レイティングデータとテキスト情報を用いて受験者の能力を推定する項目反応トピックモデルの提案

An IRT model integrating supervised LDA that estimates writing ability using rating data and textual content

宇都雅輝 *
Masaki Uto

電気通信大学
University of Electro-Communications

Abstract: In various assessment contexts, essay writing tests have been widely used to measure higher order abilities of students. A persistent difficulty is that the ability measurement accuracy depends strongly on rater characteristics. To resolve this problem, many item response theory (IRT) models have been proposed that can estimate the abilities with consideration of the rater-effects. One remaining difficulty, however, is that measurement accuracy is reduced when few raters are assigned to each essay, which is a common situation in practical testing contexts. To resolve this problem, we propose a new rater-effect IRT model integrating a supervised topic model that can estimate the abilities from raters' gradings and textual content of written essays. We evaluate the effectiveness of the proposed model through experiments using actual data.

1 はじめに

近年，論理的思考力や問題解決力といった高次の能力を測定するニーズが高まっており，これを実現する手法の一つとして論述式テストの活用が注目されている．一般に論述式テストは，受験者に複数の課題を与え，それらに対する回答文を数名の評価者によって採点する形式で実施される．しかし，この場合，得られる評点が評価者や課題の特性（評価者の甘さ/厳しさや課題困難度など）に強く依存し，これが受験者の能力測定の精度低下を引き起こすことが問題とされてきた [1, 2, 3, 4, 5, 6]．この問題を解決する手法の一つとして，評価者と課題の特性パラメータを付与した項目反応モデルが近年多数提案されている (e.g., [4, 7, 5])．これらの項目反応モデルでは評価者と課題の特性を考慮して受験者の能力を推定できるため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能となる．

しかし，これらのモデルを用いても，個々の回答文を採点する評価者数が少なくなると高精度な能力測定は困難となる．一般に論述式テストの採点プロセスでは，評価者の負担や運用の時間的・経済的コストを軽減するために，各回答文に少数名の評価者を割り当てて採点を行わせることが多い [5, 8]．

本研究では，この問題を解決するために，評価者による評点データだけでなく，受験者が執筆した回答文の内容も能力測定に利用できる新たな項目反応モデルを提案する．提案モデルは，評価者と課題の特性を考慮した項目反応モデルとトピックモデルのひとつである潜在ディリクレ配分法 [9] を統合したモデルとして定式化する．具体的には，潜在ディリクレ配分法を用いて個々の回答文のトピック分布を推定し，そのトピック分布を項目反応モデルにおける受験者の能力推定値に反映させるようにモデル化を行う．トピック分布の能力値への反映には，トピック分布と任意の目的変数の関係をモデル化した教師ありトピックモデル [10] のアプローチを用いる．提案モデルの利点は次の通りである．1）評価者が与える評点データに加えて，回答文の内容的な特徴も考慮して能力推定がなされるため，回答文あたりの評価者が少ない場合の能力測定精度を改善できると期待できる．2）評点が与えられていない回答文の得点と，それらの回答文を執筆した受験者の能力を文章情報のみから推定することができる．

本論文では，実データ実験により提案モデルの有効性を示す．

*連絡先：電気通信大学大学院情報理工学研究科
〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: uto@ai.lab.uec.ac.jp

2 データ

本研究では，J 人の受験者 $\mathcal{J} = \{1, \cdots, J\}$ に I 個の論述課題 $\mathcal{I} = \{1, \cdots, I\}$ を与え，それらの回答文を R 人の評価者集団 $\mathcal{R} = \{1, \cdots, R\}$ が K 段階カテゴリ $\mathcal{K} = \{1, \cdots, K\}$ で採点する場合を考える．ここで，課題 $i \in \mathcal{I}$ に対する受験者 $j \in \mathcal{J}$ の回答文を e_{ij} で表し，回答文 e_{ij} に対する評価者 r の評点を U_{ijr} とすると，評点データは次式で定義できる．

$$\boldsymbol{U} = \{U_{ijr} \in \mathcal{K} \cup \{-1\} \mid i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\} \quad (1)$$

