MI プログラム 卒研中間発表会

評価者特性の時間変動を考慮した項目反応モデル

1810519　林真由
指導教員 宇都 雅輝 准教授

1　はじめに

近年，大学入試や資格試験，教育評価などにおいて，パフォーマンス評価のニーズが高まっている．一方で，パフォーマンス評価では，多肢選択式試験のような客観的評価と異なり人間の評価者が採点を行うため，評価者の厳しさや一貫性などの特性差により，採点に偏りが生じ，受検者の能力測定の信頼性が低下する問題が知られている．このような問題を解決する数理的なアプローチの一つとして，評価者の特性を考慮した項目反応理論（Item response theory:IRT）[1] が近年注目されている．

それらのモデルでは，評価者の特性差の影響を考慮した能力推定を行うことができる．一方で，既存モデルのほとんどは評価者の特性が評価中に変化しないことを仮定している．しかし，この仮定は現実には成り立たないことがある．評価者の特性が採点の過程で変化する現象は評価者ドリフト (Rater Drift) と呼ばれ，これを考慮したモデルも提案されている．具体的には，一定の時間区分ごとの評価者の厳しさパラメータを導入することで，評価者特性の時間変化を捉えるモデルとなっている．しかし，このモデルでは各時間区分ごとのパラメータが独立しているため，パラメータの推定が不安定となり，評価者特性の時間変化の解釈が難しくなる問題がある．

この問題を解決するため，本研究では，時間区分ごとの評価者の厳しさパラメータにマルコフ性を仮定した新しい項目反応モデルを提案する．提案モデルでは，既存モデルよりも安定したパラメータ推定が可能となり，モデルの性能が改善すると考えられる．本研究では，シミュレーション実験と実データ実験を通して提案モデルの有効性を示す．

2　項目反応理論

本研究では，高精度な能力推定を行うために，IRT を利用する．

現在，評価者特性を最も柔軟に表現できる項目反応モデルとして，一般化多相ラッシュモデルが知られている [2]．このモデルでは，評価者 r が課題 i における受検者 j のパフォーマンスにスコア k を与える確率を次式で定義する．

$$P_{ijrk} = \frac{\exp\sum_{m=1}^{k}\{\alpha_i\alpha_r(\theta_j - (\beta_i+\beta_r) - d_{rm})\}}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\{\alpha_i\alpha_r(\theta_j - (\beta_i+\beta_r) - d_{rm})\}}$$

ここで，α_i は課題 i の識別力，α_r は評価者 r の一貫性，θ_j は受検者 j の能力，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_{rk} は評価者 r のスコア k に対する厳しさを表すステップパラメータである．モデルの識別性のために，$\sum_{i=1}^{I}\log\alpha_i = 0$，$\sum_{i=1}^{I}\beta_i = 0$，$\sum_{k=2}^{K}d_{rk} = 0$，$d_{r1} = 0$ を仮定する．なお，以降で紹介する評価者ドリフトを考慮したモデルや提案モデルでは，課題数が 1 の場合を想定し，課題パラメータは考慮しないこととする．

上記のモデルは評価者の特性が評価中に変化しないことを仮定しているが，この仮定は現実には成り立たないことがある．このような評価者ドリフトを考慮できるモデルとして，時間区分 t における評価者 r の厳しさの変化を反映させるモデルが提案されている [3]．このモデルでは，評価者

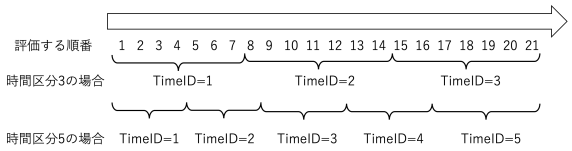

図 1: 時間区分データのイメージと例

r が時間区分 t で採点した受検者 j のパフォーマンスに，スコア k を与える確率を次式で表す．

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}(\theta_j - \beta_r - \pi_r\beta_{rt} - d_m)}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}(\theta_j - \beta_r - \pi_r\beta_{rt} - d_m)}$$

ここで，β_r は評価者 r の初期の厳しさ，π_r は評価者 r の厳しさの変化の傾き，β_{rt} は評価者 r の時間区分 t における厳しさを表す．

このモデルは，一連の採点データを一定の時間区分で区切り，時間区分ごとの評価者パラメータ β_{rt} を導入している点が特徴である．具体的には，図 1 のように，評価者が採点した順に受検者を並べ，データ全体をいくつかの時間区分数に分割し，時間区分を表すインデックス (以降では TimeID と呼ぶ) を各データに付与した上で，各評価者の厳しさパラメータを TimeID ごとに推定する．

