

目次

第 1 章	はじめに	1
第 2 章	項目反応理論	3
2.1	2 値型項目反応モデル	3
2.2	部分採点モデル (PCM)	4
2.3	一般化部分採点モデル (GPCM)	4
第 3 章	評価者特性を考慮した項目反応モデル	6
3.1	多相ラッシュモデル	6
3.2	評価者パラメータを付与した GPCM	7
3.3	一般化多相ラッシュモデル	7
第 4 章	評価者ドリフトを考慮した項目反応モデル	8
4.1	時間パラメータについて	8
4.2	評価者ドリフトを考慮した IRT	9
4.3	既存モデルの課題	9
第 5 章	提案モデル	10
5.1	パラメータの解釈	10
5.2	パラメータ推定手法	12
第 6 章	シミュレーション実験	14
6.1	パラメータ真値	14
6.2	実験結果	15
第 7 章	実データ実験	19
7.1	評価者特性	19
7.2	情報量基準におけるモデル比較	21
第 8 章	むすび	23
付録 A	stan コード	24
参考文献		26

第1章

はじめに

近年、受検者の論理的思考力や表現力などの実践的な能力を測定する方法の一つとして、パフォーマンス評価が注目されている。大学入学共通テストでも、論述式問題を採用する議論がなされていた。他方で、パフォーマンス評価における問題の一つが、採点の信頼性に関わる点である。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、採点に偏りが出てしまうことがあり、このことが評価の信頼性低下の要因となる。このような問題は大学入試や資格試験、教育評価などにおける、レポート課題、グループディスカッションやプレゼンテーション課題などの様々なパフォーマンス評価場面で起こり得る。この問題を解決するために項目反応理論 (Item response theory:IRT)[1, 2] と呼ばれる数理モデルの利用が近年注目されている。IRT は、コンピュータ・テストの普及とともに、近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つで、テストを作成・実施・評価・運用するための数理モデルである。

一般的な客観式テストに適用される IRT モデルは、受検者の能力と課題の困難度の 2 つのパラメータからなるモデル [3] である。しかし、それらのモデルは複数の評価者が採点を行うパフォーマンス評価には適用することができない。なぜなら、評価者によって一貫性や厳しさは様々であるが、従来のモデルではそのような特性差を考慮できないためである。この問題を解決するために、評価者特性を表すパラメータを加えた IRT モデルが近年多数提案されている [4, 5, 6, 7]。これらのモデルは、受検者の能力と課題の困難度の他に、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をパラメータとして加えたモデルであり、これを利用することで評価者の特性差の影響を考慮した受検者の能力推定が可能となる。

これらのモデルは評価者の基準が採点過程で変化しないという仮定のもと成り立っている。しかし、多数の受検者を長時間かけて採点するような場合、評価者の特性が採点の過程で変化する評価者ドリフト (Rater Drift) と呼ばれる現象がしばしば生じる。評価者ドリフトが生じている場合に、評価者の基準が変化しないと仮定したモデルを適用してしまうと、適切な評価者特性の推定を行うことができず、能力推定値にもバイアスが生じると考えられる。この問題を解決するために、評価者特性の時間変動を考慮した IRT モデルが提案されている [8]。このモデルは、評価者が行なった一連の評価を時間区分で区切り、評価者の初期の厳しさと厳しさの傾きを使用して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、既存モデルの問題点として次の点が挙げられる。

1. 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
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を推定できる新しい IRT モデルを提案する。具体的には、各時間区分における評価者の厳しさパラメータを導入し、そのパラメータにマルコフ性を仮定して推定するモデルを提案する。また、提案モデルには、評価者の一貫性と、各段階得点に対する各評価者の基準の差異を考慮したパラメータも付与する。提案モデルの利点は次のとおりである。

1. 時間区分ごとの評価者特性の変化をより柔軟に推定することができるため、データへのモデル適合度が向上し、モデルの性能が改善する。
2. 評価者ごとの一貫性の差異、及び各段階得点に対する基準の差異をより柔軟に考慮できるため、データへのモデル適合度が向上し、モデルの性能が改善する。

本研究では、シミュレーション実験および実データ実験を通して提案モデルの有効性を示す。

第 2 章

項目反応理論

本研究は、課題・評価者・時間区分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。

IRT は、コンピュータ・テストの普及とともに、近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つで、テストを作成・実施・評価・運用するための実践的な数理モデルである [1, 2]。IRT の利点として、次のような点が挙げられる [9]。

- 推定制度の低い異質項目の影響を小さくして能力推定を行うことができる。
- 異なる項目への受検者の反応を同一尺度上で評価できる。
- 欠測データから容易にパラメータを推定できる。

IRT はこれまで、正誤判定問題や選択式問題などの正誤を一意に判定できる 2 値型データを扱うテストへの利用が一般的であった。一方で、近年では論述式・記述式試験のような多段階カテゴリを用いた評価データに対して、多値型 IRT モデルを適用してパフォーマンスを評価する応用的な研究も進められている [10, 11]。次節では、IRT における基礎的なモデルとして、2 値型データを扱うモデルについて述べる。その後、その拡張であり、本研究で扱うリッカート型データに適用できる代表的な多値型 IRT モデルを紹介する。

2.1 2 値型項目反応モデル

最も基礎的な IRT モデルとしては、ラッシュモデル [3] が知られている。ラッシュモデルでは、受検者 j が項目 i に正答する確率を次の式で表す。

$$P_{ij} = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad (2.1)$$

ここで、 θ_j は受検者 j の能力、 b_i は項目 i の困難度を表す。

ラッシュモデルは、少数のパラメータで表現されているため、少数データからでも高精度にパラメータを推定できる [12]。しかし、そのモデルの単純さにより、複雑な特性を表現できないため、データへの当てはまりが悪い場合が多い [13]。そのため、実際には、ラッシュモデルに項目 i の識別力パラメータ α_i を加えた 2 母数ロジスティックモデルを使用することが一般的である。2 母数ロジスティックモデルでは、受検者 j が項目 i に正答する確率を次の式で表す。

$$P_{ij} = \frac{\exp(\alpha_i(\theta_j - b_i))}{1 + \exp(\alpha_i(\theta_j - b_i))} \quad (2.2)$$

ここで、 α_i は項目 i の識別力を表す。識別力 α_i の値が大きいほど、能力値 $\theta = b_i$ 付近の能力を高い精度で識別できる。

ラッシュモデルや2母数ロジスティックモデルは、テスト項目に対する受検者の反応が、正誤のような2値データで表される場合に適用できる。しかし、本研究のようなパフォーマンス評価では、段階的な評価カテゴリ(得点)が対応づけられた評価基準に基づいて採点を行うことが多いため、評価データは多値データとなることが一般的である。このような多値データに適用できるモデルとして、2値型IRTモデルを拡張した多値型IRTモデルが提案されてきた。

