多腕バンディット問題における 認知的満足化モデルのパラメータリカバリとモデル比較
Parameter and model comparison of cognitive satisficing model in multi-armed bandit problem

横須賀 天臣 †, 高橋 達二 †
Takaomi Yokosuka, Tatsuji Takahashi
† 東京電機大学
Tokyo Denki University
tatsujit@mail.dendai.ac.jp

概要

認知的満足化モデルは目的に応じた希求水準を満たすかにより，探索と活用のバランスを調整し，多腕バンディット問題において有効に働く．本研究では，4個の選択肢からなるベルヌーイ・バンディット問題における認知的満足化と Softmax による選択行動について，最尤推定によるパラメータのリカバリ性能を確認した．また，行動実験によるモデルの比較を行った．その結果，全モデルでパラメータのリカバリが確認され，データに適合する際の性質が明らかになった．

キーワード：意思決定, 強化学習, バンディット問題

1. 序論

多腕バンディット問題とは，確率的に報酬を排出する複数の選択肢の中から，エージェントが逐次的に意思決定をしていき，可能な限り多くの報酬を獲得する，という強化学習の基本的な問題の 1 つである．心理学や認知神経科学では，構造化された環境における選択行動の探索と活用を評価するために，しばしば実験課題として用いられている (Daw et al., 2006) ．情報を収集のために未知の行動を選択をする探索と，収集した情報に基づいて既知最良の行動を選択する活用があり，これらはトレードオフである．エージェントは探索と活用のバランスを調整することで，累積獲得報酬の最大化を試みる．多腕バンディット課題で得られたデータは，強化学習やベイズ推論などの計算モデルを用いてパラメータ推定が行われる．

計算モデルを用いて実験課題の評価をする上で，パラメータリカバリとモデルリカバリは，研究の信頼性と解釈性を確保するうえで重要な手順とされる (Wilson & Collins, 2019) ．パラメータリカバリとは，既知のパラメータから生成したデータに対して，パラメータ推定を用いることで，生成時のパラメータを復元できるか確認することである．モデルリカバリとは，既知のモデルから生成したデータに対して，モデル比較指標を用いることで，生成時のモデルを正しく識別できるかを確認することである．計算モデルを用いた実験をデザインする際は，事前に数値実験を実施し，モデルリカバリとパラメータリカバリを確認することが推奨されている．

本研究では，認知的満足化と呼ばれる逐次的意思決定モデルに焦点を当てて，そのパラメータとモデルのリカバリを確認する．Simon (1957) は，現実の多くの場面において，多数の選択肢の中から最適な選択肢を選ぶことは困難であることを指摘し，生物は目標に応じたある水準（希求水準）を満たすか否かにより選択を行うという満足化の原理を示唆した．この満足化のアイデアを取り入れたモデルが認知的満足化である (高橋他, 2016; Tamatsukuri & Takahashi, 2019) ．認知的満足化モデルは，各選択肢の報酬の期待値と目的に応じた水準との大小関係により，探索するか活用するかが決まる．これまでに，多腕バンディット問題での有効性が証明されているが，複数のパラメータや他のモデルとのリカバリ率の確認は行われていない．また，認知的満足化モデルから，Softmax の逆温度パラメータ β を直感化し，目標設定の意味で設定しやすくした形式の Softsatisficing モデル (Kamiya & Takahashi, 2022) がある．このモデルを用いることで，通常の Softmax 行動選択を用いたモデルと同様に，実験データとのフィッティングを行うことができる．特に，参加者の行動と獲得した報酬の系列から，参加者が満たそうとした希求水準を推定することができる．

本研究では，多腕バンディット問題における認知的満足化モデルのシミュレーションを行い，モデルリカバリとパラメータリカバリを網羅的に確認する．また，行動実験データにモデルを適合することで，モデルの比較を検討した．

2. 多腕バンディット問題

本研究で扱う多腕バンディット問題について説明する．多腕バンディット問題では，K 個の行動 $\{a_1, a_2, ..., a_K\}$ について，確率 $\{p_1, p_2, ..., p_K\}$ のベルヌーイ分布に従い報酬として 1 が排出される．エージェントは累積獲得報酬を可能な限り大きくすることを目的とする．報酬確率はエージェントから未知であり，各試行 t におけるそれまでに得られた報酬の平均として選択行動 a_i の価値 E_i が評価される．

$$E_i \leftarrow E_i + \frac{1}{n_i}(r_t - E_i) \tag{1}$$

r_t は試行 t にエージェントが得た報酬を表し，逐次的に更新される．例えば，同一の a_i を選択し続ければ，E_i は真の確率へと近づくが，確率の低い腕を選択し続けてしまうと累積獲得報酬は減少してしまう．累積獲得報酬の最大化のためには，素早くかつ正確に報酬確率が最大の腕を見つけ出すことが重要となるが，そのためには探索と活用のバランスが重要となる．

