ユーザー・アイテムの応答から構成された確率構造を持つ不完全マトリクスからのユーザーとアイテムの評価法

ユーザー・アイテムの応答から構成された
確率構造を持つ不完全マトリクスからの
ユーザーとアイテムの評価法

作　村　建　紀

目次

第 1 章

はじめに

1.1 背景

現在の日本では，少子高齢化の加速や高等教育への進学率増加を背景に，受験者層の学習習熟度の多様化が進んでいる（図 1.1 参照）．同時に PC の普及と情報インフラの充実により，インターネットを通じたビデオ講義やオンラインによる学習支援システムを利用した，個人に合わせた学習スタイルが増加している．この新しい学習スタイルは、個人の習熟度をデータベース化することを容易にし，そのデータを学習あるいは評価に有効利用されるようになってきている．

情報端末を利用した学習スタイルは，個人レベルに合わせた学習を可能にするという意味で，受験者層の多様化に適応している反面，そのためには事前に項目が持つ特性を知っておく必要があり，これまでの配点方式による評価方法では対応できない事態を引き起こしている．これに伴い，データから項目の特性を見出し，受験者の習熟度をより効率良く評価する問題提出システムへの要求が高まっている．このような背景から，受験者の習熟度評価に適した e-learning システムの確立が急務となっている．

ここで問題となるのは，項目特性および受験者習熟度を決定づける評価手法として世界的に広く用いられている項目反応理論（IRT: item response theory）[9,10,27] が，日本においてはあまり普及していないという点である．従来の伝統的な配点方式の評価は，項目配点の妥当性の検証が不十分であり，その合計得点で評価される習熟度もまた適切とは

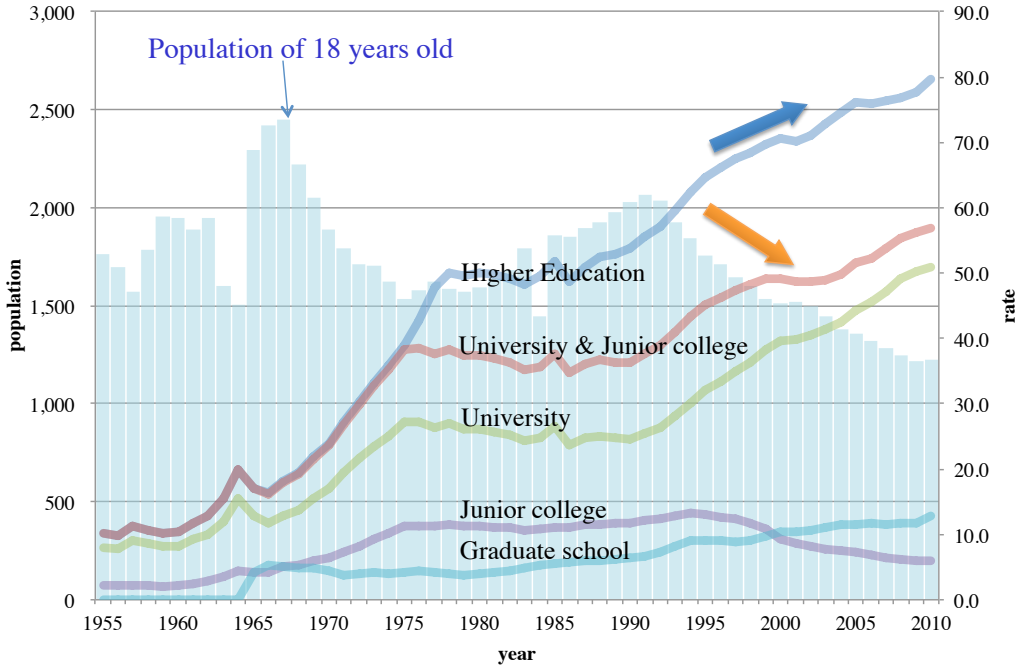


図 1.1: 18 歳人口における高等教育機関への進学率の推移 (*www.mext.go.jp*) [41]