2.2 部分採点モデル (PCM)

Mastersによって提案された多値型IRTモデルとして、部分採点モデル (Partial Credit Model)[14]が知られている。PCMでは、受検者 j が項目 i に対してカテゴリ $k \in \{1 \dots K\}$ と反応する確率 P_{ijk} を次の式で表す。

$$P_{ijk} = \frac{\exp \sum_{m=1}^k (\theta_j - \beta_{im})}{\sum_{l=1}^K \exp \sum_{m=1}^l (\theta_j - \beta_{im})} \quad (2.3)$$

ここで、 β_{ik} は項目 i においてカテゴリ $k-1$ からカテゴリ k に遷移する困難度を表し、ステップパラメータと呼ばれる。PCMではモデル識別性のために $\beta_{i0} = 0$ を所与とする。

PCMでは、カテゴリ k への反応確率 P_{ijk} とカテゴリ $k-1$ への反応確率 P_{ijk-1} のロジット $\log(P_{ijk}/P_{ijk-1})$ を受検者の能力とカテゴリ k における項目困難度の線形和 $\theta_j - \beta_{ik}$ で定義しており、項目への正答確率と誤答確率のロジットを $\theta_j - \beta_i$ と定義するラッシュモデルの多値への一般化と解釈できる。

2.3 一般化部分採点モデル (GPCM)

Murakiは、PCMにおける項目識別力一定の制約を緩和したモデルとして、一般化部分採点モデル (Generalized Partial Credit Model:GPCM)[15]を提案している。GPCMでは受検者 j が項目 i に対してカテゴリ k と反応する確率 P_{ijk} を次の式で表す。

$$P_{ijk} = \frac{\exp \sum_{m=1}^k (\alpha_i(\theta_j - \beta_{im}))}{\sum_{l=1}^K \exp \sum_{m=1}^l (\alpha_i(\theta_j - \beta_{im}))} \quad (2.4)$$

PCMと同様に、モデル識別性のため、 $\beta_{i0} = 0$ を所与とする。

GPCMは、Andrichによる評定尺度モデル[16]と同様に、ステップパラメータ β_{ik} を $\beta_i + d_k$ 、あるいは $\beta_i + d_{ik}$ と分解することができ、以下のようなモデルで表されることもある。

$$P_{ijk} = \frac{\exp \sum_{m=1}^k (\alpha_i(\theta_j - \beta_i - d_{im}))}{\sum_{l=1}^K \exp \sum_{m=1}^l (\alpha_i(\theta_j - \beta_i - d_{im}))} \quad (2.5)$$

$$P_{ijk} = \frac{\exp \sum_{m=1}^k (\alpha_i(\theta_j - \beta_i - d_m))}{\sum_{l=1}^K \exp \sum_{m=1}^l (\alpha_i(\theta_j - \beta_i - d_m))} \quad (2.6)$$

ここで、 β_i は項目 i の困難度パラメータ、 d_{ik} は項目 i のカテゴリ k に対する閾値パラメータ、 d_k はカテゴリ k のカテゴリパラメータと呼ばれる。モデルの識別性のために、 $d_{i1} = 0$, $\sum_{k=2}^K d_{ik} = 0$, $d_1 = 0$, $\sum_{k=2}^K d_k = 0$ を所与とする。

第3章

評価者特性を考慮した項目反応モデル

これまで紹介してきた IRT モデルは、受検者とテスト項目の 2 相データへの適用を想定している。一方で、本論で想定するパフォーマンス評価データ X は、パフォーマンス課題 $i \in \{1 \dots I\}$ に対する受検者 $j \in \{1 \dots J\}$ のパフォーマンスに評価者 $r \in \{1 \dots R\}$ が与える評点 $k \in \{1 \dots K\}$ の集合であり、以下のような 3 相データとして定義される。

$$X = \{x_{ijr} | x_{ijr} \in \{-1, 1, \dots, K\}\} (i \in \{1 \dots I\}, j \in \{1 \dots J\}, r \in \{1 \dots R\}) \quad (3.1)$$

このような 3 相データに対して、これまで紹介した一般的な IRT モデルを直接適用することはできない。この問題の解決策として、評価者特性を表すパラメータを加えた IRT モデルが提案されてきた (e.g., [4, 5, 6])。これらの IRT モデルは、従来の IRT モデルにおいて、項目パラメータを課題の特性を表すパラメータとみなし、評価者特性を表すパラメータを付与したモデルとして定式化される。

3.1 多相ラッシュモデル

多相データのための IRT モデルとして、最も広く知られているモデルは、Linacre が提案した多相ラッシュモデル (MFRM: many-facet Rasch model) [4] である。MFRM は、ラッシュモデルに、課題と受検者以外の要因を表すパラメータを付与したモデルである。例えば、評価者 r の評価の厳しさを表すパラメータ β_r を付与した多相ラッシュモデルでは、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ポジティブな判定を与える確率を次式で与える。

$$P_{ijr} = \frac{\exp(\theta_j - b_i - \beta_r)}{1 + \exp(\theta_j - b_i - \beta_r)} \quad (3.2)$$

ここで、 b_i は課題 i の困難度を表す。MFRM は 2 値データを扱うモデルであるが、ラッシュモデルと同様に、PCM を用いた多値への拡張モデルも提案されている。MFRM の PCM による拡張モデルは、受検者・課題・評価者・評点間にどのような作用を仮定するかによって、いくつかのモデル化が考えられる [17]。ここでは、最も単純な多値型 MFRM である Common step モデルを紹介する。

Common step モデルは、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。

$$P_{ijk} = \frac{\exp \sum_{m=1}^k (\theta_j - b_i - \beta_r - d_m)}{\sum_{l=1}^K \exp \sum_{m=1}^l (\theta_j - b_i - \beta_r - d_m)} \quad (3.3)$$

ここで、 d_k は評点 $k-1$ から評点 k に遷移する困難度を表す。パラメータ識別性のために $\beta_1 = 0$, $d_1 = 0$ を仮定する。

MFRM では、全ての課題で識別力が一定であること、また、全ての評価者の一貫性が一定であることが仮定される。しかし、一般に、パフォーマンス評価ではこれらの仮定は成り立たないことが指摘されている [18, 9]。そこで、この制約を緩めたモデルとして、MFRM を GPCM により拡張したモデルが提案されてきた。

3.2 評価者パラメータを付与した GPCM

Patz and Junker は、課題 i における評価者 r の評価の厳しさを表すパラメータ ρ_{ir} を付与した GPCM の拡張モデルを提案している [7]。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。