ここで，$U_{ijr} = -1$ は欠測データを表す．

また，回答文集合 $\boldsymbol{E} = \{e_{ij} \mid i \in \mathcal{I}, j \in \mathcal{J}\}$ に含まれる語彙集合を $\mathcal{V} = \{1, \cdots, V\}$ とすると，回答文 e_{ij} 内の単語系列は次式で定義できる．

$$\boldsymbol{W}_{ij} = \{W_{ijn} \in \mathcal{V} \mid n = \{1, \cdots, N_{ij}\}\} \quad (2)$$

ここで，W_{ijn} は回答文 e_{ij} 内の n 番目の単語を表し，N_{ij} は e_{ij} 内の単語数を表す．

本研究の目的は，これらのデータを用いて各受験者の能力を高精度に推定することである．このために本研究では項目反応理論とトピックモデルを用いる．

3 項目反応理論

項目反応理論（IRT: Item Response Theory）は数理モデルを用いたテスト理論のひとつである．IRT では，受験者のテスト項目への反応を，受験者の能力を表す潜在変数と項目の特性（困難度や識別力など）を表すパラメータで定義される確率モデルで表現する．このようなモデルを用いることで，IRT は，1) 異なる項目で構成されたテストを受験しても同一尺度上で能力を測定できる，2）個々の項目やテスト全体の能力測定精度を分析できる，3）欠測データの扱いが容易である，などの多くの利点を持つ．このような利点から，IRT は現代のテスト運用の基礎として，IT パスポート試験や医療系大学間共用試験などの大規模試験を含む，様々な評価場面で広く実用化されている．

一般的な項目反応モデルでは，テスト項目に対する受験者の反応や正誤答をデータとして扱うため，データは受験者 × 項目の 2 相データとなる．他方で，2 で定義したように，本研究で扱うデータは受験者 × 課題 × 評価者の 3 相データとなる．従来の項目反応モデルは，このような 3 相データに直接には適用できない．この問題を解決するために，項目反応モデルにおける項目特性パラメータを課題の特性パラメータとみなし，評価者の特性を表すパラメータを付与したモデルが近年多数提案されている [4, 7, 5].

本研究では，既存モデルの中で，評価者特性を最も柔軟に捉えることができる宇都・植野のモデル [6] を基礎モデルとして採用する．このモデルでは，課題 i に対する受験者 j の回答文に評価者 r が評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]} \quad (3)$$

ここで，θ_j は受験者 j の能力，α_i は課題 i の識別力，α_r は評価者 r の一貫性，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_{rk} は評価カテゴリ k に対する評価者 r の厳しさを表す．ただし，パラメータの識別性のために，$\sum_{i=1}^{I} \log \alpha_i = 0$，$\sum_{i=1}^{I} \log \beta_i = 0$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する．これらのモデルパラメータと能力値は，評点データ $\boldsymbol{U}$ から推定することができる．

1 で述べたように，このような項目反応モデルでは，受験者の能力を評価者や課題の特性の影響を取り除いて推定できるため，素点の合計や平均といった単純な得点化法より高精度な能力測定が可能となる [3, 4, 6, 7]. しかし，これらのモデルを用いても，個々の回答文を採点する評価者数が少なくなると，受験者あたりの評点データが減少するため，能力推定の精度が低下する．本研究のアイディアは，この問題を解決するために，受験者の能力 θ_j の推定に，評点データだけでなく回答文の内容も利用する点にある．本研究では，回答文の内容を扱う手法としてトピックモデルを用いる．