3. 行動の計算モデル

エージェントは行動価値 E_i と計算モデルに算出される行動確率により，逐次的な意思決定を行う．本研究では，Softmax (SM)，Risk-sensitive Satisficing (RS) と Softsatisficing (SS) という 3 個の計算モデルを用いる．これらの計算モデルは，試行回数を考慮したり，確率的に振る舞ったりすることで探索と活用のバランスを調整する．また，モデル固有のパラメータがあり，その大きさにより行動の探索と活用の傾向が変化する．

SM は価値の大きさに応じて確率的に行動を選択するモデルである．Q 学習をはじめとした多くのモデルに用いられている基本的なモデルの 1 つである．

$$\mathrm{Softmax}(a_i) = \frac{e^{E_i\beta}}{\sum_{j=1}^{k} e^{E_j\beta}} \tag{2}$$

逆温度 β が大きいほど, 活用的な選択をしやすくなる．

RS は満足化に基づいて行動を選択するモデルである．次のような価値関数が最大となる行動 a_i を選択する．

$$\delta_i = E_i - \aleph_{\mathrm{RS}} \tag{3}$$

$$\mathrm{RS}_i = n_i \delta_i \tag{4}$$

$$a_i \leftarrow \arg\max_i \mathrm{RS_i} \tag{5}$$

n_i は a_i の選択回数, $\aleph_{\mathrm{RS}}$ は希求水準を表す．$E_i < \aleph_{\mathrm{RS}}$ の場合を非満足状態と呼び，エージェントは探索的になり，n_i の少ない a_i を選びやすい．$E_i > \aleph_{\mathrm{RS}}$ の場合を満足状態と呼び，エージェントは活用的になり，n_i の多い a_i を選びやすい．また，$\aleph_{\mathrm{RS}}$ が以下の式で定められる最適切基準である場合，累積報酬が最大となる行動をする．

$$\aleph_{\mathrm{opt}} = \frac{p_{1\mathrm{st}} + p_{2\mathrm{nd}}}{2} \tag{6}$$

報酬確率が最大のものを $p_{1\mathrm{st}}$，二番目に大きいものを $p_{2\mathrm{nd}}$ とする．

SS は満足化に基づいて確率的に行動を選択するモデルである．

$$\mathrm{Softsatisficing}(a_i) = \frac{e^{E_{\mathrm{RS_i}}}}{\sum_{j=1}^{k} e^{E_{\mathrm{RS_i}}}} \tag{7}$$

$$E_{\mathrm{RS_i}} = -\ln(\aleph_{\mathrm{SS}} - E_i) \tag{8}$$

希求水準 $\aleph_{\mathrm{SS}}$ と E_i の差に応じて，探索行動の調整を行う．また，$\aleph_{\mathrm{SS}} < E_i$ となる a_i が存在する場合は，それを選択する．RS モデルと同様に，$\aleph_{\mathrm{SS}}$ は最適切基準が存在する．

4. 数値実験

数値実験を実施し，3 つの計算モデルのパラメータとモデルのリカバリを確認した．本研究では，4 個の選択行動からなるベルヌーイ・バンディット問題を用いた．報酬確率 P_k は $\{.20, .20, .40, .80\}$，試行回数 T は 150 とした．データ生成を行った後，生成したデータに対してパラメータ推定を行った．

本研究で用いた計算モデルはパラメータを 1 つまで持つ．SM モデルの逆温度 β は $[0, 20]$ の範囲で 1 ずつ，RS モデルの希求水準 $\aleph_{\mathrm{RS}}$ は $[0, 1]$ の範囲で 0.05 ずつ，SS モデルのパラメータ $\aleph_{\mathrm{SS}}$ は $[0, 1]$ の範囲で 0.05 ずつ，設定した．各モデルは上記 21 水準について，データ生成およびパラメータ推定を実行した．