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k \{\alpha_i(\theta_j - \rho_{ir} - \beta_{im})\}}{\sum_{l=1}^K \exp \sum_{m=1}^l \{\alpha_i(\theta_j - \rho_{ir} - \beta_{im})\}} \quad (3.4)$$

ここで、 α_i は課題 i の識別力を表し、 β_{ik} は課題において評点 $k-1$ から評点 k に遷移する困難度を表す。モデルの識別性のために、 $\beta_{i1} = 0, \rho_{i1} = 0$ を仮定する。

さらに、宇佐美 (2010) は、評価者内・評価者間で評価が一貫している保証がないことを指摘し、これに対応する評価者パラメータを加えた以下の GPCM 拡張モデルを提案している。

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k \{\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{im} d_r)\}}{\sum_{l=1}^K \exp \sum_{m=1}^l \{\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{im} d_r)\}} \quad (3.5)$$

ここで、 α_r は評価者 r の評価の一貫性、 β_i は課題 i の位置パラメータ、 d_{ik} は課題 i における評点 k に対する閾値パラメータ、 d_r は評価者 r による評点のばらつきの大きさを表す。モデルの識別性のために、 $\prod_r \alpha_r = 1, \sum_r \beta_r = 0, \prod_r d_r = 1, d_{i1} = 0, \sum_{k=1}^K d_{ik} = 0$ を仮定する。このモデルでは、 α_r により評価者の一貫性を考慮できるだけでなく、 d_r により評価者の中心化傾向も考慮できる点が特徴と言える。

3.3 一般化多相ラッシュモデル

宇佐美のモデルは、全ての評価者が等間隔の評価尺度を持っていると仮定しているが、Uto and Ueno[19] はこの仮定は成り立たないとして、一般化多相ラッシュモデル (g-MFRM: generalized MFRM) を提案している。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。

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k \{\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{rm})\}}{\sum_{l=1}^K \exp \sum_{m=1}^l \{\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{rm})\}} \quad (3.6)$$

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。このモデルは、現在最も評価者特性を柔軟に表現できるモデルである。そのため本研究では、このモデルを拡張する形でモデルを提案する。

第 4 章

評価者ドリフトを考慮した項目反応モデル

3 章で紹介したモデルでは，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．この章では，時間による特性の変化を考慮した IRT モデルを紹介する．

4.1 時間パラメータについて

評価者ドリフトを考慮した IRT 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，時間区分ごとのデータを使用して推定を行う．具体的には，図 4.1 のように，評価者が受検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，受検者数を時間区分数で割った数になる．例えば受検者数 21，時間区分が 3 の場合はデータを 3 つに分割し，時間区分ごとのデータ数は 7 となる．

なお，本研究では，割り切れない場合に，最後の時間区分にあまりのデータを含むこととする．例えば受検者数 21，時間区分が 5 の場合は TimeID=1 ~ 4 では時間区分ごとのデータ数は 4 であるが，最後の TimeID=5 の時に余りのデータを含めるため，TimeID=5 におけるデータ数は 5 となる．

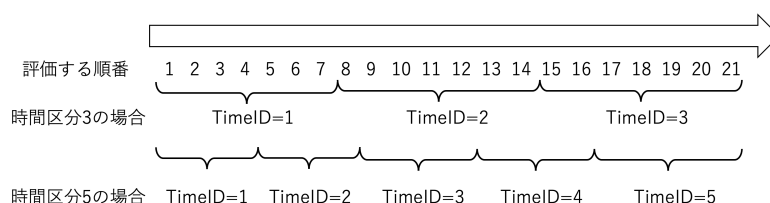


図 4.1: 時間区分データのイメージと例

4.2 評価者ドリフトを考慮した IRT

Wolfe et. al[20] は、一般的な多相ラッシュモデルに時間のパラメータを追加したモデルを提案している。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時間区分 t において評点 k を与える確率を次の式で定義する。

$$P_{ijrtk} = \frac{\exp \sum_{m=1}^k (\theta_j - \beta_i - \beta_r - \beta_t - d_m)}{\sum_{l=1}^K \exp \sum_{m=1}^l (\theta_j - \beta_i - \beta_r - \beta_t - d_m)} \quad (4.1)$$

ここで、 β_t は時間区分 t における全ての評価者の平均的な評価の厳しさを表す。このモデルは、全ての評価者の特性が同じ傾向で時間変動すると想定しているが、実際には変化の度合いも評価者によって異なる。そこで、評価者ごとの変化の度合いも考慮できるモデルとして、Raudenbush と Bryk[21] 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。

$$P_{ijrtk} = \frac{\exp \sum_{m=1}^k (\theta_j - \beta_i - \beta_r - \pi_r t - d_m)}{\sum_{l=1}^K \exp \sum_{m=1}^l (\theta_j - \beta_i - \beta_r - \pi_r t - d_m)} \quad (4.2)$$

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。また、 t は時間区分である。このモデルでは、各評価者の初期の厳しさと、厳しさの傾きが変化し、個々の評価者ごとに結果を得ることができる。重要なのは、時間区分ごとの変化である π_r で、どの評価者が試験サイクルごとに厳しさが増加する傾向があるかを示している。

4.3 既存モデルの課題

本節では、4.2 で紹介したモデルが抱える問題点について述べる。

4.2 で紹介したモデルは、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. 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
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を推定できる新しい IRT モデルを次節で提案する。

第 5 章

提案モデル

提案モデルでは、受検者 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。なお、本研究では、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。

$$P_{jrtk} = \frac{\exp \sum \alpha_r (\theta_j - \beta_{rt} - d_{rk})}{\sum \exp \sum \alpha_r (\theta_j - \beta_{rt} - d_{rk})} \quad (5.1)$$

$$\beta_{rt} \sim N(\beta_{r(t-1)}, \sigma)$$

$$\beta_{r1} \sim N(0, 1)$$

$$\sigma \sim \text{LN}(-3, 0)$$

ここで、 α_r は評価者 r の一貫性、 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_{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。また、 $N(\mu, \sigma)$ は平均 μ 、標準偏差 σ の正規分布を表し、 $\text{LN}(\mu, \sigma)$ は平均 μ 、標準偏差 σ の対数正規分布を表す。

提案モデルでは、 β_{rt} が $\beta_{r(t-1)}$ を平均とする正規分布に従うと仮定している。ただし、ある時間区分における評価者の厳しさは直前の時間区分時点での厳しさから一般には大きく変化しないと考えられる。そこで、ここでは β_{rt} が従う分布の標準偏差 σ が小さくなるように、 σ の事前分布に $\text{LN}(-3, 0)$ を採用している。

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。提案モデルの利点として、以下の点が挙げられる。

1. 時間区分ごとの評価者特性の変化をより柔軟に推定することができるため、データへのモデル適合度が向上し、モデルの性能が改善する。
2. 評価者ごとの一貫性の差異、及び各段階得点に対する基準の差異をより柔軟に考慮できるため、データへのモデル適合度が向上し、モデルの性能が改善する。