4 トピックモデル

トピックモデルは，文書集合が与えられたとき，個々の文書が複数の潜在的な話題（トピック）を持つと仮定し，それらのトピックの出現分布を文書ごとに推定する教師なし機械学習手法である．また，トピックモデルでは，各トピックに対して語彙の出現分布を推定するため，それらの語彙分布を解釈することで個々のトピックの意味を解釈することができる．代表的なトピックモデルとしては，潜在意味解析法（LSA: Latent Semantic Analysis）や確率的潜在意味解析法（PLSA: Probabilistic Latent Semantic Analysis），潜在ディリクレ配分法（LDA: Latent Dirichlet Allocation）[9] が知られている．LDA は LSA と PLSA の上位モデルであり，LSA や PLSA に比べて高精度なトピック推定が可能であることから，テキストを扱う様々なタスクで活用されている（e.g., [10, 11, 12, 13, 14]）．そこで，本研究では，トピックモデルとして LDA を利用する．

LDA では回答文 e_{ij} 内の各単語 W_{ijn} がどのトピックから生成されたかを示す潜在変数を導入する．ここで，単語 W_{ijn} に対応するトピックを $Z_{ijn} \in \mathcal{T} = \{1, \cdots, T\}$（$T$ はトピック数）で表し，回答文 e_{ij} におけるトピック t の生起確率を ψ_{ijt}，トピック t における語彙 v の生起確率を ϕ_{tv} で表す．このとき，LDA

では，各単語 W_{ijn} とトピック Z_{ijn} が以下の多項分布（$Multi(\cdot)$ と表記する）で表されるトピック分布と語彙分布に従って生起すると仮定する．

$$Z_{ijn} \sim Multi(\boldsymbol{\psi}_{ij}),\ W_{ijn} \sim Multi(\boldsymbol{\phi}_{z_{ijn}}) \quad (4)$$

ただし，$\boldsymbol{\psi}_{ij} = \{\psi_{ij1}, \cdots, \psi_{ijT}\}$，$\boldsymbol{\phi}_t = \{\phi_{t1}, \cdots, \phi_{tV}\}$．

また，各分布のパラメータ $\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ は多項分布の共役事前分布であるディリクレ分布（$Dir(\cdot)$ と表記する）に従うと仮定する．ここで，γ と η を $\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ のディリクレ事前分布のパラメータとすると，$\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ は以下の式に従って生成すると仮定される．

$$\boldsymbol{\psi}_{ij} \sim Dir(\gamma),\ \boldsymbol{\phi}_t \sim Dir(\eta) \quad (5)$$

LDA によって推定されるトピック分布 $\boldsymbol{\psi}_{ij}$ は，回答文 e_{ij} の内容的な特徴を T 次元のベクトルで表現したものと解釈できる．近年では，このように文書ごとに推定されるトピック分布を他の変数の予測に利用する教師ありトピックモデル [10] と呼ばれる手法が提案されている．本研究では，トピック分布を受験者の能力値に反映させるために教師ありトピックモデルのアプローチを用いる．

5 教師ありトピックモデル

一般に，教師ありトピックモデルでは，個々の文書 e_{ij} に対応する任意の目的変数 y_{ij} を，その文書のトピック情報を説明変数とする回帰モデルによって予測するようにモデル化する．回帰モデルには様々なモデルが利用できるが，最も一般的な正規回帰モデルを想定し，変数 y_{ij} が実数値をとると仮定すると，y_{ij} の生起確率は以下のように定義される．

$$y_{ij} \sim N(\boldsymbol{\omega}^{\mathrm{T}} \bar{\boldsymbol{Z}}_{ij}, \sigma_0^2) \quad (6)$$

ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を表し，$\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_T\}$ は目的変数に対する各トピックの重み集合を表す．σ_0^2 は目的変数の分散を表すハイパーパラメータである．また，$\bar{\boldsymbol{Z}}_{ij} = \{\bar{Z}_{ij1}, \cdots, \bar{Z}_{ijT}\}$ であり，$\bar{Z}_{ijt} \in \bar{\boldsymbol{Z}}_{ij}$ は次式で定義される．

$$\bar{Z}_{ijt} = \frac{\sum_{n=1}^{N_{ij}} \delta(Z_{ijn}, t)}{N_{ij}} \quad (7)$$

