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，教育評価などの場において，パフォーマンス評価は重要な役割を果たしている．その中で，パフォーマンス評価では，評価者の厳しさや一貫性の違い，各得点の使用傾向の差などにより，採点に偏りが生じ，受検者の能力を正確に測ることができないという問題が発生することがある．この問題を解決するために項目反応理論 (Item response theory:IRT)[1] と呼ばれる数理モデルの利用が近年注目されている．

多くのモデルでは，課題と評価者の特性を考慮した能力推定を行うことができるが，これらのモデルは評価者の特性が評価中に変化しないという仮定のもと成り立っている．しかし，その仮定は現実には成り立たないことがある．この評価者の採点基準が採点過程で変化する特性は評価者ドリフト (Rater Drift) と呼ばれ，これを考慮したモデルについても提案がなされている．既存モデルでは，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が，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難しいと言う問題点がある．

この問題を解決するために，本研究では，時間区分ごとの評価者の厳しさ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．また，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

本研究は，課題・評価者・時間の特性を考慮した高精度な能力推定を行うことを目的とする．このような能力推定を実現するために，本研究では IRT を利用する．

現在最も評価者特性を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評点 k を与える確率を次式で定義する．

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

ここで，d_{rm} 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．

上記のモデルでは，課題と評価者の特性を考慮した能力推定を行うことができるが，これらのモデルは評価者の特性が評価中に変化しないという仮定のもと成り立っている．しかし，実際には評価の仮定で特性が変化する評価者ドリフトが生じる場合がある．評価者特性の時間変化を考慮できるモデルとして，Raudenbush と Bryk[3] は，評価者の厳しさの安定性を調べるために，時間 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を次の式で提案した．

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

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 の時刻 t における時間区分を表す．評価者ドリフトを考慮した項目反応モデルでは，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を導入する．具体的には，図 1 のように，評価者が学習者を評価する順番が早い順にデータを並べ，早い方から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) を付与する．時間区分ごとのデータ数は，学習者数を時間区分数で割った数になる．このモデルは，パフォーマンス評価において，評価者と時間におけ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，このモデルの問題点として次の点が挙げられる．(1) 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難しい．(2) 評価者の一貫性と各得点に対する基準の差異を考慮できていない．以上の問題を解決するため，時間区分ごとの評価者の厳しさにマルコフ性を仮定した新しい項目反応モデルを次節で提案する．

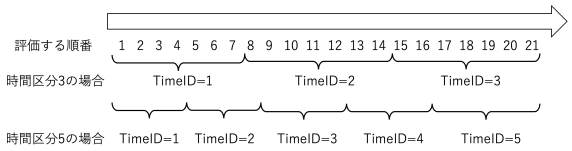

図 1: 時間区分データのイメージと例

3　提案モデル

提案モデルでは，受験者 j のパフォーマンスに，評価者 r が時間区分 t において評点 k を与える確率 P_{jrtk} を次式で定義する．今研究では，課題数は 1 と仮定し，課題パラメータは考慮しないこととする．

$$P_{jrtk} = \frac{\exp\sum\alpha_r(\theta_j - \beta_{rt} - d_{rk})}{\sum\exp\sum\alpha_r(\theta_j - \beta_{rt} - d_{rk})}$$

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

ここで，α_r は評価者 r の一貫性，θ_j は受験者 j の能力，β_{rt} は 評価者 r の時間区分 t における厳しさ，d_{rk} は評価者 r からスコア k を得る困難度を示すステップパラメータである．

提案モデルでは，β_{rt} は，時間変化における評価者の厳しさの変化を表すため，$\beta_{r(t-1)}$ に基づいて β_{rt} を決定されるように設計されている．そのため上記のように仮定する．なお，σ の事前分布に $\mathrm{LN}(-3, 0)$ を設定した理由は次のとおりである．σ は 0 から 1 の間で，なおかつ，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．上記を満たすような分布として，ここでは，$\mathrm{LN}(-3, 0)$ を採用した．

また，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．

本研究では提案モデルのパラメータ推定手法として MCMC 法を用いる．パラメータの事前分布は θ_j, d_{rk}, $\log\alpha_r$,$\beta_{rt} \sim N(0.0, 1.0^2)$ とした．ここで，$N(\mu,\sigma^2)$ は平均 μ，標準偏差 σ の正規分布を表す．本研究では，MCMC のバーンイン期

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

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 アルゴリズム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，モデル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，データを生成する．(3) 手順 (2) で生成したデータから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，平均平方二乗誤差 (RMSE) 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，学習者数 $j = 50, 100$，評価者数 $r = 10, 15$，時間区間 $t = 3, 5, 10$ の場合において行った．カテゴリ数は $K = 5$ とした．実験結果を表 1 に示す．

表 1 から，関連研究と同様に，学習者数・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，β_{rt} のパターンごとの推定結果の例を図 2 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，作成した真値にある程度従って，パラメータが推定さ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 34 名の評価者が 5 段階得点で採点したデータを使用する．本研究では，この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 に示す．縦軸が β_{rt}，横軸が TimeID である．赤色の実線で示した評価者は評価中に厳しさが減少しており，緑色の実線で示した評価者は評価中に厳しさが増加している．対して青色の実線で示した評価者は評価中にあまり厳しさが変化していないことがわかる．

以降では，提案モデルの性能を評価するために，**2.** で紹介した既存モデルとの性能比較を行う．

本節では，情報量規準によるモデル比較により提案モデルの性能を評価する．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．ここで，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基準である．どちらの場合も，値が小さい方が適したモデルであるということを示す．

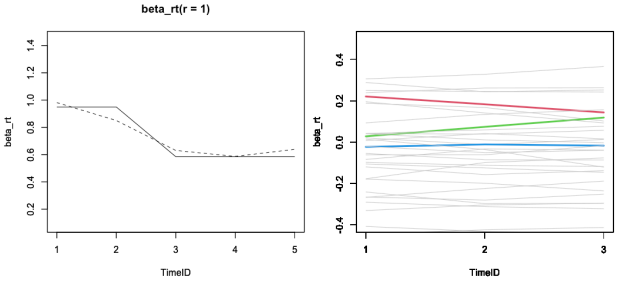

図 2: β_{rt} の推定結果の例

表 2: モデル比較の結果

	時間区分	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**

実験結果を表 2 に示す．縦軸は使用したデータの時間区分であり，横軸は各モデルの WAIC，WBIC の値である．

表 2 から，各時間区分における 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，時間区分ごとの評価者の厳しさ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．

- さらに多くの評価者データを集めて推定を行う．
- 課題特性も考慮して推定を行う．
- 既存モデルとの比較をさらに細かく行う．

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.