これらの特徴は、第 4 章で述べた既存モデルの問題点を解決することができるものである。

5.1 パラメータの解釈

本節では、提案モデルの評価観点パラメータと評価者パラメータの解釈について説明する。このために、カテゴリ数 $K = 5$ において、表 5.1 のパラメータを所与とした時の項目反応曲線 (ICC: Item Characteristic Curve) を図 5.1 に示す。なお、表 5.1 では、パラメータの意味が理解しやすい例を示

表 5.1: 図 5.1 で使用するパラメータ

評価者	時間区分	α_1	β_{rt}	d_{r1}	d_{r2}	d_{r3}	d_{r4}	d_{r5}
評価者 1	時間区分 1	1.0	0.0	0.0	-1.0	0.0	0.5	1.0
	時間区分 2	1.0	0.5	0.0	-1.0	0.0	0.5	1.0
	時間区分 3	1.0	-0.5	0.0	-1.0	0.0	0.5	1.0
		α_r	β_{rt}	d_{r1}	d_{r2}	d_{r3}	d_{r4}	d_{r5}
評価者 2	時間区分 1	1.0	0.0	0.0	-1.0	0.0	0.2	1.0
評価者 3	時間区分 1	2.0	0.0	0.0	-1.0	0.0	0.5	1.0
評価者 4	時間区分 1	0.5	0.0	0.0	-1.0	0.0	0.5	1.0

すために、モデルの識別性の条件式を必ずしも満たさないパラメータ値を用いているが、条件式を満たす値でも解釈は同様である。各図は、横軸が受検者の能力 θ_j を表し、縦軸が各評点への反応確率 P_{ijrtk} を表す。図 5.1 から、いずれの ICC においても、能力が低いほど低い評点を得る確率が高く、能力が高いほど高い評点を得る確率が高くなっていることがわかる。

ここで、表 5.1 の (a), (b), (c) は β_{rt} 以外のパラメータ一定のもとで β_{rt} を変更した場合に対応し、(a) と (d), (e), (f) は β_{rt} が一定のもとで他のパラメータを変更した場合に対応している。まず、時間区分に関わるパラメータの解釈を説明するために (a) を基準に (b) と (c) を比較する。

(b) は (a) と比較して β_{rt} の値が大きくなっている。これは、時間区分 1 の時よりも時間区分 2 の時の方が評価者の厳しさが大きくなっているという意味である。(b) の ICC では、(a) と比べて全体的に右に移動していることが確認できる。これは、 β_{rt} が大きくなっている時間区分では、良い評価を得るためにより高い能力が必要となっていることを示している。

(c) は (a) と比較して β_{rt} が小さくなっている。これは、時間区分 1 の時よりも時間区分 3 の時の方が評価者の厳しさが小さくなっているという意味である。(c) の ICC では、(a) と比べて全体的に左に移動していることが確認できる。これは、 β_{rt} が小さくなっている時間区分では、良い評価を得るために要求される能力値が低くなっていることを示している。提案モデルでは、このように各時間区分における評価者の厳しさを、評価者ごとに表現する。

次に、評価者特性の解釈を説明するために (a) を基準に (d), (e), (f) を比較する。

(d) は (a) と比較して、 d_{r3} と d_{r4} の差が小さく、 d_{r4} と d_{r5} の差が大きくなっている。このパラメータは、隣接する $d_{rk+1} - d_{rk}$ の差が大きくなるほど、評点 k と評点 $k+1$ の基準の乖離が大きいことを意味する。(a) の ICC と比較すると、評点 4 への反応確率が高くなる能力値の範囲が広く、評点 3 への反応確率が高くなる能力値の範囲が狭くなっている。提案モデルでは、このように各カテゴリに対する評価基準を、評価者ごとに表現する。

(e) は (a) と比較して、 α_r の値が大きくなっている。ICC を比べると、(a) よりも曲線の勾配が大きくなっている。これは、一貫性の高い評価者は、受検者の能力と関連した評点を与えるとともに、同等の能力の受検者には安定して同一の評点を与える傾向が強いことを表現している。

(f) は (a) と比較して、 α_r の値が小さくなっている。ICC を比べると、(a) よりも曲線の勾配が小さくなっている。これは、一貫性の低い評価者は、評価にばらつきがあり、受検者の能力と評点の相関が小さくなることを示している。