$\delta(a, b)$ は二つの値 a と b が一致するとき 1，そうでないとき 0 をとる関数とする．

教師ありトピックモデルは，個々の文書を T 次元のトピック分布パラメータで表現し，それを用いて目的変数に回帰するモデルとみなせる．教師ありトピックモデルでは，各文書の内容的な意味を考慮した予測が可能となるため，単語の出現頻度ベクトルを用いた単純な回帰モデルと比べて，高い予測精度が期待できることが報告されている [10, 12, 15]．このような利点から，教師ありトピックモデルのアプローチは，テキスト情報を予測に活用する様々な応用問題に適用され，その有効性が示されてきた．本研究でも，教師ありトピックモデルのアプローチを用いて，トピック分布を IRT モデルにおける受験者の能力推定値に反映させる．

6 提案手法

提案モデルでは，IRT における受験者の能力値 θ_j が，その受験者の回答文のトピック分布に依存すると考えることで，文章情報を能力値に反映する．具体的には，式（3）における能力 θ_j の分布として次式を考える．

$$\theta_j \sim N(\boldsymbol{\omega}^{\mathrm{T}} \bar{\boldsymbol{Z}}_j, \sigma_0^2) \quad (8)$$

ここで，$\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_T\}$ は能力推定値に対する各トピックの重みを表す．また，$\bar{\boldsymbol{Z}}_j = \{\bar{Z}_{j1}, \cdots, \bar{Z}_{jT}\}$ を表し，$\bar{Z}_{jt} \in \bar{\boldsymbol{Z}}_j$ は次式で定義される．

$$\bar{Z}_{jt} = \frac{\sum_{i=1}^{I} \sum_{n=1}^{N_{ij}} \delta(Z_{ijn}, t)}{\sum_{i=1}^{I} N_{ij}} \quad (9)$$

本研究の条件では，各受験者が複数の回答文を有するのに対し，目的変数は受験者ごとに一つのみ推定される能力値 θ_j となるため，通常の教師ありトピックモデルとは異なり，$\bar{Z}_{jt}$ が複数回答文のトピック情報を累積した形で定義されている点に注意されたい．また，式（8）中の σ_0^2 は能力値の分散を表す．IRT では，能力値に標準正規分布を仮定することが一般的であるため，本研究でも $\sigma_0^2 = 1.0$ を用いる．

式 (8) から明らかなように，提案モデルでは，文章のトピック分布から推定される能力値を，項目反応モデルにおける能力推定値 θ_j の事前分布として反映している．このとき，トピック分布と能力値の関係は，式 (8) の重み $\boldsymbol{\omega}$ によって学習される．これにより提案モデルでは，文章の内容的な特徴を能力推定に反映できるため，評点データのみを利用する IRT に比べて能力測定精度が改善されると期待できる．また，提案モデルでは，語彙分布と評価者特性，課題特性および重みのパラメータが既知であれば，評点データが与えられていない受験者の能力を，文章情報のみを用いて推定することができる．さらに，そのように推定された能力値を所与として回答文の期待得点を求めることで未採点回答文の自動評価も可能である．これらの具体的な手順は 6.2 節で述べる．

6.1 パラメータ推定

IRT における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，式（3）のような複雑な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）アルゴリズムを用いた期待事後確率 (EAP: Expected A Posteriori) 推定法が一般に高精度である．また，LDAのパラメータ推定においては，変分ベイズ法を用いたEAP法とMCMCを用いたEAP法が一般的である．MCMCは変分ベイズ法に比べて計算効率は劣るものの，実装が容易であり推定精度も高いことが知られている．