このモデルでは，時間区分ごとの評価者特性を分析することが可能である．しかし，このモデルでは時間区分ごとのパラメータが独立していると仮定しているため，パラメータの推定が不安定となり，評価者特性の時間変化について解釈が難しくなる問題がある．この問題を解決するため，時間区分ごとの評価者の厳しさパラメータにマルコフ性を仮定した新しい項目反応モデルを提案する．

3　提案モデル

提案モデルでは，評価者 r が時間区分 t で採点した受検者 j のパフォーマンスに，スコア k を与える確率 P_{jrtk} を次式で定義する．

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}\alpha_r(\theta_j - \beta_{rt} - d_{rm})}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\alpha_r(\theta_j - \beta_{rt} - d_{rm})}$$

$$\begin{aligned}\beta_{rt} &\sim \mathrm{N}(\beta_{r(t-1)},\sigma)\\ \beta_{r1} &\sim \mathrm{N}(0,1)\\ \sigma &\sim \mathrm{LN}(-3,0)\end{aligned}$$

ここで，$N(\mu,\sigma^2)$ は平均 μ，標準偏差 σ^2 の正規分布，$LN(\mu,\sigma^2)$ は平均 μ，標準偏差 σ^2 の対数正規分布を表す．

提案モデルでは，β_{rt} が $\beta_{(r-1)t}$ に依存して決まると仮定している点が特徴である．また，提案モデルでは，より柔軟に評価特性を表現するために，一般化多相ラッシュモデルでも採用されている評価者の一貫性パラメータ α_r と各スコアに対する厳しさパラメータ d_{rk} も導入している．

なお，提案モデルにおける β_{rt} の事前分布のパラメータである σ は，できる限り小さい値とすることで，$\beta_{rt}(t>1)$ の事後分布が縮小するため，パラメータ推定が安定すると期

表 1: パラメータ・リカバリ実験の結果

J	R	T	RMSE				BIAS			
			θ	α_r	β_{rt}	d_{rk}	θ	α_r	β_{rt}	d_{rk}
50	10	3	0.24	0.27	0.20	0.40	0.02	0.02	0.06	0.00
		5	0.30	0.24	0.31	0.38	-0.01	-0.00	-0.03	-0.00
		10	0.32	0.36	0.36	0.41	0.01	0.04	-0.02	0.00
	15	3	0.25	0.30	0.25	0.37	0.03	0.01	0.06	0.00
		5	0.25	0.23	0.19	0.34	0.03	0.01	0.01	-0.00
		10	0.24	0.26	0.33	0.39	0.02	0.02	0.03	0.00
100	10	3	0.26	0.19	0.13	0.28	-0.02	0.00	-0.01	0.00
		5	0.27	0.19	0.13	0.27	0.01	0.00	0.00	0.00
		10	0.26	0.18	0.19	0.30	-0.02	0.02	-0.03	-0.00
	15	3	0.23	0.22	0.20	0.30	0.00	0.02	0.00	0.00
		5	0.23	0.20	0.18	0.31	0.03	0.01	0.06	0.00
		10	0.24	0.25	0.40	0.38	0.01	0.02	-0.01	0.00
	Avg.		0.26	0.24	0.24	0.34	0.01	0.01	0.01	0.00

待できる．この事前知識に合わせて，ここでは，LN(−3, 0) を σ の事前分布として採用した．

また，提案モデルではモデルの識別性のために，$\theta_j \sim N(0,1)$，$\prod_r \alpha_r = 1$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する．

提案モデルのパラメータ推定手法にはマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo methods：MCMC）を用いる．パラメータの事前分布は θ_j，d_{rk}，$\log\alpha_r$，$\beta_{rt} \sim N(0, 1^2)$ とした．本研究では，MCMC のバーンイン期間は 1000 とし，1000~2000 時点までの 1000 サンプルを用いる．

4 シミュレーション実験

本節では，MCMC による提案モデルのパラメータ推定精度をシミュレーション実験により評価する．実験手順は以下の通りである．(1) パラメータの真値を，前節に記載したパラメータの分布に従って生成する．(2) 手順 (1) で生成したパラメータを用いて，提案モデルに従ってデータを生成する．(3) 手順 (2) で生成したデータから MCMC を用いてパラメータ推定を行う．(4) 手順 (3) で得られたパラメータ推定値と手順 (1) で生成したパラメータ真値において，RMSE（Root Mean Square Error）とバイアスを求める．(5) 以上を 5 回繰り返し実行し，RMSE とバイアスの平均値を求める．