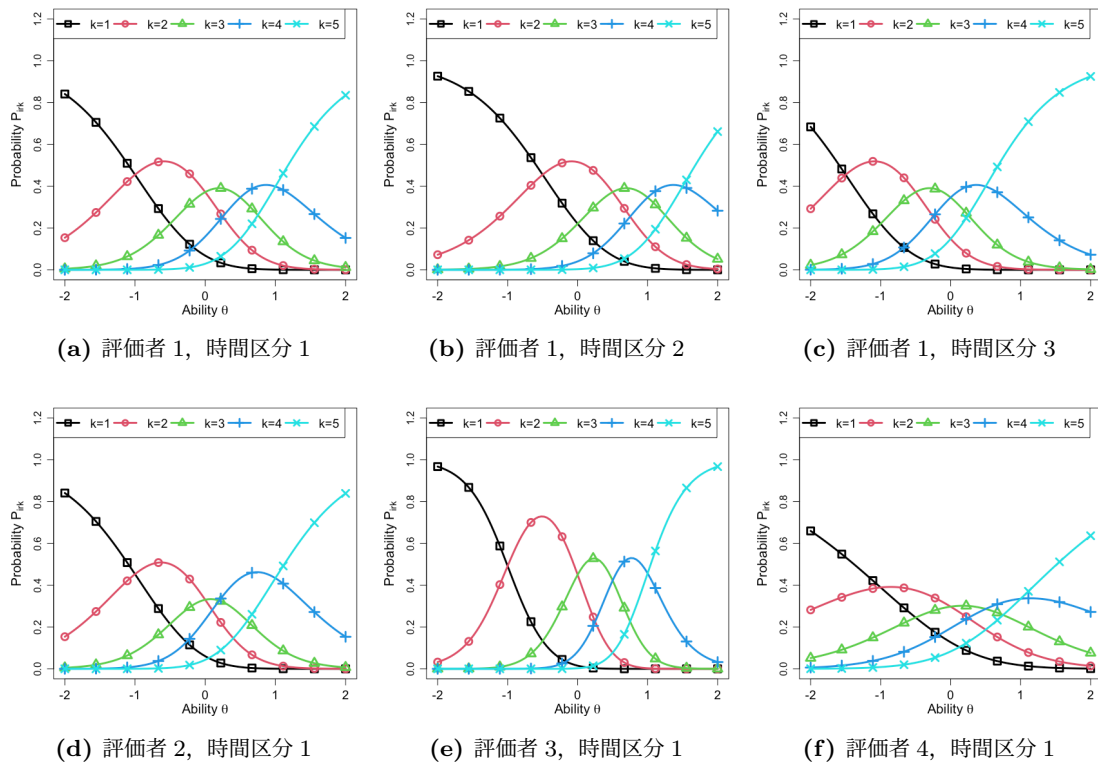


図 5.1: 表 5.1 のパラメータを適用した場合の ICC

5.2 パラメータ推定手法

本節では提案モデルのパラメータ推定手法について述べる．IRT のパラメータ推定手法としては，EM アルゴリズムを用いた周辺最尤推定法やニュートンラフソン法による事後確率最大化推定法が広く用いられてきた．一方で，本研究で扱うような複雑な IRT モデルの場合には，マルコフ連鎖モンテカルロ (MCMC: Markov Chain Monte-Carlo) を用いた期待事後確立推定法が高精度であることが知られている [9, 22]．

IRT における MCMC アルゴリズムとしては，メトロポリタンヘイスティングスとギブスサンプリングを組み合わせたアルゴリズム (Gibbs/MH)[9, 7, 23] が利用されてきた．このアルゴリズムは，単純で実装が容易である反面，目標分布への収束が遅いという問題がある [24, 25]．

Gibbs/MH より効率の良い MCMC アルゴリズムとして，ハミルトニアンモンテカルロ (HMC) が知られている [26]．HMC では，ステップサイズとシミュレーション長という 2 つの決定変数を適切に選択することで，自己相関の低い良質なサンプルを得ることができ，高速に目標分布に収束することが知られている [24, 27]．近年では，HMC の決定変数をサンプリングの過程で最適化できる No-U-TrunSampler(NUT)[24] と呼ばれる手法が提案されている．NUT による MCMC は，Stan[28] と呼ばれるライブラリの整備により，さまざまな数理モデルに容易に適用できるようになったため，IRT を含む様々な統計・機械学習モデルの推定に近年広く利用されている [29, 30, 31]．

以上より, 本研究では提案モデルのパラメータ推定手法として Stan を用いた NUT による MCMC 法を用いる. 実装は RStan[32] を用いて行った. 提案モデルの Stan コードは付録に示した. パラメータの事前分布は $\theta_j, d_{rk}, \log \alpha_r, \beta_{r1} \sim N(0.0, 1.0^2)$ 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 のバーンイン期間は 1000 とし, 1000~2000 時点までの 1000 サンプルを用いる.

第 6 章

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と、式 (4.2) の既存モデル (以下では「既存モデル」と書く) 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、受検者数 $J = 60, 90, 120$ 、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。

6.1 パラメータ真値

この実験で使用するパラメータの真値について説明する。 β_{rt} 以外のパラメータの真値は、主に 5 章にて説明した分布に従って生成したが、一方で、 β_{rt} については個別でパターンを作成し、生成するようにした。作成したパターンとしては、

採点中に一回厳しさが変化する

初期値 β_{r1} を正規分布に従う乱数で生成し、採点の真ん中のタイミング ($T=10$ なら $t=5$ から $t=6$ に変わる時) で $\beta_{r1} + \delta$ に変化させる。なお、 δ は $N(0, 0.2)$ に従う乱数である。

採点中に時間区分ごとに厳しさがほとんど変化しない

初期値 β_{r1} を正規分布に従う乱数で生成し、次の時間区分からは $\beta_{r1} + \delta_t$ となる。なお、 δ_t は $N(0, 0.01)$ に従う 0 に近い値を取る乱数であり、 t が変わることに変化する。

採点中に厳しさが線形変化する

初期値 β_{r1} を正規分布に従う乱数で生成し、次の値からは $\delta_r \times t + \beta_{r1}$ となる。なお、 δ_r は $N(0, 0.01)$ に従う乱数である。

の 3 種類を作成した。これらのパターンは、それぞれ全評価者数の 3 分の 1 ずつになるように作成

した.

6.2 実験結果

提案モデルの実験結果を表 6.1 に示す. 表 6.1 から, 全パラメータの RMSE の平均値は 0.2 ~ 0.3 程度となり, パラメータ別の最大値でも 0.38 までに収まっていることがわかる. 0.2 や 0.38 という値は, 標準正規分布に従うサンプルの 99.73% が含まれる範囲 $(-3 \sim 3)$ の 3.3% と 6.3% に相当し, 十分に小さい値と解釈できる. また, 関連研究 [33, 34, 9] 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も読み取れる. 時間区分数の増加に伴って推定精度が改善する傾向は読み取れないが, これは時間区分数の増加はサンプル数が増えるわけではなく, データの中身が変化するだけであるからと考えられる. バイアスの平均については, いずれのパラメータも 0 に非常に近い値を示しており, 系統的な過大 (または過少) 推定の傾向もないことが確認できる. また, MCMC の収束を示す Gelman-Rubin の収束判定指標 \hat{R} [35, 36] を確認したところ, すべての場合で一般的な収束判定基準値である 1.1 を下回っていた.

また, β_{rt} のパターンごとの推定結果の例を図 6.1 に示す. 実線が作成したパラメータ真値, 波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.

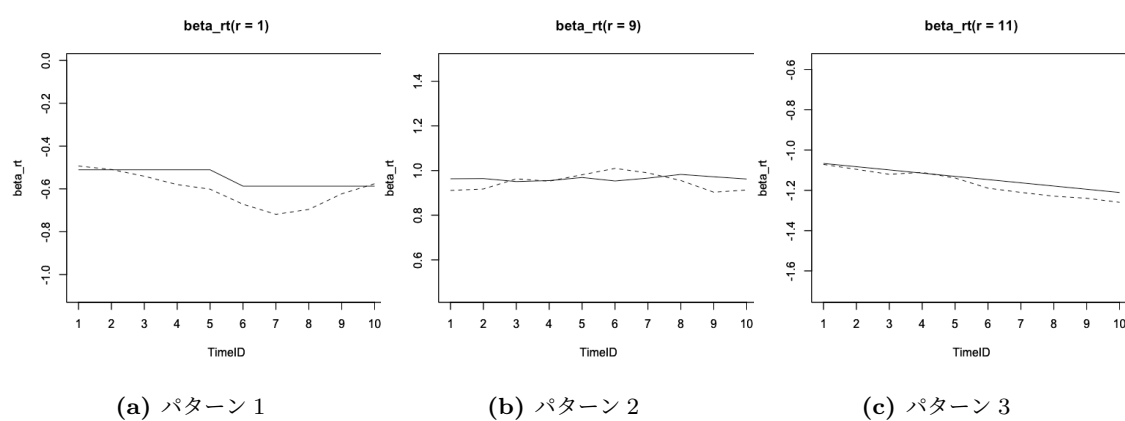
また, 表 6.2 に既存モデルの実験結果を示す. 表 6.2 から, 提案モデルと比べると RMSE, バイアスともにやや悪い傾向があるが, 全体としては受検者数や評価者数の増加に伴う推定精度の改善傾向は確認でき, 適切に推定されていると言える. また, 収束判定指標の \hat{R} [35, 36] も全て 1.1 を下回っており, MCMC の収束も確認できた. 以上の結果から, MCMC により既存モデルのパラメータを適切に推定できることが確認できた.