IRTにおけるMCMCアルゴリズムとしては，メトロポリス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が一般的であり，LDAでは周辺化ギブスサンプリングを用いたアルゴリズムが一般に採用されている．周辺化ギブスサンプリングは，特定のパラメータ集合を周辺化することでMCMCの推定効率を高めることができる手法であり，提案モデルでもLDAと同様に利用できる．以上より，本研究では，提案モデルのパラメータ推定アルゴリズムとして，メトロポリスヘイスティングスと周辺化ギブスサンプリングを組み合わせたMCMCアルゴリズムを用いる．本アルゴリズムでは，トピック分布と語彙分布のパラメータである $\boldsymbol{\psi} = \{\boldsymbol{\psi}_{ij} | i \in \mathcal{I}, j \in \mathcal{J}\}$ と $\boldsymbol{\phi} = \{\boldsymbol{\phi}_t | t \in \mathcal{T}\}$ を周辺化し，トピック $\boldsymbol{Z} = \{Z_{ijn} | i \in \mathcal{I}, j \in \mathcal{J}, n \in \{1, \cdots, N_{ij}\}\}$ とIRTのモデルパラメータ $\boldsymbol{\xi} = \{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r, \boldsymbol{d}, \boldsymbol{\theta}\}$，重みベクトル $\boldsymbol{\omega}$ を，それぞれの条件付き事後分布からサンプリングする．ここで，$\boldsymbol{\alpha}_i = \{\log \alpha_{i=1}, \cdots, \log \alpha_{i=I}\}$，$\boldsymbol{\beta}_i = \{\beta_{i=1}, \cdots, \beta_{i=I}\}$，$\boldsymbol{\alpha}_r = \{\log \alpha_{r=1}, \cdots, \log \alpha_{r=R}\}$，$\boldsymbol{\beta}_r = \{\beta_{r=1}, \cdots, \beta_{r=R}\}$，$\boldsymbol{d} = \{d_{11}, \cdots, d_{RK}\}$，$\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_J\}$ とする．アルゴリズムの詳細については紙面の都合上割愛する．

6.2 文章データのみを用いた能力値推定と得点予測

上述の通り，提案モデルでは，語彙分布と評価者特性，課題特性および重みのパラメータが既知であれば，評点データが与えられていない受験者の能力を文章情報のみから推定することができる．具体的には，上記のMCMCアルゴリズムにおけるトピック Z_{ijn} のサンプリング式を，語彙分布と評価者特性，課題特性および重みのパラメータを所与とした条件付き事後分布に変更し，評価者特性と課題特性および重みのパラメータについては更新を行わないようにしたアルゴリズムで推定できる．

また，提案モデルでは，このように推定された能力値を所与として未採点回答の期待得点を求めることも可能である．具体的には，文章 e_{ij} の期待得点 $\hat{U}_{ij}$ は，事前に推定された評価者・課題の特性パラメータを所与として次式で求められる．

$$\hat{U}_{ij} = \sum_{r=1}^{R} \frac{1}{R} \sum_{k=1}^{K} k \cdot P_{ijrk} \tag{10}$$

7 評価実験

ここでは，実データ実験を通して提案モデルの有効性を評価する．本実験で利用する実データは次の被験者実験により収集した．

34名の大学生と大学院生に対して，4つの論述式課題を行わせ，各課題に対して提出された回答文を10名の評価者に採点させた．本実験で利用した論述式課題は，National Assessment of Educational Progress（NAEP）の2002年と2007年で出題された課題を日本語に翻訳したものであり，専門知識や特別な事前知識を必要としない内容となっている．また，評価者による採点は，NAEP grade 12で使用されたルーブリックを日本語に訳して作成した5段階カテゴリの評価基準を用いて行わせた．執筆された回答文の文字数は，平均が600.41，標準偏差が104.41であった．