言い難い．項目特性は試験が持つ性質から特徴付けられるべきものであり，それに相応しい処理がなされてこそ，公正な評価法の進展が望めるのである．つまり，問題項目が持つ潜在的な性質を具現化させ，それを積極的に活用する IRT こそが e-learning システムの核となるのである．

IRT に対する研究は 20 世紀半ばから主に欧米を中心に盛んに行われてきた．IRT は，受験者と問題項目との間に現れる正答確率をパラメトリックな関数としてモデル化することで，テストの回答結果から受験者能力と問題項目特性を同時に評価する．当初，IRT ではテスト回答結果が増えると，推定対象であるパラメータも増加してしまう問題があり，パラメータ推定は困難を極めたが，周辺最尤推定法（MMLE: marginal maximum likelihood estimation）[4, 5, 17] が考案されて現実的なものとなった．その後，EM アルゴリズム，ベイズ理論 [18]，マルコフ連鎖モンテカルロ法 [22] などの適用が研究されて，パラメータ推定における困難さは大きく改善されてきた．その計算の複雑さから，コンピュータを用いた演算は必要不可欠であり，近年の PC の普及と高性能化に伴って，世界

的に広く利用されるようになってきている．20 世紀後半には，IRT を用いたテスト解析ツールとして BILOG-MG [3] が開発された．

この IRT を利用し，受験者の能力に最も適切な問題を自動的に選択して出題する e-learning システムをオンライン適応型にすることで，より速くより正確な受験者の能力を自動的に判定できるオンライン適応型試験システムを構築することが可能である．このシステムは，受験者ごとの能力に合わせたレベルの問題を逐次出題し，その都度，受験者の能力を評価する．

これまでのオンライン適応型試験システムでは，受験者の能力に合わせたレベルの問題を出題するために，あらかじめ出題される問題の項目特性を知っておく必要があった．そのためには，事前に項目バンクに登録される項目群から構成されるテストを実施し，その回答結果から問題項目特性を推定しておかなければならない．これを予備テストという．予備テストを受験する受験者集団は，オンライン適応型試験システムを受験する集団とは異なる必要がある．そのため，予備テストによって項目バンクを作成することは，大変な労力がかかる．

1.2 本論文の目的

IRT を利用した解析ツールとして，SSI 社が提供している有償ソフトウェア BILOG-MG [3] が有名であり，詳細な分析を可能としている反面，IRT の専門的な知識無しでは扱いが難しく，その入出力もテキスト形式で行われるため管理しづらいという問題があった．本研究では，IRT の一般的な評価法に触れることを目的とした，より容易な操作で簡単に扱える解析ツールの開発を行う．データの入出力形式は 一般に広く普及している Microsoft Office Excel を利用することで，データ管理面での改善も行う．また，開発したツールを Web アプリケーションとすることで，IRT の利用機会の拡大を狙う．

また，IRT を用いた評価法において核となる働きをするのが EM アルゴリズムを用いた周辺最尤推定法 [4,5,17] であり，ある受験者集団が同一の項目群を回答した場合に限り，問題項目特性および受験者能力を両方同時に評価できる．オンライン適応型試験システムのような e-learning による試験では，受験者によって回答する項目群が異なるため，

ある受験者集団により得られる試験結果はスパースな値を持つ不完全なマトリクスとなり，通常の評価法では両方同時に評価できない．そこで，試験結果の期待値は、受験者能力と項目特性から得られる正答確率で表されるという統計的性質を利用して，スパースな値を補完する方法を開発し，e-learning データへの IRT の新しい適用法を試みる．

1.3 本論文の構成

本論文では，こういった背景をもとに，II 部構成で議論を進める．まず，第 I 部では，IRT の基本的な概念と，これまでに考案されてきたパラメータの推定方法について紹介するとともに，IRT に触れる機会を改善する試みとして，容易に利用可能な Web アプリケーションの開発を行う．次に，第 II 部では，IRT を利用した適応型試験について議論する．ここでは，適応型試験による能力評価は，一部の大規模な検定試験などでのみ利用されており，大学などの教育機関ではあまり利用されていない理由について述べるとともに，その解決策として，新たな手法を提案し，それにもとづいた適応型試験の運用法について，実例を交えて述べる．