表 6.1: 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
90	10	3	0.26	0.23	0.14	0.30	-0.01	-0.02	-0.02	0.00
		5	0.26	0.25	0.16	0.30	0.01	0.02	0.03	0.00
		10	0.29	0.27	0.23	0.31	0.00	0.03	0.02	0.00
	15	3	0.21	0.21	0.17	0.32	0.03	0.01	0.06	0.00
		5	0.24	0.21	0.21	0.34	0.00	0.00	-0.01	0.00
		10	0.23	0.24	0.26	0.35	-0.01	0.02	-0.01	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

表 6.2: 既存モデルのパラメータ・リカバリ実験の結果

J	R	T	RMSE				BIAS			
			θ	β_r	π_r	d_k	θ	β_r	π_r	d_k
60	10	3	0.32	0.41	0.50	0.08	-0.06	0.00	-0.07	0.00
		5	0.34	0.57	0.67	0.10	0.12	0.00	0.22	0.00
		10	0.35	0.47	0.64	0.10	0.06	0.00	0.05	0.00
	15	3	0.28	0.58	0.57	0.07	-0.06	0.00	-0.03	0.00
		5	0.25	0.50	0.57	0.08	-0.03	0.00	-0.04	0.00
		10	0.28	0.47	0.72	0.08	0.04	0.00	0.01	0.00
	90	3	0.32	0.50	0.57	0.06	-0.07	0.00	-0.03	0.00
		5	0.31	0.47	0.55	0.07	-0.03	0.00	-0.11	0.00
		10	0.32	0.52	0.61	0.10	-0.02	0.00	-0.03	0.00
120	10	3	0.31	0.54	0.59	0.07	0.10	0.00	0.07	0.00
		5	0.33	0.45	0.53	0.07	0.05	0.00	0.08	0.00
		10	0.32	0.45	0.61	0.08	-0.02	0.00	-0.08	0.00
	15	3	0.27	0.48	0.46	0.06	-0.03	0.00	-0.11	0.00
		5	0.27	0.46	0.58	0.07	-0.01	0.00	0.01	0.00
		10	0.28	0.51	0.64	0.07	-0.02	0.00	-0.05	0.00
	Avg.		0.30	0.51	0.60	0.07	0.00	0.00	-0.02	0.00

図 6.1: β_{rt} の推定例

第 7 章

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。これらの 6 人を評価者 A, B, C, D, E, F とすると、それぞれに与えた指示は以下の通りである。

評価者 A 日ごとにだんだん厳しくなるように採点する

評価者 B 日ごとにだんだん優しくなるように採点する

評価者 C 日によって厳しさを変えて採点する。具体的には、2 日目は 1 日目より厳しく、3 日目は 1 日目より甘く、4 日目は 2 日目よりさらに厳しくなるように採点する。

評価者 D 使用する得点に偏りを持たせ、2, 3, 4 点を多く使用するように採点する

評価者 E 使用する得点に偏りを持たせ、1, 3, 5 点を多く使用するように採点する

評価者 F 採点基準を用いずに採点する

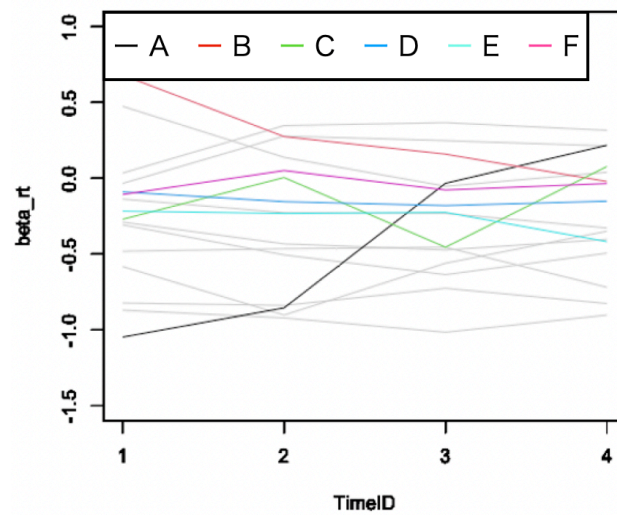
本研究では、このように収集したデータに対して提案モデルを適用する。

7.1 評価者特性

実データから推定された評価者パラメータを表 7.1 に示す。また、 β_{rt} の推定結果をグラフにしたものを図 7.1 に示す。縦軸が β_{rt} 、横軸が時間区分である。色がついているものは、前述した指示を与えた評価者のデータであり、黒が評価者 A、赤が評価者 B、緑が評価者 C、青が評価者 D、水色が評価者 E、紫が評価者 F である。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示していることを推定できている。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。例えば $r = 2$ の評価者は時間区分 1 から 2 に変化するとき厳しさが大幅に上昇し、そのあとは緩やかに低下していることがわかる。また、評価者 D と評価者 E の評価者パラメータから作成した ICC を表 7.2 に示す。これを見ると、どちらも多く使うように指示した得点を多く使用していることがデータから推定できていることがわかる。また、 α_r の値を評価者ごとに比べると、評価者ごとに一貫性に差があることがわかる。さらに、評価者ごとにステップパラメータ d_{rk} の値に差があることが読み取れる。

