Learning under uncertainty: a comparison between R-W and Bayesian approach

Authored by:

He Huang
Martin Paulus

Abstract

Accurately differentiating between what are truly unpredictably random and systematic changes that occur at random can have profound effect on affect and cognition. To examine the underlying computational principles that guide different learning behavior in an uncertain environment, we compared an R-W model and a Bayesian approach in a visual search task with different volatility levels. Both R-W model and the Bayesian approach reflected an individual's estimation of the environmental volatility, and there is a strong correlation between the learning rate in R-W model and the belief of stationarity in the Bayesian approach in different volatility conditions. In a low volatility condition, R-W model indicates that learning rate positively correlates with lose-shift rate, but not choice optimality (inverted U shape). The Bayesian approach indicates that the belief of environmental stationarity positively correlates with choice optimality, but not lose-shift rate (inverted U shape). In addition, we showed that comparing to Expert learners, individuals with high lose-shift rate (sub-optimal learners) had significantly higher learning rate estimated from R-W model and lower belief of stationarity from the Bayesian model.

1 Paper Body

Learning and using environmental statistics in choice-selection under uncertainty is a fundamental survival skill. It has been shown that, in tasks with embedded environmental statistics, subjects use sub-optimal heuristic Win-Stay-Lose-Shift (WSLS) strategy (Lee et al. 2011), and strategies that can be interpreted using Reinforcement Learning model (Behrens et al. 2007), or Bayesian inference model (Mathys et al. 2014; Yu et al. 2014). Value-based model-free RL model assumes subjects learn the values of chosen options using a prediction error that is scaled by a learning rate (Rescorla and Wagner, 1972; Sutton and Barto, 1998). This learning rate can be used to measure an individual?s reaction to environmental volatility (Browning et al. 2015). Higher learning rate

is usually associated with a more volatile environment, and a lower learning rate is associated with a relatively stable situation. Different from traditional (model-free) RL model, Bayesian approach assumes subjects make decisions by learning the reward probability distribution of all options based on Bayes? rule, i.e., sequentially updating the posterior probability by combining the prior knowledge and the new observation (likelihood function) over time. To examine how environment volatility may influence this inference process, Yu & Cohen 2009 proposed to use a dynamic belief model (DBM) 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

that assumes subjects update their belief of the environmental statistics by balancing between the prior belief and the belief of environmental stationarity, in which a belief of high stationarity will lead to a relatively fixed belief of the environmental statistics, and vice versa. Though formulated under different assumptions (Gershman 2015), those two approaches share similar characteristics. First, both the learning rate in RL model and the belief of stationarity in DBM reflect an individual?s estimation of the environmental volatility. In a highly volatile environment, one will be expected to have a high learning rate estimated by RL model, and a belief of low stationarity estimated by DBM. Second, though standard RL only updates the chosen option?s value and DBM updates the posterior probability of all options, assuming maximization decision rule (Blakely, Starin, & Poling, 1988), both models will lead to qualitatively similar choice preference. That is, the mostly rewarded choice will have an increasing value in RL model, and increasing reward probability in DBM, while the often-unrewarded choice will lead to a decreasing value in RL model, and decreasing reward probability in DBM. Thirdly, they can both explain Win-Stay strategy. That is, under the maximization assumption, choosing the option with maximum value in RL model, the rewarded option (Win) will reinforce this choice (i.e. remain the option with the maximum value) and thus will be chosen again (Stay). Similarly, choosing the option with the maximum reward probability in DBM, the rewarded option (Win) will also reinforce this choice (i.e. remain the option with the maximum reward probability) and thus will be chosen again (Stay). While both approaches share some characteristics as mentioned above and have showed strong evidence in explaining the overall subjects? choices in previous studies, it is unclear how they differ in explaining other behavioral measures in tasks with changing reward contingency, such as decision optimality, i.e., percentage of trials in which one chooses the most likely rewarded option, and lose-shift rate, i.e., the tendency to follow the last target if the current choice is not rewarded. In a task with changing reward contingency (e.g., 80%:20% to 20%:80%), decision optimality relies on proper estimation of the environmental volatility, i.e., how frequent change points occur, and using proper strategy, i.e. staying with the mostly likely option and ignoring the noise before change points (i.e. not switching to the option with lower reward rate when it appears as the target). Thus it is important to know how the parameter in each model (learning rate vs. the belief of stationarity) affects decision optimality in tasks with different volatility. On the other hand, lose-shift can be explained as a heuristic decision policy that is used to reduce

