Empirical Test

Zijian Zark Wang

2023-04-16

1 Data

To test the capacity of the proposed model in explaining experimental findings under the "Money Earlier or Later" (MEL) paradigm, I select two open datasets. The first dataset is from Ericson et al. (2015), containing 23,131 observations from 939 participants; the second is from Chávez et al. (2017), containing 34,515 choices from 1,284 participants. Hereafter, I term the first dataset as *Ericson* data, and the second dataset as *Chávez* data. Each dataset has been used in more than one previous study.[1] Readers interested in the empirical method and the results of this paper can easily compare them with those of other papers. For both datasets, the corresponding experiments require the participants to answer a series of choice questions. For each question, participants need to make a choice between an small sooner reward (denoted by SS) and a large later reward (denoted by LL). I denote the reward magnitude and delay by x_s and t_s for option SS, and by x_l and t_l for option LL, where $x_l > x_s > 0$, $t_l > t_s$.

Notably, there are significant differences in experimental procedure underlying the datasets. In *Ericson* data, the participants are recruited via Mechanical Turk. Rewards are framed in US dollars, ranging from USD 0.03 to USD 10100. For 25.4% of the observations, option SS has a reward less than USD 3. Delays are framed in weeks. Delays for SS range from

[1] For example, *Ericson* data is used by Wulff and van den Bos (2018) for discussing the proper ways to compare different choice models. *Chávez* data is used by Gershman and Bhui (2020) for testing their proposed attention-based choice model.

0 to 2 weeks, and the maximum delay for LL is 5 weeks. In *Chávez* data, the participants are Mexican high school and first-year university students, and attending the experiment is a class requirement. Rewards are framed in Mexican pesos, ranging from MXD 11 (approx. USD 0.6) to MXD 85 (approx. USD 4.7). Delays are framed in days. All sooner rewards are delivered "today" ($t_s = 0$), and the maximum delay for LL is 186 days. Given that the two datasets employ different frames and ranges for rewards and delays, a same model fitted on these datasets may have different parameters and goodness of fit.

I mainly focus on out-of-sample model performance. For each dataset, I randomly draw the responses from 20% of the participants as the test sample, and set the rest as the train sample. To mitigate the overfitting issue, I implement a 10-fold cross-validation procedure on the train sample.

2 Empirical Strategy

I test three types of intertemporal choice model: discounted utility model, trade-off model, and heuristic model.

The discounted utility model assumes that the decision maker tends to choose the option with greater discounted utility. Let the discounted utility for option j ($j \in \{l, s\}$) be $v_j = d(t_j)u(x_j)$, where $d(.)$ is the discounting function and $u(.)$ is the instantanoues utility function. Suppose the decision maker's perceived discounted utility for each option, denoted by is $\tilde{v}_l$ and $\tilde{v}_s$, is noisy. I set $\tilde{v}_l = v_l + \eta_l$, $\tilde{v}_s = v_s + \eta_s$. When η_l and η_s are independent noises and both follow $Gumble(0, \rho)$, where the scale parameter $\rho \in (0, \infty)$, the probability that the decision maker chooses LL is

$$P\{\tilde{v}_s \leq \tilde{v}_l\} = \frac{1}{1 + \exp\{-\frac{1}{\rho}(v_l - v_s)\}}$$

The trade-off model (Scholten and Read, 2010; Scholten et al., 2014) assumes that when thinking of whether to choose LL, the decision maker tends to make a comparison between attributes (reward and time), rather than between options (LL and SS). If the benefit of receiving a larger reward exceeds the cost of waiting longer, she will choose LL; otherwise, she will choose SS. Let B denote the benefit of receiving a larger reward, Q denote the cost of waiting longer. The value of B can be simply represented by $u(x_l) - u(x_s)$. Following Scholten et al. (2014), I represent Q by

$$Q = \frac{\kappa}{\zeta_1} \ln\left(1 + \zeta_1 \left(\frac{w(t_l) - w(t_s)}{\zeta_2}\right)^{\zeta_2}\right)$$