表 7.1: パラメータ推定値

r	α_r	β_{r1}	β_{r2}	β_{r3}	β_{r4}	d_{r1}	d_{r2}	d_{r3}	d_{r4}	d_{r5}
1	0.71	-0.48	-0.46	-0.46	-0.72	0.00	-2.13	-0.23	0.63	1.73
2	1.07	-0.03	0.28	0.25	0.21	0.00	-0.86	-0.51	0.54	0.83
3	1.57	-0.87	-0.92	-1.02	-0.90	0.00	-0.92	-1.26	0.39	1.79
4	0.97	-0.58	-0.90	-0.56	-0.35	0.00	-1.62	-1.27	0.77	2.12
5	1.14	-0.14	-0.23	-0.23	-0.33	0.00	-2.30	-0.28	0.80	1.78
6	0.93	0.47	0.14	-0.05	0.04	0.00	-1.63	-0.64	0.48	1.80
7	0.82	-0.29	-0.43	-0.47	-0.40	0.00	-1.18	-0.09	0.71	0.56
8	1.32	0.03	0.35	0.37	0.32	0.00	-1.41	-0.65	0.75	1.31
9	0.77	-0.82	-0.84	-0.73	-0.83	0.00	-1.00	-0.45	0.47	0.98
10	1.00	-0.31	-0.51	-0.64	-0.49	0.00	-1.52	-0.13	0.38	1.26
11(評価者 A)	1.16	-1.05	-0.85	-0.03	0.22	0.00	-1.66	-0.66	0.92	1.40
12(評価者 B)	1.23	0.68	0.28	0.16	-0.02	0.00	-0.96	-0.10	0.21	0.84
13(評価者 C)	1.02	-0.27	0.01	-0.45	0.08	0.00	-1.08	-0.70	0.73	1.05
14(評価者 D)	0.68	-0.09	-0.15	-0.18	-0.15	0.00	-1.69	-0.71	0.42	1.97
15(評価者 E)	1.16	-0.22	-0.23	-0.22	-0.42	0.00	0.19	-1.20	1.37	-0.36
16(評価者 F)	0.87	-0.11	0.05	-0.08	-0.03	0.00	-0.90	-0.46	0.06	1.29

図 7.1: β_{rt} の推定結果

以上のことから、提案モデルでは、時間区分ごとの評価者の厳しさの変化と、評価者ごとの一貫性、得点の使用傾向の違いを推定することができているとわかる。

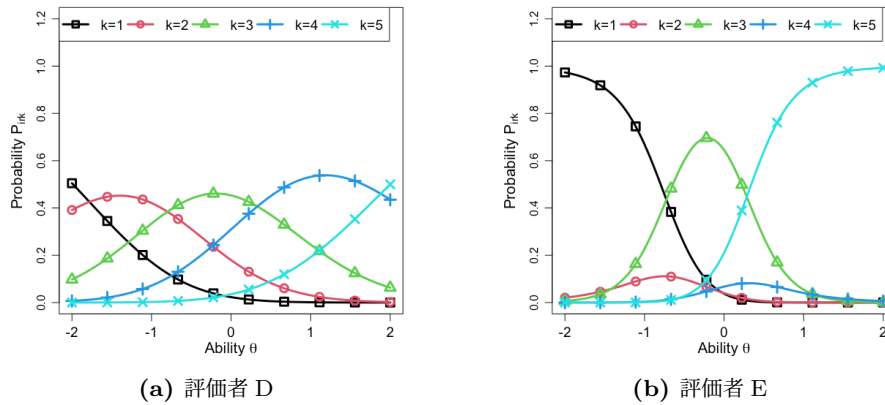


図 7.2: 指示を与えた評価者の ICC

7.2 情報量基準におけるモデル比較

本節では、提案モデルの性能を評価するために、4. で紹介した既存モデルとの情報量規準による性能比較を行う。

また、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 提案モデルにおいて d_{rk} を d_k に入れ替えたモデル

比較モデル 2 提案モデルにおいて β_{rt} を $\beta_r - \pi_r t$ に置き換えたモデル

比較モデル 3 提案モデルにおいて α_r を抜いたモデル

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 は汎化誤差 (まだ手に入れていないデータを予測した時の誤差) の近似であり、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 は周辺尤度の近似であり、データを生成した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。

今回は、全ての評価者データを使用した場合と、指示を与えた評価者を除外した場合に分けて情報量基準を求め、その結果をそれぞれ表 7.2, 7.3 に示す。

表 7.2 と 7.3 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。また、提案モデルと比較モデル 1 の結果を比較すると、提案モデルの方が WAIC の値が小さいことから、ステップパラメータを評価者に依存させたことによって、提案モデルの性能が向上していることがわかる。加えて、提案モデルと比較モデル 2 の結果を比較すると、提案モデルの方が WAIC の値が小さいことから、時間区分ごとにおける評価者特性を考慮できるようにしたこと、提案モデルの性能が向上していることがわかる。さらに、提案モデルと比較モデル 3 についても、提案モデルの方が WAIC の値が小さいことから、評価者の一貫性を考慮したことによって、提案モデルの性能が向上していることがわかる。以上より、提案モデルでは、既存モデルよりも時間区

表 7.2: モデル比較の結果 (全評価者のデータ)

WAIC	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
	5361.581	5027.951	5225.050	5104.463	5032.362
WBIC	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
	3071.706	3033.753	3038.600	3056.349	3028.649

表 7.3: モデル比較の結果 (指示ありの評価者を除外したデータ)

WAIC	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
	3279.444	3134.700	3190.802	3154.331	3137.137
WBIC	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
	1927.215	1900.211	1863.710	1924.210	1910.699

分における評価者特性の依存関係に加え，評価者の一貫性や時間区分ごとにおける特性を表現できるようになり，データへの当てはまりが最も高くなったと解釈できる．

次に，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，比較モデル 2 よりも性能が高いため，時間区分ごとの厳しさ β_{rt} を導入したことの有効性は確認できる．

第 8 章

むすび

本研究では、時間区分ごとの評価者の厳しさを推定できる新しい IRT モデルを提案した。また、提案モデルのパラメータ推定手法として、Stan を用いた No-U-turn sampler による MCMC アルゴリズムを提案し、シミュレーション実験によるアルゴリズムの妥当性を示した。更に、情報量規準に基づくモデル選択のアプローチを提案モデルに適用することで、能力尺度の最適な次元数を推定できることを、シミュレーション実験により示した。また、シミュレーション実験と実データを用いた実験では、提案モデルが評価者の時間区分ごとの厳しさを考慮した高精度な能力推定が実現できることを従来のモデルとの比較により示した。

なお、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、今後の課題としては、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。

付録 A

stan コード

```

data{
  int <lower=0> J;//n_examinee
  int <lower=0> R;//n_rater
  int <lower=2> K;//n_score
  int <lower=0> T;//n_time
  int <lower=0> N;//n_samples
  int <lower=1, upper=J> ExamineeID [N];
  int <lower=1, upper=R> RaterID [N];
  int <lower=1, upper=K> X [N]; //Score
  int <lower=1, upper=T> TimeID[N];
}
transformed data{
  vector[K] c = cumulative_sum(rep_vector(1, K)) - 1;
}
parameters {
  vector[J] theta;
  real<lower=0> alpha_r [R-1];
  matrix[R,T] beta_rt;
  vector[K-2] beta_rk[R];
  real<lower=0> sigma_beta_rt;
}
transformed parameters{
  vector[K-1] category_est[R];
  vector[K] category_prm[R];
  real<lower=0> trans_alpha_r[R];
  trans_alpha_r[1] = 1.0 / prod(alpha_r);
  trans_alpha_r[2:R] = alpha_r;
  for(r in 1:R){
    category_est[r, 1:(K-2)] = beta_rk[r];
    category_est[r, K-1] = -1*sum(beta_rk[r]);
    category_prm[r] = cumulative_sum(append_row(0, category_est[r]));
  }
}
model{
  theta ~ normal(0, 1);
  trans_alpha_r ~ lognormal(0.0, 0.4);
  sigma_beta_rt ~ lognormal(-3, 1);
  for (r in 1:R){
    beta_rt[r,1] ~ normal(0, 1);
    for (t in 2:T){
      beta_rt[r,t] ~ normal(beta_rt[r,t-1], sigma_beta_rt);
    }
  }
}