4.1 データ生成手順

シミュレーションにより，選択行動系列 a_t と獲得報酬系列 r_t を生成した．シミュレーションは各モデル・各パラメータごとに 100 回ずつ，合計 6300 回実行した．SM モデルの逆温度 β は $[0, 20]$ の範囲で 1 ずつ，RS モデルの希求水準 $\aleph_{\mathrm{RS}}$ は $[0, 1]$ の範囲で 0.05 ずつ設定し，SS モデルのパラメータ $\aleph_{\mathrm{SS}}$ は $[0, 2]$ の範囲で 2 ずつ設定した．

4.2 パラメータ推定手順

選択行動系列 a_t に対して，負対数尤度 (negative log-likelihood, NLL) を算出した．

$$\mathrm{NLL} = -\sum_{t=1}^{T} \log P(a_t) \tag{9}$$

各モデルのパラメータ 21 水準ごとに負対数尤度を算出し，負対数尤度が最小となる値を推定値として求めた．

4.3 結果

SM モデルにおける負対数尤度を図 1，RS モデルおける負対数尤度を図 2，SS モデルにおける負対数尤度を図 3 に示す．すべてのモデルで，パラメータが真値と等しいときに対数尤度が最小となった．

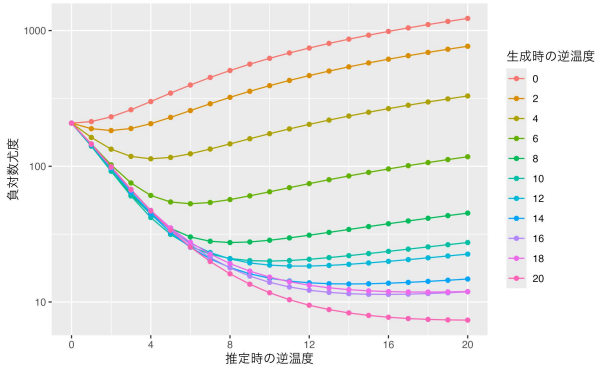

図 1 SM モデルによる生成データと同モデルでの負対数尤度の平均

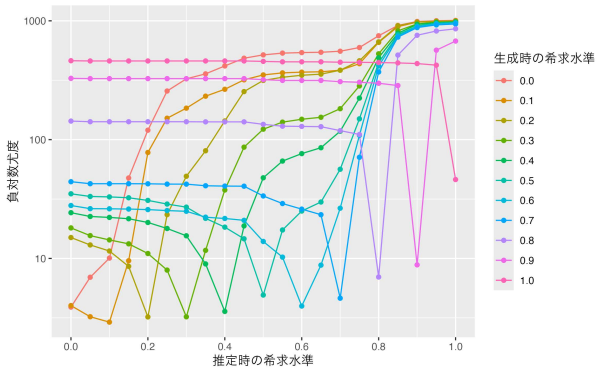

図 2 RS モデルによる生成データと同モデルでの対数尤度

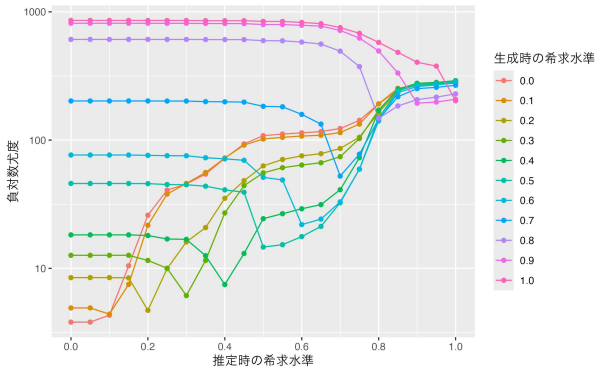

図 3 SS モデルによる生成データと同モデルでの対数尤度

データ生成時のパラメータ (真値) と推定されたパラメータ (推定値) の相関分析の結果を表 1 に示す．各モデルにおいて同一モデルが最大の値となった．

表 1 各モデルにおける真値と推定値の相関係数

	推定値		
真値	SM	RS	SS
SM	.886	-.495	-.777
RS	-.805	.851	.826
SS	-.831	.596	.864

データに対して推定値での負対数尤度が最小となったモデルを適合モデルと呼ぶ．各モデルの生成データに対して適合したモデルの割合を表 2 に示す．これは混合行列とも呼ばれ，単位行列に近くほどモデルがリカバリされていることを示す．各モデルにおいて同一モデルが最大となり，モデルのリカバリが確認された．