まず，第 2 章で現代テスト理論として一般化し普及しつつある IRT について説明する．IRT は受験者個人の能力とテスト項目の特性を独立なものとして仮定することで，従来の古典的テスト理論とは全く異なる，受験したテストに依存しない個人の能力測定を可能とする．さらに，IRT における，各問題項目の特性と受験者個人の能力尺度値の推定方法について述べる．推定方法としては，まず，受験者能力を確率変数として積分消去し項目特性のみを推定可能とする周辺最尤推定法 [4,5,17] を紹介する．次に，ベイズ理論を利用した周辺ベイズ推定（MBE: marginal baysian estimation）法 [18] について解説する．さらに，マルコフ連鎖モンテカルロ（MCMC: Malcov chain Monte Carlo）法を IRT に用いる方法 [22] について述べる．最後に，これらの手法を比較，検証することで，それぞれの推定法の特徴を見る．

第 3 章では，IRT をもとに作成した Web アプリケーションを提案する [1]．これは，インターネットを通じて，クライアント側の PC 上で動作するもので，データの入出力には一般的に普及している Microsoft Office Excel 2003 を利用する．誰もがより簡単な操

作で IRT に触れることができるものを目指したアプリケーションである．世界的に著名な BILOG-MG [3] による推定結果との比較も行う．

第 4 章では，IRT を利用した e-learning システムとして，オンライン適応型試験システムについて説明する．このシステムを運用する上で必要な事前準備について述べるとともに，実際に運用して得られる不完全マトリクスについて説明する．

第 5 章では，オンライン適応型試験システムによって得られる不完全マトリクスに対して IRT を適用するための新たな手法を提案する．オンライン適応型試験システムを受験する受験者数が増加すると，その受験者らに合わせた項目特性のキャリブレーションが必要になってくる．しかし，得られる回答結果は受験者によって回答した項目が異なるため，全体としては不完全マトリクスとなる．従来の IRT は完全マトリクスを想定した推定を行っているため，直接は適用できない．本提案法によって，不完全マトリクスから項目特性を推定することを試みる．まず，テストの問題項目特性および受験者能力が分かっているときに不完全マトリクスを模擬したデータを用いて，提案法が元のパラメータを再現できることを確認し，次に実際に行ったテストに対して提案法による予測を行う．

第 6 章では，本提案法による不完全マトリクス問題の解決内容は，ユーザーとアイテムから構成される評価パターンの不完全マトリクスからユーザーの嗜好を予測して商品を推薦する推薦システムとよく似ていることに注目し，推薦システムにおける不完全マトリクスの予測法の 1 つとして一般に有名なマトリクス分解法と本提案法とを比較することで，本提案法による不完全マトリクス予測の有効性を述べる．

第 7 章では，これらの研究の総評を述べる．

第 I 部

応答に連続値を許す項目反応理論とそのオンラインシステムへの実装

第 2 章

応答に連続値を許す項目反応理論

本章では項目反応理論 [9, 10, 27] の基礎概念について述べる．項目反応理論は，受験者の学力に対して潜在的要因の存在を仮定することから始まる．現代テスト理論として世界中で利用されており，その研究も盛んである．ここでは，基本的な項目反応理論の考え方を紹介する．また，項目反応理論で扱う試験データは本来は正答を 1，誤答を 0 とした 0/1 の二値データであるが，ここでは，それを連続値に拡張できることについても述べる．

2.1 項目反応理論

2.1.1 反応パターン

表 2.1 の 0 と 1 の数字の並びは，テストの反応パターンの例である．反応パターンを $\boldsymbol{\delta}$ で表す．$\boldsymbol{\delta}$ は，N 人の受験者が n 問の項目から構成されるテストを受験した結果を表している．つまり，$\boldsymbol{\delta}$ は $N \times n$ の行列である．ここで 0 は不正解，1 は正解を意味している．例えば，$\delta_{i,j} = 1$ であるとき，受験者 i が項目 j に正解したことを表す．