where $w(t) = \ln(1 + \omega t)/\omega$. The parameter ω measures the magnitude in which time is distorted in the decision maker's mind; κ measures the relative importance of reducing waiting time compared with increasing reward magnitude; ζ_1, ζ_2 jointly determine the curvature of changes in Q relative to $t_l - t_s$.[2] I assume the decision maker's perception of B and Q, denoted by $\tilde{B}$ and $\tilde{Q}$, is noisy. $\tilde{B} = B + \eta_B$, $\tilde{Q} = Q + \eta_Q$, where η_B and η_Q are noise terms. Again, assume η_B and η_Q are independent and follow $Gumble(0, \rho)$, then the probability that the decision maker chooses LL is

$$P\{\tilde{Q} \leq \tilde{B}\} = \frac{1}{1 + \exp\{-\frac{1}{\rho}(B - Q)\}}$$

For the heuristic model, I employ a decision tree algorithm called XGBoost (Chen and Guestrin, 2016). The intuition underlying XGBoost is that, the decision-maker uses a chain of if-then rules to make a choice, and repeats this process for several times, adding up the results of each iteration to make the final decision. This algorithm has been widely used in solving classfication problems (including predicting human risky choices, see Plonsky et al. 2019). To better fit the data, I tune the hyper-parameters of this algorithm via grid search. For *Ericson* data, the features I use are x_s, x_l, t_s, t_l, the absolute and relative differences between x_s and x_l, the absolute and relative differences between t_l and t_s, the interest rate of LL when SS is invested as principal. For *Chávez* data, given that $t_s = 0$, I omit t_s and the differences between t_s and t_l from the features.[3]

[2] Scholten et al. (2014) use these two parameters $\{\zeta_1, \zeta_2\}$ to ensure that Q follow a S-shape curve in relation to $t_l - t_s$, and that the decision-maker's behavioral pattern can shift between sub-additivity and super-additivity.

[3] The features are extracted following the methods in Read et al. (2013) and Ericson et al. (2015).

The attention-adjusted discounting factor is dependent on the decision maker's initial attention allocation. I test the model under the assumptions that initial attention allocation is exponential and uniform (I term the former as "*attention*", the latter as "*attention_uni*"). Along with the attention-adjusted discounting, I employ 8 other methods to draw discounting factors, which are:

1. **exponential**, denoted by "*expo*"

$$d(t) = \delta^t$$

where the parameter is δ and $\delta \in (0, 1]$.

2. **double exponential**, denoted by "*expo2*" (van den Bos and McClure, 2013)

$$d(t) = \omega \delta_1^t + (1 - \omega)\delta_2^t$$

where the parameters are $\{\delta_1, \delta_2, \omega\}$, and $\delta_1, \delta_2 \in (0, 1]$.

3. **hyperbolic**, denoted by "*hb*"

$$d(t) = \frac{1}{1 + kt}$$

where the parameter is k.

4. **dual-parameter hyperbolic**, denoted by "*hb2*" (Loewenstein and Prelec, 1992)

$$d(t) = \frac{1}{(1 + kt)^a}$$

where the parameters are $\{k, a\}$.

5. **magnitude-dependent hyperbolic**, denoted by "*hbmd*" (Gershman and Bhui, 2020)

$$d(t) = \frac{1}{1+kt}, \quad k = \frac{1}{bu(x_t)}$$

where the parameter is b.

6. **quasi-hyperbolic**, denoted by "*quasihb*"(Laibson, 1997)

$$d(t) = \mathbf{1}\{t = 0\} + \beta\delta^t \cdot \mathbf{1}\{t > 0\}$$

where the parameters are $\{\beta, \delta\}$, and $\beta, \delta \in (0, 1]$.

7. **quasi-hyperbolic plus fixed delay cost**, denoted by "*quasihb_fc*" (Benhabib et al., 2010)

$$d(t) = \mathbf{1}\{t = 0\} + (\beta\delta^t - \frac{c}{u(x_t)}) \cdot \mathbf{1}\{t > 0\}$$

where the parameters are $\{\beta, \delta, c\}$, and $\beta, \delta \in (0, 1]$.

8. **homogeneous costly empathy**, denoted by "*hce*" (Noor and Takeoka, 2022)

$$d_t = \kappa_t u(x_t)^{\frac{1}{m}}$$

where κ_t is decreasing in t. I set $\kappa_t = \delta^t$, where the parameters are $\{m, \delta\}$ and $\delta \in (0, 1]$.