表 2 モデルリカバリの混合行列

	適合モデル		
生成モデル	SM	RS	SS
SM	.916	.053	.031
RS	.271	.704	.025
SS	.350	.099	.551

4.4 考察

SM モデルは，全データの真値と高い相関を示す．RS モデルの希求水準 $\aleph_{RS}$ の真値が大きくなるほど，SM モデルの逆温度 β の推定値は小さくなり，それぞれのモデルにおいて活用的な行動になる傾向として一致している．SS モデルの希求水準 $\aleph_{SS}$ の真値とも同様の結果を示しており，より大きい相関となっている．SS モデルも，全データの真値と高い相関を示している．RS モデルの $\aleph_{RS}$ の真値に対して，SS モデルの $\aleph_{SS}$ の推定値は，SM モデルよりも大きい相関がある．これは SS モデルが，SM モデルをベースに RS モデルの結果を再現するように定式化されているためであり，妥当な結果といえる．RS モデルの $\aleph_{RS}$ の推定値は他のモデルの真値とほとんど相関が無いが， 図 2 より，リカバリされたときに負対数尤度がとても小さくなることが示された．これは RS が価値関数によって求められた値を確率 1 で選ぶ決定論的モデルであるためこのような結果になったと考えられる．

5. 行動実験

本研究では，人間を対象とした 4 本腕ベルヌーイ・バンディット課題の行動実験データによって 3 つのモデルを検討する．実験は Web で実施し，クラウドソーシングサービス CrowdWorks (https://crowdworks.jp/) を用いて実験参加者を募集した．参加者は PC を操作してタスクを完遂した．本研究では，参加者 100 名の

データを分析に用いる．

5.1 実験手順

参加者には，選択行動 a_i として 4 個のデッキを提示した．各デッキには 10 枚の同色のカードが積まれており，カードの裏面は単色であり，表面は「あたり $(r = 1)$」か「はずれ $(r = 1)$」の 2 種類である．デッキによって，当たりの枚数（報酬確率 p_k）は異なっている．各試行 t において，参加者はデッキを 1 つ選択し，トップのカードが裏返され，その表面が結果として表示された．その後，カードは元のデッキに戻され，デッキの並び順序と各デッキ内のカードの順序はランダムにシャッフルされた．このような一連の手順を既定の試行回数 T に到達するまで，毎試行繰り返した．シミュレーションと同様に，各デッキの報酬確率 p_k を $\{.20, .20, .40, .80\}$，試行回数 T を 150 とした．

5.2 結果

数値実験と同様の手順でパラメータ推定を行った．実験データに対する適合モデルの割合と適合時の推定と負対数尤度の平均を表 3 に示す．

表 3 行動実験データへのモデルの適合結果

モデル	SM	RS	SS
適合したデータの割合	.670	.120	.210
適合時の平均推定値	6.89	.288	.546
適合時の平均負対数尤度	91.8	20.6	42.3

5.3 考察

表 3 より，人間の行動データに適合した割合が最大のモデルは SM モデルであった．SM モデルは過半数以上の参加者のデータに適合したが，適合時の平均負対数尤度は最大である．それに対して，RS モデルは適合したデータの割合が最小であるが，適合時の平均負対数尤度も最小であり，少数ながらも一部の参加者の行動をよりよく説明していると考えられる．また，SS モデルは適合したデータの割合と適合時の平均負対数尤度がともに次善の結果となっており，$\aleph_{\mathrm{SS}}$ の平均推定値は最適切基準に近い値を示す．このことから，人間を対象とした多腕バンディット課題において，希求水準を満たすか否かにより選択を行う参加者と，そうでない参加者がいると考えられる．

6. 結論

本研究では，Softmax, Risk-sensitive Satisficing, Soft-satisficing という 3 つの計算モデルについて，4 本腕ベルヌーイ・バンディット問題におけるパラメータとモデルのリカバリを確認した．その結果，各モデルは最尤推定により真値と等しい推定値を求め，データ生成時と同一のモデルが最も適合することを示した．また，行動実験より，認知的満足化モデルを用いることで説明できる参加者のデータが 33% あったことが分かった．

文 献

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441(7095), 876-879.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. Elife, 8, e49547.

Simon, H. (1957). A behavioral model of rational choice. Models of man, social and rational: Mathematical essays on rational human behavior in a social setting, 6(1), 241-260.

高橋達二, 甲野佑, & 浦上大輔. (2016). 認知的満足化 限定合理性の強化学習における効用. 人工知能学会論文誌, 31(6), AI30-M_1.

Tamatsukuri, A., & Takahashi, T. (2019). Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. Biosystems, 180, 46-53.

Kamiya, T., & Takahashi, T. (2022). Softsatisficing: Risk-sensitive softmax action selection. Biosystems, 213, 104633.