MI プログラム 卒業論文

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．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が，多数の受験者を長時間かけて採点するような場合にはこの仮定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．しかし，このモデルでは評価者特性の変化を直線的にしか捉えることができないという問題点がある．

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を推定できる新しい IRT モデルを提案する．提案モデルでは，既存モデルよりも柔軟な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 IRT 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．

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

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 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．

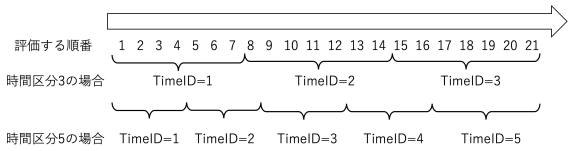

図 1: 時間区分データのイメージと例

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}(\theta_j - \beta_r - \pi_r t - d_m)}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}(\theta_j - \beta_r - \pi_r t - d_m)} \quad (2)$$

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，時間区分ごとのデータを使用して推定を行う．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，インデックス (以降では TimeID と呼ぶ) を各データに付与する．

このモデルでは，時間経過による評価者特性の変化を考慮することが可能である．しかし，このモデルでは，評価者特性の変化を直線的にしかとらえることができないという問題点がある．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を推定できる新しい IRT 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．

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}\alpha_r(\theta_j - \beta_{rt} - d_{rm})}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\alpha_r(\theta_j - \beta_{rt} - d_{rm})} \quad (3)$$

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

ここで，β_{rt} は評価者 r の時間区分 t における厳しさである．また，$N(\mu,\sigma^2)$ 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

表 1: 提案モデルのパラメータ・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}
60	10	3	0.28	0.23	0.19	0.34	-0.01	0.01	-0.01	0.00
		5	0.29	0.26	0.25	0.37	0.00	0.02	-0.03	0.00
		10	0.29	0.23	0.23	0.37	0.00	0.01	0.00	0.00
	15	3	0.23	0.27	0.24	0.38	0.01	0.01	0.01	0.00
		5	0.22	0.22	0.23	0.36	0.00	0.01	0.02	0.00
		10	0.26	0.27	0.31	0.38	-0.02	0.01	-0.04	0.00
120	10	3	0.26	0.19	0.15	0.27	-0.01	0.01	0.00	0.00
		5	0.28	0.21	0.18	0.32	0.00	0.01	-0.01	0.00
		10	0.28	0.20	0.25	0.29	-0.01	0.00	-0.03	0.00
	15	3	0.22	0.19	0.13	0.24	0.02	0.00	0.03	0.00
		5	0.23	0.17	0.18	0.31	0.02	0.02	0.05	0.00
		10	0.24	0.20	0.26	0.30	0.00	0.01	0.00	0.00
Avg.			0.26	0.24	0.24	0.34	0.01	0.01	0.01	0.00

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 =60，90，120，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．実験結果の一部を表 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．

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，134 名分のエッセイ課題を，16 名の評価者が 5 段階得点で採点したデータを使用する．採点は 4 日に分け，日ごとに全体の 1/4 ずつ採点するように指示した．以降では日ごとに時間区分 1，2，3，4 として扱う．全 16 名の被験者のうち，6 人には採点の際に指示を与え，人為的に評価者バイアスのあるデータを作成するようにした．

本研究では，このように収集したデータに対して提案モデルを適用する．

β_{rt} の推定結果をグラフにしたものを図 2 に示す．縦軸が β_{rt}，横軸が時間区分である．また，例として，だんだん厳しくなるように指示した評価者の結果を赤色の線で示した．

これを見ると，指示を与えた評価者は，その指示通りの評価者バイアスを推定できていることがわかる．また，指示を与えていない評価者でも，評価者ごとにパラメータの変化を推定できていることがわかる．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．

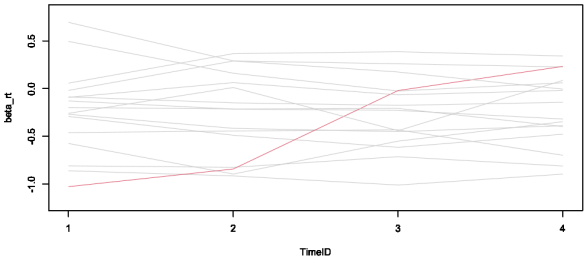

図 2: 実データ実験による β_{rt} の推定結果

表 2: モデル比較の結果

	既存モデル	提案モデル	比較モデル 1	比較モデル 2	比較モデル 3
WAIC	5361.581	**5027.951**	5225.050	5104.463	5032.362
WBIC	3071.706	3033.753	3038.600	3056.349	**3028.649**

また，提案モデルの性能を評価するために，式 (2) の既存モデルとの情報量規準によるモデル比較を行う．

加えて，既存モデルから提案モデルへのいくつかの変更点の中で，どの変更が効果を持っていたかを確認するために，以下の 3 つのモデルとの比較を行う．

比較モデル 1 提案モデルの d_{rk} を d_k に入れ替え

比較モデル 2 提案モデルの β_{rt} を $\beta_r - \pi_r t$ に置き換え

比較モデル 3 提案モデルの α_r を削除

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．表 2 と 3 から，WAIC 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．次に，表 2 と 3 から，WBIC の最小値を比較すると，比較モデル 3 と比較モデル 1 が最も高い性能を示しており，提案モデルより単純なモデルが最適なモデルとして選択されていることがわかる．一方で，提案モデルはどちらにおいても 2 番目に高い性能を示しており，β_{rt} にマルコフ性を導入したことの有効性は確認できる．

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を推定できる新しい IRT 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．今後の課題としては，今回提案したモデルは課題数 1 の問題を想定しており，課題ごとの特性を考慮することが出来ていない．そのため，課題パラメータを追加して課題ごとの特性を考慮したモデルを作ることが挙げられる．

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.

表 3: モデル比較の結果 (指示なしの評価者)

	既存モデル	提案モデル	比較モデル 1	比較モデル 2	比較モデル 3
WAIC	3279.444	**3134.700**	3190.802	3154.331	3137.137
WBIC	1927.215	1900.211	**1863.710**	1924.210	1910.699

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