Besides, I employ 2 types of utility functions to obtain $u(.)$: exponential or CARA utility, where $u(x) = 1 - e^{-\gamma x}$; power utility, where $u(x) = x^\gamma$. In each utility function, γ is the parameter, $\gamma \in (0, \infty)$. For parameters in discounting function, except for those explicitly marked as having a domain between 0 and 1, the domain of all other parameters is $(0, \infty)$.

In model fitting, if a parameter has a lower bound of 0, I set its lower bound to 0.001; if a parameter has a upper bound of infinity, I set its upper bound to 100.

I use the maximum likelihood method to estimate the parameters, and apply L-BFGS-B method for optimization. As the solutions of L-BFGS-B are sensitive to initial points and often converge to local optima, I use the basin-hopping algorithm to achieve global optimization. [4] Finally, I compare the goodness of fit and out-of-sample performance of 20 discounted utility models, 2 trade-off models, and 1 heuristic model on the two datasets.

3 Result

3.1 Results for ***Ericson*** data

Table 1 shows the goodness of fit for each model in cross-validation. The heuristic model has the highest accuracy rate, the lowest log loss, and the lowest MAE. The trade-off (*trade*) model with power utility performs the lowest MSE. The magnitude-dependent hyperbolic (*hbmd*) model with power utility ranks the second or third in all evaluation metrics. On the test sample, these three models also perform the best in MSE, MAE, log loss and accuracy rate (see Table 2).

To test the correlation between the models, I randomly draw 1,000 choice questions from *Ericson* data, and set the choices predicted by the heuristic model as labels, letting the other models to predict them. The accuracy rate for *trade* model with power utility is the highest, which is 95.7%, and the second is *hbmd* with power utility (95.0%).

Notably, the heuristic model has the most parameters that need to be fitted. Apart from the heuristic model, the trade-off model has the highest number of parameters, which is 6. The *hbmd* model has only 3 parameters to be fitted.

[4]The basin-hopping algorithm runs a local optimizer for several times. After each iteration, the solution randomly drifts to a new point. This new point is taken as the initial point for the next iteration. The algorithm compares the solution of the next iteration with the original solution, and is more likely to accept the better solution between them (note there is still some probability of accepting an inferior solution). The magnitude of drifting is dependent on a stepwise parameter, which I set as 0.5; the probability of accepting the inferior solution is dependent on a temper parameter, which I set as 1.0. I also set the maximum number of iterations as 500.

Table 1: Cross-Validation Results on Ericson Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
heuristic	–	0.298	0.298	0.581	0.702
trade	power	0.204	0.407	0.595	0.693
hbmd	power	0.206	0.413	0.602	0.692
quasihb_fc	power	0.208	0.418	0.606	0.685
quasihb	power	0.209	0.417	0.607	0.687
expo2	power	0.210	0.420	0.609	0.685
attention_uni	power	0.211	0.422	0.611	0.679
hb2	power	0.211	0.422	0.611	0.682
hb	power	0.211	0.422	0.612	0.681
expo	power	0.211	0.422	0.612	0.681
hce	power	0.211	0.422	0.612	0.681
attention	power	0.215	0.431	0.621	0.673
trade	cara	0.218	0.437	0.628	0.666
attention	cara	0.228	0.456	0.648	0.638
attention_uni	cara	0.229	0.458	0.650	0.631
hbmd	cara	0.229	0.458	0.650	0.631
quasihb	cara	0.229	0.459	0.651	0.630
quasihb_fc	cara	0.229	0.459	0.651	0.630
expo2	cara	0.229	0.458	0.651	0.630
expo	cara	0.229	0.459	0.651	0.629
hce	cara	0.229	0.459	0.651	0.629
hb2	cara	0.229	0.459	0.651	0.629
hb	cara	0.229	0.459	0.651	0.629

Note: "mae" denotes mean absolute error, "mse" denotes mean squared error.

