the highest predicted value obtaining the highest probability. The unknown parameters α and β also have a psychological explanation with α corresponding to rate of learning or speed of conditioning (Rescorla & Wagner, 1972), while β indicate the tradeoff between exploration and exploitation (Pleskac, 2015).

Model recovery and parameter recovery

In the following we have implemented three agents. A random bias agent, and 2 Q-learning agents, the choice kernel and the Rescorla Wagner agent, with the only difference being in their delta rule, where the Rescorla Wagner model delta rule uses the prediction error between reward and expected reward the choice kernel model utilises the delta rule to model the prediction error between its actual choice and its 'choice kernel', i.e. its choice preference. For more on these model see (Wilson & Collins, 2019).

For each agent were simulated acting in a bandit task with 100 trials, and two bandits, with the reward 1 and 0.5 with a chance probability winning 0.3 and 0.8 respectively. For each of the agents its start parameters - α and β for the Q-learning agents and a bias term θ for the biased agent - were randomised within reasonable bounds. Lastly, for each of the three simulation all three models were fitted using JAGS. The parameter estimates used the slightly restrictive priors:

$$\theta, \alpha \sim uniform(0,1)$$
$$\beta \sim gamma(1,1)$$

The above process were repeated 100 times and the best model for each repetition was recorded with the results seen in figure 1A. Notably, the model are distinguishable by each other with the best fit generally being the model itself. For each of the model fit to its respective data a parameter recovery was performed for each of the agent. As seen in figure 1B-1F. All parameters seem to be well recovered, albeit with the α parameter containing a fair bit of noise. Notably β parameter seem to be generally be predicted lower than actual, which might be due to the gamma prior.

References

Pleskac, T. J. (2015). Decision and choice: Luce's choice axiom. *International Encyclopedia of the Social & Behavioral Sciences*, *5*, 895–900.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Wilson, R., & Collins, A. (2019). Ten simple rules for the computational modeling of behavioral data. https://doi.org/10.31234/osf.io/46mbn