7.1 能力推定精度の評価

本節では，提案モデルによる能力測定精度の評価を行う．このために，トピック数 T を $[1, 15]$ の区間で変化させながら，次の実験を行った．

1. 実データを用いてMCMCによるパラメータ推定を行なった．ただし，$T = 1$ のときには $\omega_1 = 0$ と固定し，ω_1 の推定は行わなかった．パラメータの事前分布とハイパーパラメータは先行研究の設定 [13, 6] に合わせて次の通りとした．$\log \alpha_i \sim N(0.1, 0.4)$，$\log \alpha_r \sim N(0.0, 0.5)$，$\beta_i, \beta_r, d_{rk}, \omega_t \sim N(0.0, 1.0)$，$\eta = 1/T$，$\gamma = 1/VT$．また，回答文集合から抽出する語彙の集合としては，ストップワードを除去した名詞，動詞，形容詞，接続詞，副詞を用いた．ストップワードの判定基準は，1）全回答文のうち2つ以下の回答文でしか利用されていない，2）全回答文の半分以上の回答文で利用されている，とした．結果として，語彙数は201となった．

2. 各受験者に $n \in \{1, 2\}$ 名の評価者をランダムに割り当て，評価者が割り当てられていない回答文の評点データを欠測させた．

3. 手順2で作成された欠測データを用いて，各学習者の能力値をMCMCにより再推定した．推定は，語彙分布と評価者特性，課題特性および重みのパラメータを所与として，6.2節の方法で行なった．

4. 手順 3 で推定された能力値と手順 1 で推定された能力値との平均平方二乗誤差（RMSE: Root Mean Square Error）を計算した．

5. 手順 2〜4 を 10 回繰り返し，RMSE の平均を求めた．

実験結果を図 1 に示す．図の横軸はトピック数を表し，縦軸は RMSE の値を表す．また，図中の One Rater と Two Raters のプロットが，それぞれ評価者が 1 名と 2 名のときの結果を表す．なお，$T = 1$ の提案モデルは，式（3）で与えられる従来の IRT モデルと一致する点に注意されたい．

実験結果から，従来モデルに対応する $T = 1$ の場合に比べて，提案モデルでは RMSE が大幅に低下していることがわかる．これは提案モデルが，回答文の内容的な特徴を能力測定値に適切に反映できたためと考えられる．また，提案モデルでは，トピック数が 4 までは単調に RMSE が低下し，以降では概ね同程度の性能を示している．概ね性能が収束したとみられるトピック数 $T \geq 4$ の提案モデルと従来モデルの性能を比較すると，提案モデルにおける評価者 1 名のときの誤差が，従来モデルにおける評価者 2 名のときの誤差と同程度となっている．これは，提案モデルでは，文章情報を利用したことで，従来モデルにおいて評価者を 1 名追加した場合と同程度の能力測定精度の改善が達成できたことを示している．

以上の実験結果から，対象物あたりの評価者数が少ないとき，提案モデルが能力測定精度の改善に有効であることが示された．

7.2 文章情報のみを用いた能力測定精度

ここでは，評点データが与えられていない受験者の能力を文章情報のみから推定した場合の能力測定精度について評価する．このために，トピック数 T を $[1, 15]$ の区間で変化させながら次の手順の実験を行なった．

1. 7.1 節の実験手順 1 と同様に，実データを用いて MCMC によるパラメータ推定を行なった．

2. 評点データを全て欠測させ，手順 1 で推定された語彙分布と評価者特性，課題特性および重みのパラメータを所与として，6.2 節の方法で各受験者の能力を再推定した．この手順は，受験者の能力を文章情報のみから推定していることに対応する．