上記の実験を，受検者数 J =50，100，評価者数 R =10，15，時間区分数 T =3，5，10 の場合において行った．得点の段階数は K =5 とした．実験結果を表 1 に示す．

表 1 から，先行研究と同様に，受検者数・評価者数の増加に伴い推定精度が改善する傾向が読み取れる．また時間区分の総数 T が多くなると性能が低下する傾向も読み取れる．これはパラメータ数に対するデータ数が減少するためと考えられる．

また，J =100，R =15，T =5 における β_{rt} の推定結果の例を図 2 の左に示す．縦軸が β_{rt}，横軸が TimeID であり，実線が作成したパラメータ真値，点線が推定したパラメータである．このグラフより，作成した真値に近い β_{rt} の推定値が得られていることがわかる．

以上の結果から，MCMC により提案モデルのパラメータを適切に推定できることが確認できた．

5 実データ実験

本章では，実データの適用を通して，提案モデルの有効性を評価する．

本研究では，34 名の被験者にエッセイ課題を与え，そのエッセイを時間区分数 3，5，10 において 34 名の評価者が 5 段階得点で採点したデータに対して提案モデルを適用する．

実データから推定された β_{rt} の例を，図 2 の右に示す．縦軸が β_{rt}，横軸が TimeID であり，各線が一人一人の評価者の β_{rt} の推定値を表す．代表的な特性の評価者については，

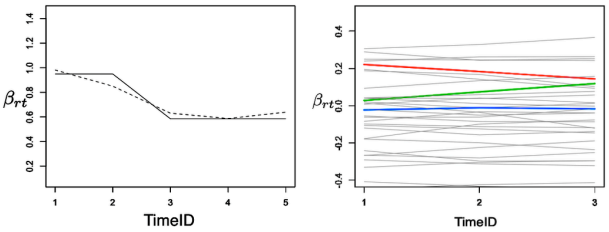

図 2: β_{rt} の推定結果の例

表 2: モデル比較の結果

	時間区分数	既存モデル	提案モデル
WAIC	3	3106.633	**2956.466**
	5	3171.920	**2962.159**
	10	3296.762	**2961.461**
WBIC	3	1825.648	**1808.674**
	5	1916.035	**1807.423**
	10	2016.664	**1806.655**

ハイライトして表示している．図から，赤色で示した評価者は評価中に β_{rt} が減少しており，対して緑色で示した評価者は増加しているとわかる．さらに，青色で示した評価者は評価中に β_{rt} がほぼ変化しないことがわかる．このように，評価者ドリフトの傾向を推定できていることがわかる．

また，提案モデルの性能を評価するために，情報量規準によるモデル比較を行った．ここでは，MCMC により各モデルのパラメータを推定し，得られた推定値を用いて情報量規準を求めた．情報量規準には MCMC のパラメータサンプルから算出できる WAIC（Widely Applicable Information Criterion）と WBIC（Widely Applicable Bayesian Information Criterion）を用いた．WAIC は将来のデータの予測に優れたモデルを選択する規準である．他方で，WBIC は真のモデルを漸近的に選択できる規準である．どちらの場合も，値が小さい方が適したモデルであることを示す．

時間区分数を T = 3，5，10 と変えて実験した結果を表 2 に示す．表 2 から，各時間区分数における WAIC と WBIC の最小値を比較すると，提案モデルが最適モデルとして選択されたことが確認できる．

6 まとめと今後の課題

本研究では，評価者の厳しさパラメータの時間変化にマルコフ性を仮定した新しい項目反応モデルを提案した．また，シミュレーション実験と実データを用いた実験を通して，提案モデルの有効性を示した．今後の課題として，以下のものがある．

- より大規模なデータを収集して，提案モデルの性能を評価する．
- 既存モデルと比べて提案モデルで採用した α_r や d_{rk} の影響について分析を行う．
- 一般化多相ラッシュモデルのように，課題の特性も考慮できるように拡張を行う．

参考文献

[1] F.M. Lord, “Applications of item response theory to practical testing problems,” 1980.

[2] M.Uto and M.Ueno, “A multidimensional generalized many‐facet rasch model for rubric‐based performance assessment,” 2021.

[3] S.W. Raudenbush and A.S. Bryk, Hierarchical linear models: Applications and data analysis methods, vol.1, Sage Publications, 2002.