a cognitively difficult problem (Kahneman & Frederick, 2002), or as an artifact of learning that can be interpreted in a principled fashion (using RL: Worthy et al. 2014; using Bayesian inference: Bonawitz et al. 2014). Intuitively, when experiencing a loss in the current trial, in a high volatility environment where change points frequently occur, one may tend to shift to the last target; while in a stable environment with fixed reward rates, one may tend to stay with the option with the higher reward rate. That is, the frequency of using lose-shift strategy should depend on how frequent the environment changes. Thus it is also important to examine how the parameter in each model (learning rate vs. the belief of stationarity) affects lose-shift rate under different volatility conditions. However, so far little is known about how a model-free RL model and a Bayesian model differ in explaining decision optimality and lose-shift in tasks with different levels of volatility. In addition, it is unclear if parameters in each model can capture the individual differences in learning. For example, if they can provide satisfactory explanation of individuals who always choose the same choice while disregarding feedback information (No Learning), individuals who always choose the most likely rewarded option (expert), and individuals who always use the heuristic win-stay-lose-shift strategy. Here we aim to address the first question by investigating the relationship between decision optimality and lose-shift rate with parameters estimated from an Rescorla-Wagner (R-W model) and a Bayesian model in three volatility conditions (Fig 1a) in a visual search task (Yu et al. 2014): 1) stable, where the reward contingency at three locations remains the same (relative reward frequency at three locations: 1:3:9), 2) low volatility, where the reward contingency at three locations changes (e.g. from 1:3:9 to 9:1:3) based on N (30, 1) (i.e. on average change points occur every 30 trials), and 3) high volatility, where the reward contingency changes based on N (10, 1) (i.e. on average change points occur every 10 trials). For the second question, we will examine how the two models differ in explaining three types of behavior: No Learning, Expert, and WSLS (Fig 1b).

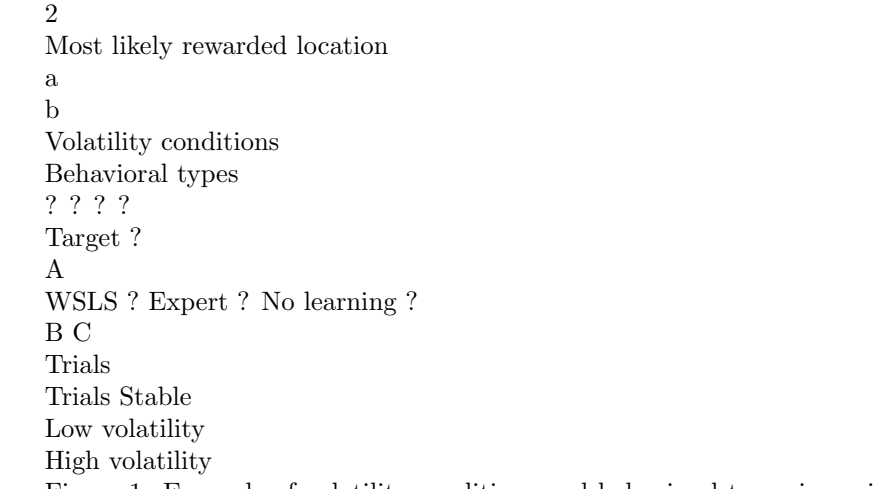

Figure 1: Example of volatility conditions and behavioral types in a visual search task. a. Example of three volatility conditions. b. Example of three

behavioral types. Colors indicate target location in each trial. WSLS: Win-stay-Lose-shift (follow last target). Expert: always choose the most likely location. No Learning: always choose the same location.