2.1.2 配点方式の問題点

従来の古典的テスト理論の代表的な評価法の一つに，配点方式による評価がある．この評価方法は日本では現在もなお広く用いられており，最も慣れ親しんだものである．この

表 2.1: 反応パターンの例

受験者 1	11111110100111001111010010001	$\cdots$
受験者 2	00111110110111100010111010011	$\cdots$
受験者 3	00000110000000100000011110000	$\cdots$
受験者 4	00000000011110011111011110101	$\cdots$
受験者 6	11101111001111111010111010100	$\cdots$
受験者 7	10111110111111011110110010000	$\cdots$
受験者 8	00000100001000000000000010000	$\cdots$
受験者 9	00111110111110000000001010000	$\cdots$
$\vdots$	$\vdots$	$\ddots$

評価方法では，個人の能力の大小はテスト得点の大小で対応づけられる．つまり，各受験者のテスト合計得点 T_i は，各項目 j の配点を w_j とすると，

$$T_i = \sum_{j=1}^{n} w_j \delta_{i,j}$$

で表される．

ここで，4 問の項目で作成されたテストがあるとしよう．もし各問の配点を 1 とすると，テスト得点は正解した問題数で定義されるから，各反応パターンに対応するテスト得点は表 2.2 の得点 1 の列になる [29]．ここでは，パターン 2-5 の人（1 点），6-11 の人（2 点），12-15 の人（3 点）は，それぞれ同一の得点が与えられるため，区別することができない．

ここで，問 1，2，3，4 の配点をそれぞれ，2，2，3，3 に変えると，満点が 10 点となる．各反応パターンの人の得点は表 2.2 の得点 2 の列になる．そこでは 0-10 点までの 9 種類の得点に分けられ，各配点 1 のときに比べて同点となる割合は少なくなる．

このように，テスト問題への受験者の反応パターンとその配点から定義されるテスト得点による評価は，配点の違いによって変化する．標準得点や偏差値などの評価も，配点の合計であるテスト得点にもとづいたものであるため同様である．つまり，受験者の「能力」と「配点合計のテスト得点」とを異なるものとして扱わなければならない．

また，受験者集団についても問題がある．能力の高い集団が受験すれば，その問題の正答率は高まり，見かけ上問題が易しかったという結果となる．逆に能力の低い集団が受験

表 2.2: 4 問のテストの回答パターンとテスト得点 [29]

パターン 配点 1 配点 2	問 1 1 2	問 2 1 2	問 3 1 3	問 4 1 3	得点 1 満点 4	得点 2 満点 10	生起確率 $\prod P_j^{\delta_{i,j}} Q_j^{1-\delta_{i,j}}$
1	0	0	0	0	0	0	0.038
2	1	0	0	0	1	2	0.154
3	0	1	0	0	1	2	0.058
4	0	0	1	0	1	3	0.026
5	0	0	0	1	1	3	0.010
6	1	1	0	0	2	4	0.230
7	1	0	1	0	2	5	0.102
8	1	0	0	1	2	5	0.038
9	0	1	1	0	2	5	0.038
10	0	1	0	1	2	5	0.014
11	0	0	1	1	2	6	0.006
12	1	1	1	0	3	7	0.154
13	1	1	0	1	3	7	0.058
14	1	0	1	1	3	8	0.026
15	0	1	1	1	3	8	0.010
16	1	1	1	1	4	10	0.038
能力が θ の人 の正答確率	$P_1(\theta)$ 0.8	$P_2(\theta)$ 0.6	$P_3(\theta)$ 0.4	$P_4(\theta)$ 0.2			計 1.000

した場合は正答率が下がり，問題は難しかったという結果となる．つまり，正答率をもとに問題の特性を測る場合，受験者集団に依存することになる．そのため，テストの問題項目の本質的な「難しさ」と「正答率」もまた異なるものとして扱わなければならない．