```

```
}
for (p in 1:R) category_est [p,] ~ normal(0, 1);
for (n in 1:N){
  X[n] ~ categorical_logit(1.7*trans_alpha_r[RaterID[n]]*(c*(
    theta[ExamineeID[n]]-beta_rt[RaterID[n],TimeID[n]])
    -category_prm[RaterID[n]]));
}
}

generated quantities {
  vector[N] log_lik;
  for (n in 1:N){
    log_lik[n] = categorical_logit_log(X[n], 1.7*trans_alpha_r[RaterID[n]]*(c*(
      theta[ExamineeID[n]]-beta_rt[RaterID[n],TimeID[n]])
      -category_prm[RaterID[n]]));
  }
}
```

参考文献

- [1] 豊田秀樹, 項目反応理論: 入門編, 朝倉書店, Reading, Massachusetts, 2007.
- [2] F.M. Lord, “Applications of item response theory to practical testing problems,” 1980.
- [3] G. Rasch, “Probabilistic models for some intelligence and attainment tests,” 1993.
- [4] J.M. Linacre, “Many-faceted rasch measurement,” 1990.
- [5] L.T. DeCarlo, Y. Kim, and M.S. Johnson, “A hierarchical rater model for constructed responses, with a signal detection rater model,” *Journal of Educational Measurement*, vol.48, no.3, pp.333–356, 2011.
- [6] Y. Lu and X. Wang, “A hierarchical Bayesian framework for item response theory models with applications in ideal point estimation,” 2006.
- [7] R.J. Patz and B.W. Junker, “Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses,” *Journal of educational and behavioral statistics*, vol.24, no.4, pp.342–366, 1999.
- [8] T.M. McNaughton, “Rater stability in a high-stakes performance assessment: A longitudinal investigation,” 2018.
- [9] M. Uto and M. Ueno, “Item response theory for peer assessment,” *IEEE Transactions on Learning Technologies*, vol.9, no.2, pp.157–170, 2016.
- [10] M. Matteucci and L. Stracqualursi, “Student assessment via graded response model,” *Statistica*, vol.66, no.4, pp.435–447, 2006.
- [11] L.T. DeCarlo, “A model of rater behavior in essay grading based on signal detection theory,” *Journal of Educational Measurement*, vol.42, no.1, pp.53–76, 2005.
- [12] 村木英治, 項目反応理論, 朝倉書店, 2011.
- [13] 豊田秀樹, 項目反応理論: 中級編, 朝倉書店, 2013.
- [14] G.N. Masters, “A rasch model for partial credit scoring,” *Psychometrika*, vol.47, no.2, pp.149–174, 1982.
- [15] E. Muraki, “A generalized partial credit model,” *Handbook of modern item response theory*, pp.153–164, 1997.
- [16] D. Andrich, “A rating formulation for ordered response categories,” *Psychometrika*, vol.43, no.4, pp.561–573, 1978.
- [17] C.M. Myford and E.W. Wolfe, “Detecting and measuring rater effects using many-facet rasch measurement: Part i,” *Journal of applied measurement*, vol.4, no.4, pp.386–422, 2003.
- [18] 宇佐美慧, “論述式テストの運用における測定論的問題とその対処,” *日本テスト学会誌*, vol.9,

- no.1, pp.145–164, 2013.
- [19] M.Uto and M.Ueno, “A multidimensional generalized many - facet rasch model for rubric - based performance assessment,” 2021.
- [20] E.W. Wolfe, B.C. Moulder, and C. Myford, Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model, *Journal of Applied Measurement*, 2001.
- [21] S.W. Raudenbush and A.S. Bryk, *Hierarchical linear models: Applications and data analysis methods*, vol.1, Sage Publications, 2002.
- [22] J.-P. Fox, “Bayesian item response modeling: Theory and applications,” 2010.
- [23] 宇佐美慧, “採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル,” *教育心理学研究*, vol.58, no.2, pp.163–175, 2010.
- [24] M.D. Hoffman and A. Gelman, “The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo,” *J. Mach. Learn. Res.*, vol.15, no.1, pp.1593–1623, 2014.
- [25] M. Girolami and B. Calderhead, “Riemann manifold langevin and hamiltonian monte carlo methods,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol.73, pp.123–214, 2011.
- [26] J. Rosenthal, S. Brooks, A. Gelman, G. Jones, and X. Meng, “Handbook of markov chain monte carlo,” vol.2000, chapter Optimal Proposal Distributions and Adaptive, pp.93–112, CRC Press, Boca Raton, FL, 2011.
- [27] R.M. Neal, et al., “MCMC using hamiltonian dynamics,” *Handbook of markov chain monte carlo*, vol.2, no.11, p.2, 2011.
- [28] B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J.Guo, P. Li, and A. Riddell, “Stan: A probabilistic programming language,” *Journal of statisticalsoftware*, vol.76, no.1, pp.1–32, 2017.
- [29] Y. Luo and H. Jiao, “Using the stan program for Bayesian item response theory,” *Educational and psychological measurement*, vol.78, pp.384–408, 2018.
- [30] Z. Jiang and R. Carter, “Using hamiltonian monte carlo to estimate the log-linear cognitive diagnosis model via stan,” *Behavior research methods*, vol.51, no.2, pp.651–662, 2019.
- [31] 松浦健太郎, *Stan と R でベイズ統計モデリング: Wonderful R 2, 第 2 巻*, 共立出版, 2017.
- [32] S.D. Team, et al., “Rstan: the r interface to stan. r package version 2.17. 3,” 2018.
- [33] M. Uto and M. Ueno, “Empirical comparison of item response theory models with rater’ s parameters,” 2018.
- [34] 八木嵩大, 宇都雅輝, “パフォーマンス評価における多次元項目反応モデル,” *電子情報通信学会論文誌 D*, vol.102, no.10, pp.708–720, 2019.
- [35] A. Gelman, D.B. Rubin, et al., “Inference from iterative simulation using multiple sequences,” *Statistical science*, vol.7, no.4, pp.457–472, 1992.
- [36] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, “Bayesian data analysis,” 2013.