3. 手順 1 で推定された能力値と手順 2 で推定された能力値の RMSE を計算した．

実験結果を図 1 の「No Raters」のプロットとして示した．従来モデルに対応する $T = 1$ では，評点データも文章情報も能力推定に利用できないため，能力測定誤差が著しく大きくなっている．他方で，提案モデルを利用した場合（$T > 1$ の場合）には，精度が大幅に改善していることがわかる．また，前節の実験と同様に，トピック数 $T = 4$ までは単調に RMSE が減少し，以降は概ね同程度の性能を示している．さらに，トピック数 $T \geq 4$ の提案モデルでは，評点データを利用していないにも関わらず，従来モデルにおいて評価者 1 名の評点データを利用した場合を上回る能力測定精度を達成していることがわかる．本実験結果から，提案モデルでは，評点データが与えられていない場合でも，従来モデルを用いて評価者 1 名の評点データから推定する場合と同程度の能力測定が実現できることが示された．

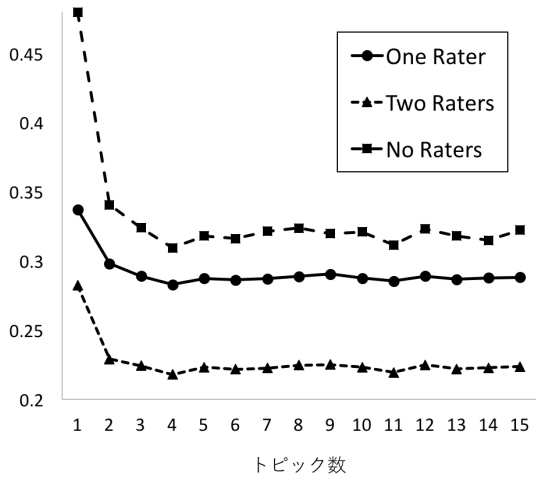

図 1: 能力推定誤差の評価結果

7.3 未採点回答の得点予測精度

本節では，提案モデルを用いた未採点回答の得点予測の性能評価を行う．このために，トピック数 T を $[1, 15]$ の区間で変化させながら，次の手順で実験を行なった．

1. 7.1 節の実験手順 1 と同様に，実データを用いて MCMC によるパラメータ推定を行なった．

2. 前節の実験手順 2 と同様に，評点データを全て欠測させたあと，手順 1 で推定された語彙分布と評価者特性，課題特性および重みのパラメータを所与として，6.2 節の方法で各受験者の能力を推定した．

3. 手順 2 で求めた能力推定値と手順 1 で得られた評価者と課題パラメータを用いて期待得点 $\hat{U}_{ij}$ を式（10）を用いて求め，期待得点 $\hat{U}_{ij}$ と完全データを用いて計算した観測平均得点 $U_{ij} = \sum_r U_{ijr}/R$ との RMSE を求めた．

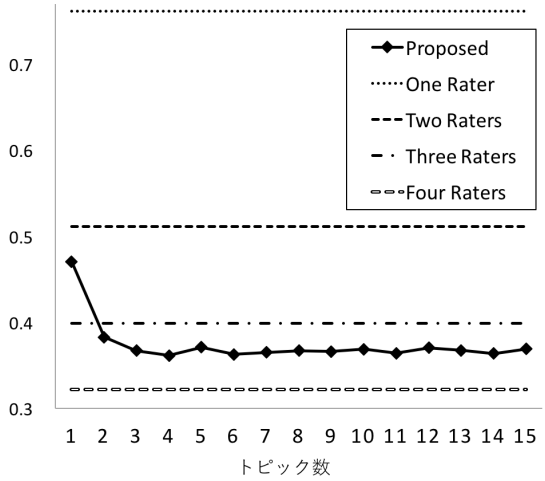

図 2: 評点予測誤差の評価結果

4. 比較のために，各回答文に $n \in \{1, \cdots, 5\}$ 名の評価者をランダムに割り当て，割り当てた評価者の評点データから求めた各回答文の平均得点と，完全データから求めた観測平均得点 U_{ij} との RMSE を計算した．この手順は評価者の割り当てを変えながら 10 回繰り返し，RMSE の平均値を求めた．