2

Value-based RL model

Assuming a constant learning rate, the Rescorla-Wagner model takes the following form (Rescorla and Wagner, 1972; Sutton and Barto, 1998): Vit+1 = Vit + ?(Rt ? Vit)

(1)

where ? is the learning rate, and Rt is the reward feedback (0-no reward, 1-reward) for the chosen option i in trial t. In this paper, we assume subjects use a softmax decision rule for all models as follows: e?Vi p(i) = P ?Vj je

(2)

where ? is the inverse decision temperature parameter, which measures the degree to which subjects use the estimated value in choosing among three options (i.e. a large ? approximates ?maximization? strategy). This model has two free parameters = {?, ?} for each subject. 2.1

Simulation in three volatility conditions

Learning rate is expected to increase as the volatility increases. To show this, we simulated three volatility conditions (Stable, Low and High volatility, Fig 2ab) and the results are summarized in Table1. For each condition, we simulated 100 runs (90 trial per run) of agents? choices with ? ranges from 0 to 1 with an increment of 0.1 and fixed ? = 20. As is shown in Fig 2a, decision optimality in a stable and a low volatility environment has an inverted U shape as a function of learning rate ?. It is not surprising, as in those conditions, where one should rely more on the long term statistics, if the learning rate is too high, then subjects will tend to shift more due to recent experience (Fig 2b), which would adversely influence decision optimality. On the other hand, in a high volatility environment, decision optimality has a linear correlation with the learning rate, suggesting that higher learning rate leads to better performance. In fact, the optimal learning rate increases as the environmental volatility increases (i.e. the peak of the inverted U should shift to the right). On the other hand, across all volatility conditions, lose shift rate increases as learning rate increases (Fig 2b), except for learning rate=0. It is not surprising as zero learning rate indicates subjects make random choices, thus it will be close to 1/3. 2.2

Simulation of three behavioral types

To examine if learning rate can be used to explain different types of learning behavior, we have simulated three types of behavior (No Learning, Expert and WSLS, Fig 1b) in a low volatility condition. In particular, we simulated 60 runs (90 trials per run) of target sequences with a relative reward frequency 1:3:9 that changes based on N (30, 1), and generated three types of behavior for 3

Table 1: R-W model: Influence of learning rate ? Condition

Decision optimality

Lose-shift rate

Stable Low volatility High volatility

Inverted U shape, ?optimal = low Inverted U shape, ?optimal = medium Positive linear relationship, ?optimal = high

Positive linear Positive linear Positive linear

each run. For each simulated behavior type, R-W model was fitted using Maximum Likelihood Estimation with ? ranges from 0 to 1 with an increment of .025 and ? = 20. Based on what we have shown in 2.1, in a low volatility condition where decision optimality has an inverted U shape as a function of learning rate, individuals that perform poorly will be expected to have a low learning rate, and individuals that use heuristic WSLS strategy will be expected to have a high learning rate. We confirmed this in simulation (Fig 2c), that agents with the same choice over time (No Learning) have the lowest learning rate, indicating their choices have little influence from the reward feedback. Expert agents have the medium learning rate indicating the effect of long-term statistics. Agents that strictly follow WSLS have the highest learning rate, indicating their choices are heavily impacted by recent experience. Results for learning rate estimation of three behavioral types in stable and high volatility condition can be seen in Supplementary Figure S1.

Optimal choice% 1

Stable Non N(30,1) N(10,1) Low volatility High volatility

0.9 0.8 0.7 0.6

0.4

0.8

0.35

0.7

0.25 0.2 0.15

0.5

0.1

0.4

0.05

0.3

0

0.2

0.4

eta

0.6

0.8

c

Lose shift%

0.3

Lose shift%

Optimal choice%

b

Learning rate ?

a

1

0

Learning rate Low volatility
0.6 0.5 0.4 0.3 0.2 0.1
0
Learning rate ?
0.2
0.4
0.6
0.8
0
1
eta
Learning rate ?
No learning
Expert
WSLS
Learning type

Figure 2: R-W simulation. a. Percentage of trials in which agents chose the optimal choice (the most likely location) as a function of learning rate in three volatility conditions. b. Lose shift rate as a function of learning rate in three volatility conditions. c. Learning rate estimation of three simulated behavior types in low volatility condition. Errorbars indicate standard error of the mean across simulation runs.