本研究で扱う項目反応理論の大きな特徴は，「受験者個人の能力特性」と使用した「テスト問題の項目特性」を異なるものとして独立に扱うことによって，従来の配点方式で評価される古典的テスト理論が抱える問題の解決を目指すということである．

2.1.3 尤度の定義

項目反応理論は，観測されたテスト問題への回答結果 $\delta_{i,j}$ と，現実には直接観測されない潜在特性 θ を結びつけるための 1 つの数学モデルである．

いま，受験者 i の能力を表す特性値を θ_i，問題項目 j の特性を表す特性値を ϕ_j とする．このとき，受験者 i が問題 j に正答する確率を $P(\theta_i, \phi_j)$ とする．誤答する確率は $Q(\theta_i, \phi_j) = 1 - P(\theta_i, \phi_j)$ となる．

n 個の項目（$j = 1, 2, \cdots, n$）からなるテストを考えたとき，受験者 i と問題項目 j の観測値（項目得点）$\delta_{i,j}$ は $\delta_{i,j} = 1$（正答），$= 0$（誤答）をとる．このとき項目反応理論では，n 個の問題 1 つ 1 つの反応パターンに着目し，能力が θ_i の受験者 i が所与の反応パターンを生起させる条件付き確率を考えると，それぞれの生起は項目ごとに独立と考えて次の形で示される．

$$P(\delta_{i,j}|\theta_i) = \prod_{j=1}^{n} P(\theta_i, \phi_j)^{\delta_{i,j}} Q(\theta_i, \phi_j)^{1-\delta_{i,j}} \tag{2.1}$$

これは反応パターン $\delta_{i,j}$ を所与としたときの，θ_i の尤もらしさを表す尤度（likelihood）と考えることができる．よって，n 個の項目に対する N 人の受験者（$i = 1, 2, \ldots, N$）の反応パターン行列 $\boldsymbol{\delta}$ が観察される確率（尤度）は

$$L = P(\delta_{i,j}|\theta_i, \phi_j) = \prod_{i=1}^{N} \prod_{j=1}^{n} P(\theta_i, \phi_j)^{\delta_{i,j}} Q(\theta_i, \phi_j)^{1-\delta_{i,j}} \tag{2.2}$$

となる．

このとき，$\delta_{i,j}$ は試験結果から得られる値であるため，あらかじめ項目の特性値 ϕ_j が既知である項目によって作成されたテストであれば，直ちに受験者 i の能力 θ_i を得ることができる．もし，項目の特性値 ϕ_j が未知であれば，2.3 節で述べる方法によって，項目の特性値 ϕ_j と能力パラメータ θ_i の両方を推定することができる．

2.1.4 項目特性曲線

テストは項目の集まりであり，テストの性質は項目の性質によって規定される．項目の性質は，縦軸に正答確率 $P(\theta, \phi_j)$ を，横軸に受験者の能力 θ を取ることで，関数の形によって特徴付けられる．このときの曲線を項目特性曲線（ICC）という．θ は直接観測することが不可能な特性であるために，具体的な関数形は一通りには決定できないが，一般的には計算が容易なロジスティック関数によって与えられる．

正規累積モデル

統計学で最も頻繁に利用される標準正規分布の密度関数

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{2.3}$$

の累積分布関数

$$\Phi(f(\theta)) = \int_{-\infty}^{f(\theta)} \phi(z)dz \tag{2.4}$$

を ICC として利用したものを正規累積モデルという．$f(\theta)$ としては，θ に関する単調増加関数を選ぶ．このとき $\Phi(f(\theta))$ も θ に関する単調増加関数となる．

ロジスティックモデル

ロジスティックモデルは，正規累積モデルをより簡単な数式で近似したモデルである．正規累積モデルでは，積分が含まれているため，パラメータ推定の際に解析的に解くことが難しくなる．そこで，ロジスティックモデルで正規累積モデルを近似することで，積分計算の煩わしさから解放され，比較的容易な計算が可能となる．