結果を図 2 に示す．図の横軸はトピック数を表し，縦軸は RMSE の値を表す．また，図 2 では，実線のプロット（「Proposed」と表記）が提案モデルで予測した得点と完全データから求めた観測平均得点の誤差を表し，破線（「n Rater(s)」と表記）が n 名の評価者のデータのみで求めた平均得点と完全データから求めた観測平均得点の誤差を表す．

図 2 から，これまでの実験と類似した傾向として，以下の結果が読み取れる．1）従来モデルに対応する $T = 1$ では予測誤差が著しく大きい．2）提案モデルを利用した場合には精度が大幅に改善する．3）トピック数 $T = 4$ までは誤差が単調に減少し，以降は概ね同程度の性能を示す．

さらに，提案モデルによる予測得点の精度を評価者 n 名の平均得点を利用した場合の精度と比較すると，提案モデルでは，評価者 3 名の平均得点を上回る予測精度を達成したことが確認できる．この結果から，提案モデルは，未採点回答の得点予測としても妥当な結果を与えることが確認できた．

8 まとめ

本研究では，評価対象物あたりの評価者数が少ない場合に IRT による能力測定の精度が低下する問題を解決するために，受験者が執筆した回答文の内容を能力測定の補助情報として利用できる新たなモデルを提案した．また，提案モデルのパラメータ推定手法として MCMC アルゴリズムによるベイズ推定法を提案した．さらに，実データ実験により，提案モデルが能力測定の精度改善に有効であり，未採点の回答文を持つ受験者の能力推定とその回答文の得点予測についても妥当な結果を与えることを示した．

参考文献

[1] 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, pp.386–422, 2003.

[2] T. Eckes, “Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis,” Language Assessment Quarterly, vol.2, no.3, pp.197–221, 2005.

[3] M. Uto and M. Ueno, “Item response theory for peer assessment,” IEEE Transactions on Learning Technologies, vol.9, no.2, pp.157–170, 2016.

[4] 宇都雅輝，植野真臣，“パフォーマンス評価のため項目反応モデルの比較と展望,” 日本テスト学会誌，vol.12，no.1，pp.55–75，2016.

[5] T. Eckes, Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments, Peter Lang Pub. Inc., 2015.

[6] 宇都雅輝，植野真臣，“ピアアセスメントにおける異質評価者に頑健な項目反応理論,” 電子情報通信学会論文誌.D，vol.101，no.1，pp.211–224，2018.

[7] M. Uto and M. Ueno, “Empirical comparison of item response theory models with rater’s parameters,” Heliyon, Elsevier, vol.4, no.5, pp.1–32, 2018.

[8] 宇都雅輝，“評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンス・テストの等化精度,” 電子情報通信学会論文誌.D，2018.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[10] D.M. Blei and J.D. McAuliffe, “Supervised topic models,” Proceedings of the 20th International Conference on Neural Information Processing Systems, pp.121–128, 2007.

[11] X. Li, J. Ouyang, and X. Zhou, “Supervised topic models for multi-label classification,” Neurocomputing, vol.149, pp.811–819, 2015.

[12] S. Jameel, W. Lam, and L. Bing, “Supervised topic models with word order structure for document classification and retrieval learning,” Information Retrieval Journal, vol.18, no.4, pp.283–330, 2015.

[13] M. Uto, S. Louvigné, Y. Kato, T. Ishii, and Y. Miyazawa, “Diverse reports recommendation system based on latent dirichlet allocation,” Behaviormetrika, vol.44, no.2, pp.425–444, 2017.

[14] S. Louvigné, M. Uto, Y. Kato, and T. Ishii, “Social constructivist approach of motivation: social media messages recommendation system,” Behaviormetrika, vol.45, no.1, pp.133–155, 2018.

[15] F. Li, S. Wang, S. Liu, and M. Zhang, “SUIT: A supervised user-item based topic model for sentiment analysis,” Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, pp.1636–1642, 2014.