Simu_all.m
Simu_Fix_TD.m
Simu_all_new.m
3
A Bayesian approach
Simu_target.m

Here we compare above R-W model to a dynamic belief model that is based on a Bayesian hidden Figuremodel, 2 where we assume subjects make decisions by using the inferred posterior target Markov probability st based on the inferred hidden reward probability ? t and the reward contingency mapping bt (Equation 3). To examine the influence of volatility, we assume (?? t , bt) has probability ? of remaining the same as the last trial, and 1-? of being drawn from the prior distribution p0 (?? t , bt) (Equation 4). Here ? represents an individual?s estimation of environmental stationarity, which contrasts with learning rate ? in R-W model. For model details, please refer to Yu & Huang 2014. ? 1 1 1 (3 , 3 , 3), ? ? ? ? (? ? h , ?m , ?l), ? ? ?(?h , ?l , ?m), P (st —?? t , bt) = (?m , ?h , ?l), ? ? ?(?m , ?l , ?h), ? ? ? ?(?l , ?h , ?m), ? (?l , ?m , ?h), 4

bk bk bk bk bk bk bk
=1 =2 =3 =4 =5 =6 =7

(3)

P (?? t , bt —st?1) = ?P (?? t?1 , bt?1 —st?1) + (1 ? ?)p0 (?? t , bt)

(4)

where ? t is the hidden reward probability, bt is the reward contingency mapping of the probability to the options, st?1 is the target history from trial

1 to trial t ? 1. We also used softmax decision rule here (Equation 2), thus this model also has two free parameters = {?, ?}. 3.1

Simulation in three volatility conditions

Belief of stationarity ? is expected to decrease as the volatility increases, as subjects are expected to depend more on the recent trials to predict next outcome (Yu &Huang 2014). We have shown this in three simulated conditions under different volatility (Fig 3ab) and the results are summarized in Table 2. For each simulated condition (stable, low and high volatility), we simulated 100 runs (90 trials per run) for agents? choices with ? ranges from 0 to 1 and fixed ? = 20. As is shown in Fig 3a, in a stable condition, decision optimality increases as ? increases, indicating a fixed belief mode (i.e. no change of the environmental statistics) is optimal in this condition. In the other two volatile environments, decision optimality also increases as ? increases, but both drop as ? approaches 1. It is reasonable as in volatile environments a belief of high stationarity is no longer optimal. On the other hand, lose shift rate in all conditions (Fig 3b) have an inverted U shape as a function of alpha, where ? = 0 leads to a random lose shift rate (1/3), and ? = 1(fixed belief model) leads to the minimal lose shift rate. 3.2

Simulation of three behavioral types

To examine if an individual?s belief of environmental stationarity can be used to explain different types of learning behavior, we fit DBM using Maximum Likelihood Estimation with the simulated behavioral data in 2.2. DBM results suggest that WSLS has a significantly lower belief of stationarity comparing to Expert behavior (Fig 3c), which is consistent with the higher volatility estimation reflected by a higher learning rate than Expert from R-W model (Fig 2c). Simulation results also suggest that No Learning agents have a significantly lower belief of stationarity than Expert learners, but not different from WSLS. However, the comparison of model accuracy between R-W and DBM (Fig 3d) shows DBM outperforms R-W in predicting Expert and WSLS behavior (p = .000), but it does not perform as well in No Learning behavior where R-W has significantly better performance (p = .000). Model accuracy is measured as the percentage of trials that the model correctly predicted subjects? choice. Thus further investigation is needed to examine the validity of using DBM in explaining poor learners? choice behavior in this task. Results for ? in stable and high volatility condition can be seen in Supplementary Figure S2.

a

b

c

Optimal choice% 1

1

0.8

Non

Stable N(30,1) N(10,1) Low volatility High volatility

0.8

0.7 0.6 0.5 0.4 0.3

0

0.2
Non N(30,1) N(10,1)
0.9
Lose Shift%
Optimal choice%
0.9
0.4
0.6
alpha
0.8
1
Belief of stationarity ?
d Belief of stationarity
Lose shift%
0.7
Model accuracy
1
1
0.9
0.9
0.8
0.8
0.7
0.7
0.6
? 0.6
0.5
0.5
0.5
0.4
0.4
0.4
0.3
0.3
0.3
0.2
0.2
0.2
0.1
0.1
0
0
0
0.2
0.4
0.6

alpha
0.8
1
Belief of stationarity ?
RW DBM
0.6
0.1 No learning
Expert
WSLS
0
No learning
Expert
WSLS
Learning type
Learning type
(Low volatility)
(Low volatility)
Compare_Simu_TD_DBM.m