Table 2: Out-of-Sample Test Results on *Ericson* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
heurstic	–	0.200	0.400	0.586	0.702	0.290
hbmd	power	0.204	0.411	0.596	0.696	0.237
trade	power	0.202	0.405	0.591	0.695	0.249
quasihb_fc	power	0.207	0.412	0.603	0.694	0.248
quasihb	power	0.207	0.422	0.604	0.694	0.248
expo	power	0.210	0.427	0.609	0.690	0.306
hce	power	0.210	0.428	0.609	0.686	0.260
hb	power	0.209	0.424	0.607	0.686	0.260
attention_uni	power	0.211	0.422	0.611	0.678	0.157
attention	power	0.215	0.431	0.623	0.673	0.142
trade	cara	0.217	0.434	0.626	0.668	0.120
attention	cara	0.231	0.457	0.654	0.629	0.091
attention_uni	cara	0.231	0.459	0.654	0.624	0.087
hbmd	cara	0.231	0.459	0.654	0.623	0.085
hb2	cara	0.233	0.454	0.658	0.620	0.046
hb	cara	0.231	0.460	0.655	0.618	0.082
hce	cara	0.231	0.460	0.655	0.618	0.080
quasihb	cara	0.231	0.460	0.655	0.618	0.078
quasihb_fc	cara	0.231	0.460	0.655	0.618	0.078
expo	cara	0.231	0.460	0.655	0.618	0.080
expo2	cara	0.240	0.447	0.679	0.616	0.013
expo2	power	0.385	0.386	4.767	0.614	0.000
hb2	power	0.386	0.386	6.562	0.614	0.000

Note: "mae" denotes mean absolute error, "mse" denotes mean squared error, "pred_ll" denotes the ratio of LL in predicted choices.

I now focus on attention-adjusted models. First, it is notable that *hbmd* can be viewed as a special case of the attention-adjusted models. [5] Magnitude-dependent hyperbolic discounting plus power utility is identical to attention-adjusted discounting under uniform initial attention allocation, plus logarithmic utility. Second, to explain why attention-adjusted models with exponential or power utility underperform some other models, see the last column in Table 2, which reports the ratio of LL choices in the predictions. The two attention-adjusted models underestimate the decision makers' tendency to choose option LL in *Ericson* data. Third, for why using logarithmic utility in attention-adjusted models outperforms power utility in the given data, the experimental design may explains.

25.3% of the small early rewards is less than $3, 19.3% of the large late rewards is less than $3.

In Chavez data, the smallest x_s is 11, the smallest x_l is 25.

1. what model performs the best
2. redundancy analysis
3. attention

    - attention with log = hbmd with power
    - not fit very small reward

3.2 Results for *Chávez* data

Table 3shows the fitness of each model in cross-validation.

Table 3: Cross-Validation Results on *Chávez* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
heuristic	–	0.2211	0.2211	0.4818	0.7789
tradeoff	power	0.1571	0.3140	0.4840	0.7826
expo2	power	0.1574	0.3146	0.4844	0.7826

[5] Note that under the assumption that the initial attention is uniformly allocated, the discounting factor in attention-adjusted model is $1/(1+kt)$, where $k = e^{-u(x_t)/\lambda}$. Suppose $u(x)/\lambda = \ln\beta + \gamma \ln x$, then we can get the magnitude-dependent hyperbolic model with power utility.

Table 3: Cross-Validation Results on *Chávez* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
quasihb	power	0.1574	0.3144	0.4845	0.7826
quasihb_fc	power	0.1576	0.3146	0.4850	0.7816
hb2	power	0.1577	0.3149	0.4854	0.7804
attention	power	0.1579	0.3170	0.4860	0.7826
hbmd	power	0.1583	0.3174	0.4868	0.7731
attention_uni	power	0.1584	0.3178	0.4881	0.7826
hbmd	cara	0.1592	0.3191	0.4900	0.7731
hb	power	0.1597	0.3209	0.4903	0.7731
attention_uni	cara	0.1601	0.3203	0.4905	0.7641
expo	power	0.1604	0.3229	0.4922	0.7731
hce	power	0.1604	0.3229	0.4922	0.7731
attention	cara	0.1625	0.3263	0.4956	0.7465
tradeoff	cara	0.1633	0.3268	0.4971	0.7430
hb2	cara	0.1659	0.3326	0.5042	0.7439
hb	cara	0.1685	0.3372	0.5093	0.7287
quasihb	cara	0.1679	0.3411	0.5099	0.7450
expo	cara	0.1692	0.3421	0.5107	0.7253
hce	cara	0.1700	0.3412	0.5124	0.7168
quasihb_fc	cara	0.1701	0.3407	0.5128	0.7203
expo2	cara	0.1747	0.3487	0.5213	0.6825

Table 4 shows the out-of-sample performance of each model.

