評価者の厳しさの時間変化を検出する時系列型ベイズ多相ラッシュモデル

Bayesian Extension of Dynamic Many-Facet Rasch model for Detecting Rater Severity Drift

宇都 雅輝 *1，林 真由 *1
Masaki Uto*1, Mayu Hayashi*1
*1 電気通信大学
*1The University of Electro-Communications
Email: uto@ai.lab.uec.ac.jp

あらまし： 人間の評価者が採点を行うパフォーマンス評価では，採点結果が個別の評価者の厳しさに依存してしまう問題がある．この問題を解決する手法の一つとして，評価者の厳しさの影響を考慮して受検者の能力を推定できる項目反応モデルが提案されてきたが，既存モデルの多くは評価者の厳しさが採点過程で変化しないと仮定している．しかし，この仮定は長時間に渡って採点作業を行う場合には成り立たないことがある．そこで本研究では，そのような評価者の厳しさの時間変化を高精度に推定できる時系列型ベイズ項目反応モデルを提案する．

キーワード： 項目反応理論，評価者バイアス，評価者特性ドリフト，教育測定

1 はじめに

近年，様々な学習評価場面において記述式試験や実技試験などのパフォーマンス評価のニーズが高まっている．一方で，人間評価者による採点を伴うこのような評価では，評価者ごとの厳しさの差異がバイアス要因となり，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．この問題を解決するアプローチの一つとして，評価者の厳しさの影響を考慮して受検者の能力を推定できる項目反応理論（IRT）モデルが提案されてきた (e.g., [1])．それらの既存モデルのほとんどは評価者の厳しさが採点過程で変化しないことを仮定しているが，多数の受検者を採点するような場合には，評価者の厳しさが採点の過程で変化する「評価者特性ドリフト」と呼ばれる現象がしばしば生じる．近年では，評価者の厳しさの時間変化を推定できる IRT モデルも提案されているが (e.g., [2])，それらのモデルは評価者の厳しさを時刻ごとに独立に推定するため，高精度なパラメータ推定が困難である．そこで本研究では，評価者の厳しさの時間依存性を考慮することで，より高精度なパラメータ推定を実現できる時系列型ベイズ項目反応モデルを提案する．

2 評価者特性ドリフトを推定する従来モデル

評価者の厳しさの時間変化を推定できる従来モデルは，伝統的な IRT モデルの一つである多相ラッシュモデルの拡張モデルとして定式化され，評価者 r がある時間区分 t において受検者 j のパフォーマンスに得点 k を与える確率 P_{jrtk} を次式で与える．

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

ここで，θ_j は受検者 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_{rm} は評価者 r の得点 m に対する厳しさを表すステップパラメータである．なお，「時間区分」とは，図 1 のように各評価者の採点データを時間方向に分割して得られた一定の時間幅を持つ区間を表す．

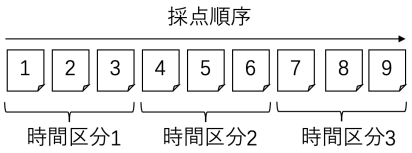

図 1 時間区分の概念図

既存モデルでは，評価者 r の各時間区分 t における厳しさ β_{rt} に独立性（i.i.d）を仮定して推定する．しかし，実際には，評価者の厳しさは時間的に強く依存することが知られているため，その時間依存性を加味することで，より高精度に厳しさパラメータを推定でき，それはモデルの全体的な性能改善に寄与すると考えられる．

3 提案モデル

本研究では，評価者の厳しさにマルコフ性を仮定することで時間的な依存関係を考慮したモデルを提案する．提案モデルでは確率 P_{jrtk} を次式で定義する．

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}(\theta_j - \beta_{rt} - d_{rm})}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}(\theta_j - \beta_{rt} - d_{rm})} \quad (2)$$

$$\begin{cases}\theta_j \sim N(0,1), d_{rm} \sim N(0,1), \beta_{r1} \sim N(0,1)\\ \beta_{rt(t\neq 1)} \sim N(\beta_{r,t-1}, \sigma_r), \sigma_r \sim LN(\mu_\sigma, 1)\end{cases} \quad (3)$$

ここで，$N(\mu,\sigma^2)$ と $LN(\mu,\sigma^2)$ は正規分布と対数正規分布を表す．

従来モデルと比べた提案モデルの主要な特徴は，評価者の厳しさパラメータ β_{rt} にマルコフ性を仮定した分布を設定している点である．具体的には，図 2 に示すように，時間区分 t での厳しさ β_{rt} が直前の時間区分の厳しさ $\beta_{r,t-1}$ に依存するように分布を設定しており，これに

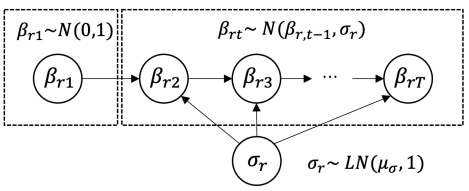

図 2 評価者の厳しさパラメータのグラフィカル表現

表 1　パラメータ推定精度（※従来モデルは σ_r を持たない）

			提案モデル				従来モデル		
J	R	T	θ_j	β_{rt}	d_{rm}	σ_r	θ_j	β_{rt}	d_{rm}
100	5	5	.471	.220	.311	.173	.478	.410	.388
200	10	5	.471	.252	.381	.223	.499	.594	.451
500	10	5	.470	.224	.248	.246	.462	.295	.258
全条件の平均			.470	.215	.310	.254	.475	.321	.329