Figure 3: DBM simulation. a. Percentage of trials in which agents chose the optimal choice (the most CompareNcp_10_30_DBM.m Simu_DBM_bk_low.m likely location) as a function of ? in three volatility conditions. b: Lose shift rate as a function of ? in three volatility conditions. c. ? estimation of three types of behavior in low volatility condition check_simuDBM_low.m DBM_Softmax_Ncp, CPbetween R-W and DBM in low volatility condition. (N (30, 1)). d. Model performance comparison

Add optimal alpha in Non, 45, 5 30, 10 CompareNcp_10_30_45_DBM.m Figure 3

check_simuDBM.m

Table 2: DBM: Influence of stationarity belief ?

4

Condition

Decision optimality

Lose-shift rate

Stable Low volatility High volatility

Positive linear relationship, ?optimal =1 Inverted U shape, ?optimal = high Inverted U shape, ?optimal = medium

Inverted U shape Inverted U shape Inverted U shape

Experiment

We applied the above models to two sets of data in a visual search task: (1) stable condition with no change points (from Yu & Huang, 2014) and (2) low volatility condition with change points based on N (30, 1). For both data sets, we fitted an R-W model and DBM for each subject, and compared learning rate ? in R-W and estimation of stochasticity (1 ? ?) in DBM, as well as how they correlate with decision optimality and lose shift rate. For (2), we also looked at how model parameters differ in explaining No Learning, Expert and WSLS behavior. 4.1 4.1.1

Results Stable condition

In a visual search task with relative reward frequency 1:3:9 but no change points, we found a significant correlation between ? estimated from R-W model and 1-? from DBM (r2 = .84, p = .0001, Fig 4a), which is consistent with the hypothesis that both the learning rate in R-W model and the belief of stochasticity in the Bayesian approach reflects subjects? estimation of environmental volatility. We also examined the relationship between decision optimality (optimal choice%) (Fig 4b) and lose-shift rate (Fig 4c) with ? and 1-? respectively. As is shown in Fig 4b, in this stable condition, decision optimality decreases as the learning rate increases, as well as the belief of stochasticity increases, which is consistent with Fig 2a (red, for ? ? .1) and Fig 3a (red). For lose-shift rate, there is a significant positive relationship between lose-shift% and ?, as shown previously in Fig 2b (red), and an inverted U shape as suggested in Fig 3b (red). There are no significant differences in the prediction accuracy of R-W model and DBM (R-W: .81+/-.03, DBM: .81+/-.03) or inverse decision parameters (p ¿ .05).

b
0.4 0.3
c
0.8 0.7 0.6 0.5
0.2
1
0.1
0.8
Lose-shift%
alpha-DBM 1-? (DBM)
0.5
0 -0.1 -0.2
0
0.2
0.4
0.6
0.8
1
?eta-TD (RW)
Optimal choice%
0.6
Optimal choice%
1 0.9
0
0.5
2 (RW) Lose-shift%
1
0.6 0.4 0.2 0
0
0.5

2 (RW)
Optimal choice%
1 0.9 0.8 0.7 0.6 0.5
0
0.8
Lose-shift%
Optimal choice%
a
1
0.5
1-, (DBM) Lose-shift%
0.6 0.4 0.2 0
0
0.5
1-, (DBM)

Figure 4: Stable condition: R-W vs. DBM. a. Relationship between learning rate ? in R-W model and 1-? in DBM. b. Optimal choice% as a function of ? and 1-?. c. Lose-shift% as a function of ? and 1-?. 4.1.2

Low volatility condition

In a visual search task with relative reward frequency 1:3:9, and change of the reward contingency based on N (30, 1) (3 blocks of 90 trials/block), we looked at the correlation between model parameters, their correlation with decision optimality and lose-shift rate, as well as how model parameters differ in explaining different types of behavior. 6

Subjects (N=207) were grouped into poor learners (optimal choice%¡ .5, n = 63), good learners (.5? optimal choice% ? 9/13, n = 108) and expert learners (optimal choice% ¿ 9/13, n = 36) based on their performance (percentage of trials started from the most likely rewarded location). Consistent with what we have shown previously (Fig 3d-No Learning), R-W model outperformed DBM in poor learners (p = .000). Similar as in stable condition (Fig 4a), among good and expert learners, there is a significant positive correlation between ? and 1-? (Fig 5b, r2 = .35, p = .000). The relationship between decision optimality and lose-shift% is shown in Fig 5c. As is shown, in this task where change points occur with relatively low frequency (N (30, 1)), lose shift% has an inverted U shape as a function of optimal choice%, indicating that a high lose-shift rate does not necessarily lead to better performance.

