令和３年度 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

電気通信大学　情報理工学域Ⅰ類
コンピュータサイエンスプログラム
学籍番号 1810519

林　真由

指導教員：宇都雅輝　准教授

2022年1月31日

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．一方で，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難しいという問題がある．この問題を解決するために，評価者特性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多数提案されている．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．このような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，評価者の厳しさの初期値と傾きを表す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，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．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を推定する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．

目次

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・テスティングの普及とともに，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，多数の受験者を長時間かけて採点するような場合には，採点の進行に伴い，評価者の特性が変化してしまう「評価者ドリフト(Rater Drift)」と呼ばれる問題が知られている．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・テスティングの普及とともに，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．

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

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．

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

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．

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

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．

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

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．

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

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

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．

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

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．

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

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．

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

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．

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

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．

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

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，$\alpha_r$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．

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

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，時間区分ごとの評価者パラメータを導入する．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．

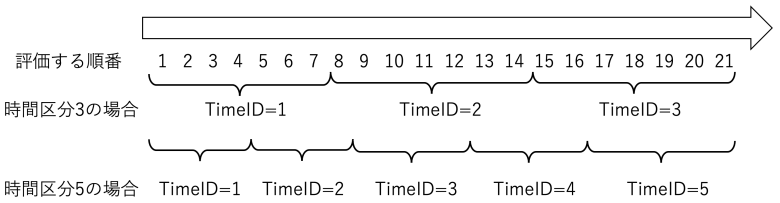


図4.1: 時間区分データのイメージと例

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．

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

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．

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

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変化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．

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

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

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，$\mathrm{N}(\mu, \sigma)$は平均μ，標準偏差σの正規分布を表し，$\mathrm{LN}(\mu, \sigma)$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}は，時間進行に伴う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．なお，σは0に近い値を取ることで，$\beta_{rt}(t > 1)$の事後分布が縮小し，パラメータ推定が安定すると期待できる．そこで，この事前知識を踏まえて，σの事前分布として，$\mathrm{LN}(-3, 0)$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．

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す

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

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なり，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．

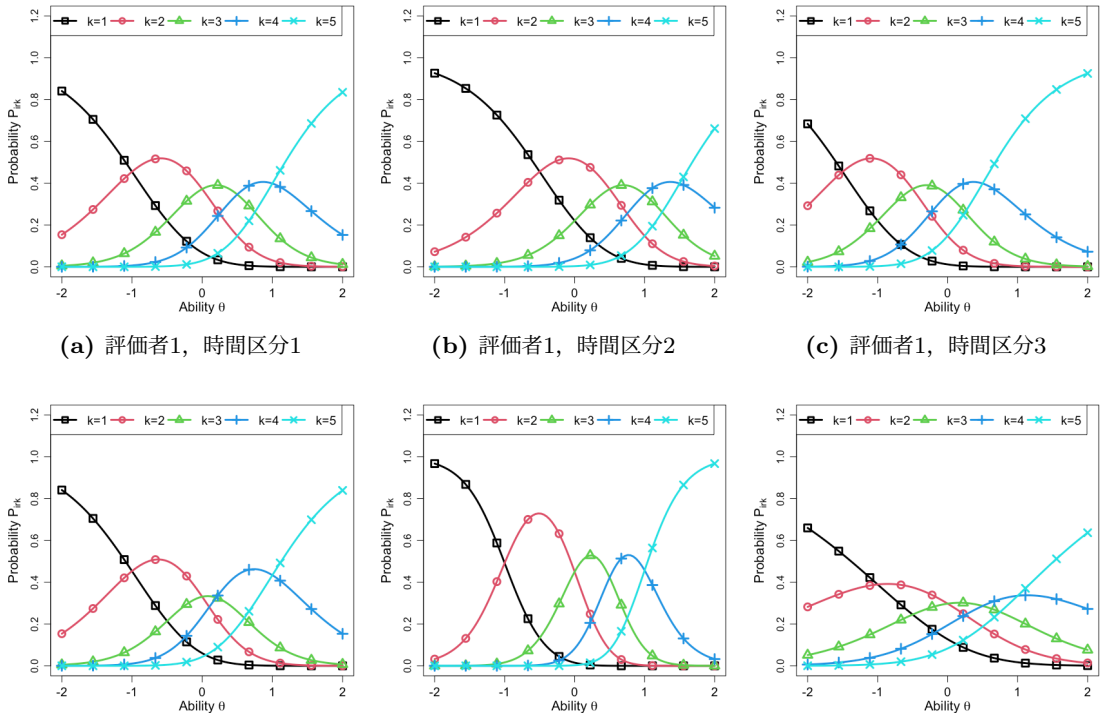
(c)は(a)と比較してβ_{rt}が小さくなっている．これは，時間区分1の時よりも時間区分3の時の方が評価者の厳しさが小さくなっているという意味である．(c)のICCでは，(a)と比べて全体的に左に移動していることが確認できる．これは，β_{rt}が小さくなっている時間区分では，良い評価を得るために要求される能力値が低くなっていることを示している．提案モデルでは，このように各時間区分における評価者の厳しさを，評価者ごとに表現する．

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$\beta_{rk+1} - d_{rk}$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，受検者の能力と相関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．



(a) 評価者1，時間区分1　**(b)** 評価者1，時間区分2　**(c)** 評価者1，時間区分3

(d) 評価者2，時間区分1　**(e)** 評価者3，時間区分1　**(f)** 評価者4，時間区分1

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$\mu$，標準偏差$\sigma$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．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，提案モデルにおける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が変わるごとに変化する．