■1 母数ロジスティックモデル　1 母数ロジスティックモデルは，Rash モデルとも呼ばれる．1 母数正規累積モデルをロジスティック近似したものであり，$\phi_j = \{b_j\}$ として，次式で表される．

$$P(\theta, \phi_j) = \frac{1}{1 + \exp\left(-Da(\theta - b_j)\right)} \tag{2.5}$$

ここで，D は定数であり，より正規累積モデルに近似するために $D = 1.7$ とすることが多い．b_j は項目 j の難しさを決める母数であり，項目困難度（item difficulty）あるいは，単に困難度と呼ばれる．ここで a はすべての項目に共通の値である．このとき，ICC は図 2.1 のようになる．曲線が左側のほうにある項目ほど困難度が低く，つまり易しい項目であったことを表している．1 母数ロジスティックモデルの場合，各 ICC は位置が異なるだけであり，傾きは等しいため交わることはない．

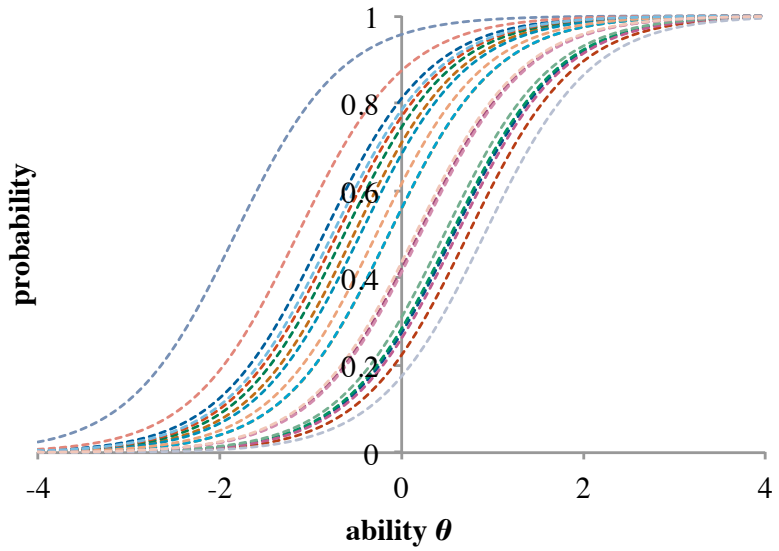


図 2.1: 1 母数ロジスティックモデルによる ICC の例

■2 母数ロジスティックモデル　2 母数正規累積モデルをロジスティック近似したものである．$\phi_j = \{a_j, b_j\}$ として，次式で表される．

$$P(\theta, \phi_j) = \frac{1}{1 + \exp\left(-Da_j(\theta - b_j)\right)} \tag{2.6}$$

a_j は項目 j の受験者能力に対する敏感さを表す母数であり，識別力（descriminating parameter）と呼ばれる．2 母数ロジスティックモデルの場合，1 母数ロジスティックモデルに加えて，各 ICC に傾きの変化が加わる．このモデルは，項目反応理論において最もポピュラーなモデルであり，本研究でもこのモデルを採用している．図 2.2 に，2 母数ロジスティックモデルの ICC を描いている．このように，ICC を見ることで各項目の特性を瞬時に知ることができる．

2.2　δ の有理数への拡張

式 (2.2) において，δ は正答のとき $\delta = 1$，誤答のとき $\delta = 0$ を表す 2 値関数であった．ここではこれを，受験者が同じ困難度を持つ異なる項目 m 問中で l 回正答したと考