より時間依存性を考慮できるようになっている．

また，提案モデルの他の特徴として，1）各評価者の厳しさの時間変化の大きさを単一の数値として表現する評価者固有標準偏差パラメータ σ_r と，2）対象評価者集団内における厳しさの時間変化の多様性に関する分析者の事前知識を反映できる事前分布 $LN(\mu_\sigma, 1)$ を導入している点も挙げられる．ただし，紙面の都合上，これらの詳細な議論は文献 [3] に譲る．なお，本稿では $\mu_\sigma = -2$ とする．また，提案モデルのパラメータはマルコフ連鎖モンテカルロ法（MCMC）で推定する．

4　シミュレーション実験

評価者の厳しさの時間依存性を考慮したことでパラメータ推定精度が向上するかを確認するために，提案モデルと従来モデルのパラメータ推定精度をシミュレーション実験で評価した．実験手順は次のとおりである．1）ランダムに生成したパラメータ真値を用いて各モデルからデータを生成した．2）生成したデータから MCMC でパラメータ推定を行い，パラメータの推定値と真値の RMSE を求めた．3）以上を 10 回繰り返し，RMSE の平均値を求めた．この実験を受検者数・評価者数・時間区分数の条件を変えながら行った．紙面の都合上，ここでは一部の条件の結果と全条件の結果の平均のみ表 1 に示す．表 1 から，提案モデルでは厳しさ β_{rt} の推定精度が従来モデルより大幅に向上しており，他のパラメータの推定精度向上も確認できる．このことから提案モデルがパラメータ推定精度向上に寄与することが示せた．

5　実データ実験

本章では，実データ実験を通して提案モデルの有効性を評価する．本実験では，あるエッセイ課題に対する 134 名の解答を，15 名の評価者が 5 段階得点で採点したデータを使用する．評価者には，採点を 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．本実験では，各採点日を時間区分として扱う．なお，15 名のうち 5 名に対しては採点時に指示を与え，人為的にバイアスを加えた．具体的には，3 名の評価者（評価者番号 11〜13）には「日ごとに徐々に厳しくせよ」，「日ごとに徐々に甘くせよ」，「2 日目は厳しく，3 日目は甘く，4 日目は厳しくせよ」という指示をそれぞれ与え，残りの 2 名（評価者番号 14，15）には「得点 2 と 3 を中心的に使用せよ」のように段階得点の使用に制限を与えた．

全 15 名の評価者のデータを用いた場合と人為的にバイアスを加えた 5 名の統制評価者を除いたデータを用いた場合のそれぞれについて，提案モデルと従来モデルに

表 2　提案モデルと従来モデルの比較結果

	全評価者		統制評価者除外	
	WAIC	WBIC	WAIC	WBIC
提案モデル	**4661.7**	**2822.30**	**3131.9**	**1906.3**
従来モデル	4686.5	2880.4	3152.2	1959.5

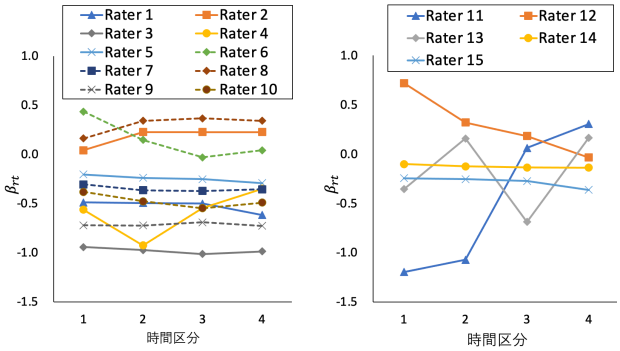

図 3　通常評価者 10 名（左）と統制評価者 5 名（右）の厳しさパラメータ β_{rt} の推定値

における情報量規準（WAIC と WBIC）を計算した．表 2 に実験結果を示す．表では最適値を意味する最小値を太字で示した．表から，いずれの場合でも提案モデルが最適モデルとして選択されており，評価者の厳しさに時間依存性を加えたことの有効性が示された．

図 3 に各評価者の厳しさ β_{rt} の推定値を示す．図の縦軸は β_{rt} の値，横軸は時間区分，各線はそれぞれの評価者を表す．図から，指示を与えた統制評価者 11〜13 については，指示通りの厳しさの遷移が推定されており，提案モデルが適切に厳しさの変化を推定できたことがわかる．さらに，指示を与えていない評価者の中にも，評価者 4 や 6 など，厳しさの変化が比較的大きい評価者が見受けらる．反対に，その他の評価者については厳しさが比較的安定している傾向も読み取れる．

また，紙面の都合上詳細は割愛するが，1）提案手法で導入した σ_r が厳しさの時間変動の大きさに比例した推定値を示したことと 2）統制評価者 14 と 15 について，指示通りの得点の使用傾向がステップパラメータ d_{rm} に推定されていたことも確認できた．

6　まとめ

本研究では，評価者の厳しさの時間変化を高精度に推定できる新しい IRT モデルを提案した．本稿で説明を省略した提案モデルの特徴は発表当日に説明する．

参考文献

[1] M. Uto. A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. *Behaviormetrika*, Vol. 48, No. 2, pp. 425–457, 2021.
[2] C. M. Myford and E. W. Wolfe. Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *J. Educ. Meas.*, Vol. 46, No. 4, pp. 371–389, 2009.
[3] 宇都雅輝・林真由．評価者特性の時間変動を推定する時系列型ベイズ多相ラッシュモデル．日本行動計量学会第 50 回大会，2022.