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，$\delta_r$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，それぞれ評価者の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(-3~3)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制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$\hat{R}$[33, 34]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．

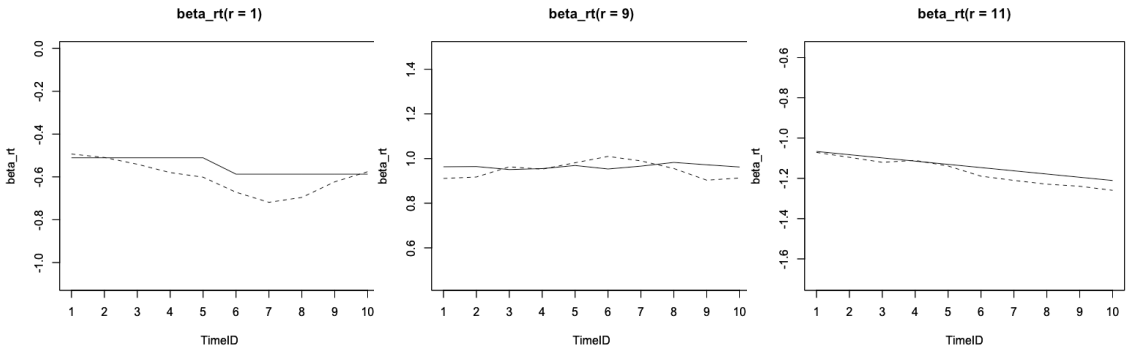
また，表6.2に既存モデルの実験結果を示す．表6.2から，提案モデルと比べるとRMSE，バイアスともにやや悪い傾向があるが，全体としては受検者数や評価者数の増加に伴う推定精度の改善傾向は確認でき，適切に推定されていると言える．また，収束判定指標の$\hat{R}$[33, 34]も全て1.1を下回っており，MCMCの収束も確認できた．以上の結果から，MCMCにより既存モデルのパラメータを適切に推定できることが確認できた．

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	10	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
	15	3	0.26	0.42	0.51	0.08	-0.05	0.00	-0.07	0.00
		5	0.31	0.78	0.69	0.08	0.00	0.00	-0.08	0.00
		10	0.28	0.53	0.59	0.08	0.04	0.00	-0.08	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



(a) パターン1　　**(b)** パターン2　　**(c)** パターン3

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に変化するときに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

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．

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．

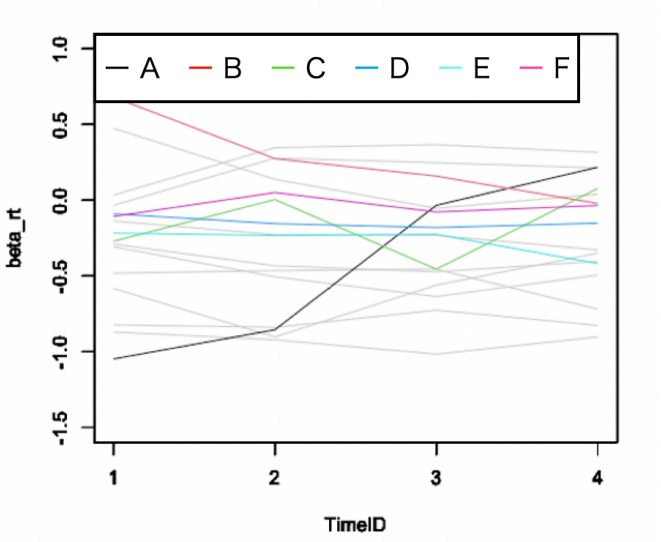
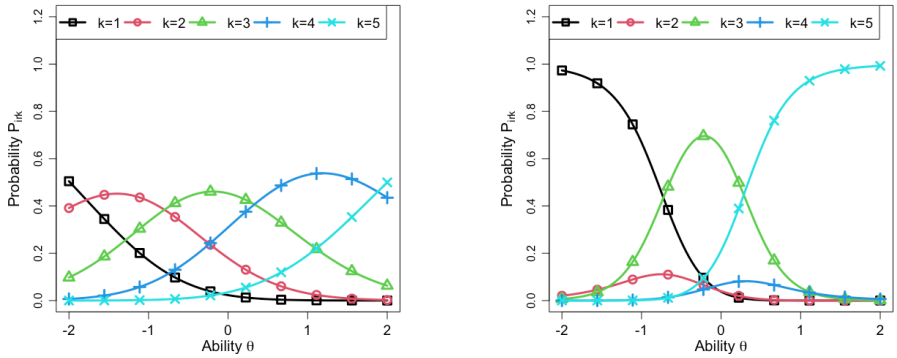


図7.1: β_{rt}の推定結果



(a) 評価者D　　**(b)** 評価者E

図7.2: 指示を与えた評価者のICC

今回は，全ての評価者データを使用して推定を行った場合と，指示を与えた評価者を除外して推定を行った場合に分けて情報量基準を求めた．全ての評価者データを使用した実験結果を表7.2に，指示を与えた評価者を抜いたデータを使用した実験結果を表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で，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

表7.2: WAICとWBACによるモデル比較の結果

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: WAICとWBACによるモデル比較の結果(指示なしの評価者)

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

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，表7.2と7.3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．

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．

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．

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 LearningTechnologies, 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 meth- ods,” 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] A. Gelman, D.B. Rubin, et al., “Inference from iterative simulation using multiple sequences,” Statistical science, vol.7, no.4, pp.457–472, 1992.

[34] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, “Bayesian data analysis,” 2013.