a
b Model accuracy
c Good & Expert learners
0.9
0.6
0.8 0.75 0.7
0.5
0.5
0.4
0.4

0.3 0.2
0
0.6
-0.1
Good(.5-9/13)
Expert(¿9/13)
0.3 0.2 0.1
0.1
0.65
Poor(¡.5)
lose shift%
RW DBM
1-, (DBM) 1-?(DBM)
Model acc
0.85
Optimal choice % vs. Lose-shift%
0.6
0 0
0.2
0.4
0.6
0.8
?2 (RW) (RW)
1
-0.1
.1-.2 .2-.3 .3-.4 .4-.5 .5-6 .6-.7 .7-.8 .8-9
optimal choice%

Figure 5: a. Prediction accuracy in poor, good and expert learners. b. Correlation between ? from R-W and 1-? from DBM. c. Correlation between optimal choice% and lose-shift%. Belief of stationarity ?

Compare_TD2_DBM_overall.m Next, we looked at how each model parameter correlates with decision optimality and lose-shift rate. For decision optimality (Fig 6ab), consistent with simulation result, it has an inverted U shape as a function of learning rate ? in R-W model (Fig 6a), while it is positively correlated with ? in DBM (Fig 6b). For lose-shift rate (Fig 6cd), also consistent with simulation result, it is positively correlated with ? in R-W (Fig 6c), while having an inverted U shape as a function of ? in DBM (Fig 6d). a

Optimal choice % (RW)
b
0.7
c
Optimal choice % (DBM)
d
Lose-shift % (RW)
0.6
0.7

0.7
0.65
0.5
0.5
0.4
0.5 0.45 0.4
0.3
0.35
0-.2
.2-.4
.4-.6
2 (RW)
.6-.8
Learning rate ?
.8-1
0.2
0.3
0.2 0.1
0.25
0.4
0.4
0.2
0.3
0.45
LS%
0.55
0.55
LS%
0.6
0.6
0.5
0.6
Optimal choice%
Optimal choice%
0.65
Lose-shift % (DBM)
0-.2
.2-.4
.4-.6
.6-.8
.8-1
,(DBM)
Belief of stationarity ?
0.1
0-.2
.2-.4

.4-.6
.6-.8
2 (RW) Learning rate ?
.8-1
0
0-.2
.2-.4
.4-.6
.6-.8
.8-1
Belief of ,(DBM) stationarity ?

Figure 6: Decision optimality and lose-shift rate. a. Optimal choice% as a function of ? in R-W model. b. Optimal choice% as a function of ? in DBM. c. Lose-shift% as a function of ? in R-W model. d. Lose-shift% as a function of ? in DBM. In addition, we examinedCompare_TD2_DBM_overall.m how poor, expert learners and individuals with a high lose-shift rate (LS, lose shift%¿ .5 and optimal choice% ¡ 9/13, n = 51) differ in model parameters (Figure 7). Consistent with what we have shown (Fig 2c), those three different behavioral types had significantly different learning rate (one-way ANOVA, p = .000) and each condition is significant from each other (p = .000 for t test across conditions), in which poor learners had the lowest learning rate while subjects with high lose-shift rate had the highest learning rate (Fig 7a). Belief of stationarity from DBM also confirmed what we have shown (Fig 3c), that expert subjects had significantly higher belief of stationarity (one-way ANOVA, p = .003, and p = .004 for t test comparing to Poor subjects and p = .000 comparing to LS subjects). It also suggested that poor learners did not differ from LS subjects (p ¿ .05), though DBM had a lower accuracy in predicting poor learners? choices (Fig 5a). No significant difference of inverse decision parameter ? was found between R-W and DBM for 7

expert and LS subjects (p ¿ .05), but it was significantly lower in poor learners estimated in DBM (Supplementary Figure S3). b