Table 4: Out-of-Sample Test Results on *Chávez* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
attention	power	0.1628	0.3230	0.4982	0.7702	0.3299
attention_uni	power	0.1635	0.3233	0.5008	0.7702	0.3299

Table 4: Out-of-Sample Test Results on *Chávez* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
tradeoff	power	0.1633	0.3206	0.4986	0.7674	0.2926
heuristic	–	0.2326	0.2326	8.3825	0.7674	0.2926
hb	power	0.1641	0.3315	0.5014	0.7603	0.3296
quasihb_fc	power	0.1876	0.4173	0.5640	0.7603	0.3296
expo	power	0.1646	0.3302	0.5032	0.7603	0.3296
hbmd	cara	0.1645	0.3283	0.5033	0.7603	0.3296
hce	power	0.1651	0.3359	0.5039	0.7603	0.3296
quasihb	power	0.1779	0.3969	0.5413	0.7603	0.3296
hbmd	power	0.1630	0.3252	0.4987	0.7603	0.3296
hb2	power	0.1650	0.3217	0.5024	0.7570	0.2184
attention_uni	cara	0.1821	0.3844	0.5425	0.6710	0.1118
tradeoff	cara	0.2500	0.5000	0.6931	0.6678	0.4780
attention	cara	0.1802	0.3552	0.5360	0.6483	0.0000
quasihb	cara	0.2060	0.3433	0.5958	0.6483	0.0000
expo	cara	0.1996	0.3290	0.5981	0.6483	0.0000
hce	cara	0.2085	0.3226	0.6464	0.6483	0.0000
hb	cara	0.2198	0.3268	0.6642	0.6483	0.0000
expo2	cara	0.2245	0.3201	0.7294	0.6483	0.0000
hb2	cara	0.2504	0.3283	0.7914	0.6483	0.0000
expo2	power	0.3517	0.3517	12.6783	0.6483	0.0000
quasihb_fc	cara	0.2019	0.3423	0.5866	0.6483	0.0000

3.3 Parametrication

Reference

Benhabib, J., Bisin, A., and Schotter, A. (2010). Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior*, 69(2):205–223.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA. ACM.

Chávez, M. E., Villalobos, E., Baroja, J. L., and Bouzas, A. (2017). Hierarchical Bayesian modeling of intertemporal choice. *Judgment and Decision Making*, 12(1):19–28. Publisher: Cambridge University Press.

Ericson, K. M., White, J. M., Laibson, D., and Cohen, J. D. (2015). Money Earlier or Later? Simple Heuristics Explain Intertemporal Choices Better Than Delay Discounting Does. *Psychological Science*, 26(6):826–833.

Gershman, S. J. and Bhui, R. (2020). Rationally inattentive intertemporal choice. *Nature Communications*, 11(1):3365. Number: 1 Publisher: Nature Publishing Group.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, 112(2):443–478.

Loewenstein, G. and Prelec, D. (1992). Anomalies in Intertemporal Choice: Evidence and an Interpretation. *The Quarterly Journal of Economics*, 107(2):573–597.

Noor, J. and Takeoka, N. (2022). Optimal Discounting. *Econometrica*, 90(2):585–623.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., Cavanagh, J. F., and Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. Publisher: arXiv Version Number: 1.

Read, D., Frederick, S., and Scholten, M. (2013). DRIFT: An analysis of outcome framing in intertemporal choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):573–588.

Scholten, M. and Read, D. (2010). The Psychology of Intertemporal Tradeoffs. *Psychological Review*, 117(3):925–944.

Scholten, M., Read, D., and Sanborn, A. (2014). Weighing Outcomes *by* Time or *Against* Time? Evaluation Rules in Intertemporal Choice. *Cognitive Science*, 38(3):399–438.

van den Bos, W. and McClure, S. M. (2013). Towards a General Model of Temporal Discounting: General Model of Time Discounting. *Journal of the Experimental Analysis of Behavior*, 99(1):58–73.

Wulff, D. U. and van den Bos, W. (2018). Modeling Choices in Delay Discounting. *Psychological Science*, 29(11):1890–1894.