Learning rate
Belief of stationarity
0.8
0.9
0.7
0.8
0.6
0.7 0.6
0.5 0.4 0.3
, (DBM)
1-? (DBM)
? (RW)
a
0
0.4 0.3

0.2 0.1
0.5
0.2
Poor
Expert
0.1
LS
0
Poor
Expert
LS
Poor
Expert
LS

Figure 7: Parameter estimation for different behavioral types. a. Learning rate in R-W model. b. Compare_RW_DBM_overall.m Belief of stationarity in DBM. Figure 7

5

Compare_TD2_DBM_overall.m

Discussion

In this paper we compared an R-W model and a Bayesian model in a visual search task across different volatility conditions, and examined parameter differences for different types of learning behavior. We have shown in simulation that both the learning rate ? estimated from R-W and the belief of stochasticity 1 ? ? estimated from DBM have strong positive correlation with increasing volatility, and confirmed that they are highly correlated with behavioral data (Fig 4a and Fig5b). This suggests that both models are able to reflect an individual?s estimation of environmental volatility. We also have shown in simulation that R-W model can differentiate No Learning, Expert and WSLS behavioral types with (increasing) learning rate, and DBM can differentiate Expert and WSLS behavioral types with (increasing) belief of stochasticity, and confirmed this with behavioral data in a low volatility condition. A few other things to note here: Correlation between decision optimality and lose-shift rate. Here we have provided a modelbased explanation of using lose-shift strategy and how it is related to decision optimality. 1) R-W model suggests that, across different levels of environmental volatility, the frequency of using loseshift is positively correlated with learning rates (Fig 2b). However, decision optimality is NOT positively correlated with lose-shift rate across conditions. 2) DBM model suggests that, across different levels of environmental volatility, there is an inverted U shape relationship between the frequency of using lose-shift and one?s belief of stationarity (Fig 3b), and a close-to-linear relationship between decision optimality and the belief of stationarity in a low volatility environment (Fig 6b). Implications for model selection. We have shown that both models have comparable prediction accuracy for individuals with good performance, but R-W model is better in explaining poor learners? choice. There are several possible reasons: 1) the Bayesian model assumed subjects would use the feedback information to

update the posterior probability of target reward distribution. Thus for ?poor? learners who did not use the feedback information, this assumption is no longer appropriate. 2) the R-W model assumed subjects would only update the chosen option?s value, thus error trials may have less influence (especially in the early stages, with low learning rate). That is, for 0 value option, it will remain 0 if not rewarded, and for the highest value option, it will remain being the highest value option even if not rewarded. Therefore, R-W model may capture poor learners? search pattern better with a low learning rate. Future directions. For future work, we will modify current R-W model with a dynamic learning rate that will change based on value estimation, and modify current DBM model with a parameter that controls how much feedback information is used in updating posterior belief and a hyper-parameter that models the dynamic of ?. Acknowledgements We thank Angela Yu for sharing the data in Yu et al. 2014, and for allowing us to use it in this paper. 8

2 References

[1] Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. Cognitive Systems Research, 12(2), 164-174. [2] Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. Nature neuroscience, 10(9), 1214-1221. [3] Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci, 8. [4] Yu, A. J., & Huang, H. (2014). Maximizing masquerading as matching in human visual search choice behavior. Decision, 1(4), 275. [5] 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. [6] Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. MIT press. [7] Browning, M., Behrens, T. E., Jocham, G., O?Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. Nature neuroscience, 18(4), 590-596. [8] Yu, A. J., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior?. In Advances in neural information processing systems (pp. 1873-1880). [9] Gershman, S. J. (2015). A Unifying Probabilistic View of Associative Learning. PLoS Comput Biol, 11(11), e1004567. [10] Blakely, E., Starin, S., & Poling, A. (1988). Human performance under sequences of fixed-ratio schedules: Effects of ratio size and magnitude of reinforcement. The Psychological Record, 38(1), 111. [11] Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. Heuristics and biases: The psychology of intuitive judgment, 49. [12] Worthy, D. A., & Maddox, W. T. (2014). A comparison model of reinforcement-learning and win-stay-loseshift decision-making processes: A tribute to WK Estes. Journal of mathematical psychology, 59, 41-49. [13] Bonawitz, E., Denison, S., Gopnik, A., & Griffiths,

T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. Cognitive psychology, 74, 35-65.
9