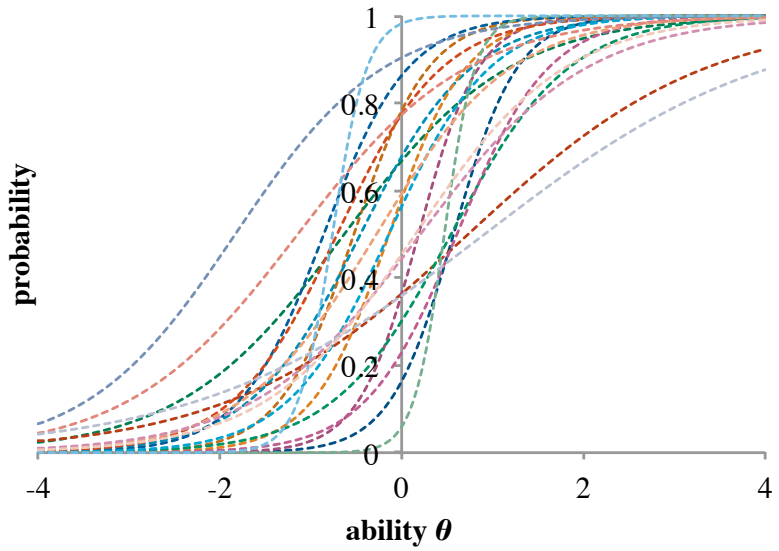


図 2.2: 2 母数ロジスティックモデルによる ICC の例

え $\delta = l/m$ とみなすことによって，δ に対し有理数を割り当てることができるように拡張する．

たとえば，$m = 5$ とすれば，l が取り得る値は，$l = \{0,\ 1,\ 2,\ 3,\ 4,\ 5\}$ であり，このときの δ は

$$\begin{aligned}\delta &= \frac{l}{m} \\ &= \left\{\frac{0}{5},\ \frac{1}{5},\ \frac{2}{5},\ \frac{3}{5},\ \frac{4}{5},\ \frac{5}{5}\right\} \\ &= \{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1\}\end{aligned}$$

となる．

2.3　項目反応理論におけるパラメータ推定

項目反応理論のパラメータ推定方法はさまざまなものが考えられており，一般的に広く使われているのは，最尤推定法を用いることである．最尤推定法では式 (2.2) をもとにパラメータ推定を行う．このとき，項目パラメータと能力パラメータを両方同時に

求める方法では，推定した値が一致推定量とならない問題が指摘されている [21]．そこで，能力パラメータを確率変数とみなし，これを周辺化によって積分消去することで，先に項目パラメータのみを求める周辺最尤推定（MMLE: marginal maximum likelihood estimation）法 [4, 5, 17] が考えられている．また，これにベイズ理論を応用した周辺ベイズ推定（MBE: marginal Bayesian estimation）法 [18] も考えられている．これには，各パラメータの事前分布を組み込むことで極端な値が推定されることを防ぐという特徴がある．項目特性を既知としたとき，受験者能力値のみを推定する方法としては，ベイズ最大事後（BMAP: Bayesian maximum a posteriori）推定法とベイズ期待事後（BEAP: Bayesian expectation a posteriori）推定法がある [18]．このとき，計算コストが低く安定した推定値が得られる MBE および BEAP を用いると良いとの報告がある [18]．最尤推定法以外では，マルコフ連鎖モンテカルロ法（MCMC: malcov chain Monte Carlo method）を用いた方法 [22] などがある．本章では，MBE，BMAP，BEAP，MCMC について，その概要を説明する．なお，ICC を表す関数としては，2 母数ロジスティックモデルを採用している．

2.3.1 周辺ベイズ推定法

ベイズの定理から，パラメータ $\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\}$ の事前確率分布 $g(\theta), g(a), g(b)$ と，与えられたパラメータ $\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\}$ のもとで応答パターンを得る条件付き確率（尤度）$L(\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})$ とを掛け合わせたものが，与えられた応答パターンのもとでパラメータ $\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\}$ を得る確率（事後分布）$g\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b} \mid \boldsymbol{\delta}\}$ に比例する．つまり，

$$g\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b} \mid \boldsymbol{\delta}\} \propto L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} g(a) g(b) g(\theta) \tag{2.7}$$

となる．周辺ベイズ推定法は，これを応用し，適切な事前分布を仮定することによって，応答パターン $\boldsymbol{\delta}$ とパラメータ $\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\}$ との周辺分布の事後確率が最も大きくなるようなパラメータを探していく方法である．

超パラメータ

パラメータ $\{\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\}$ の事前分布を決めるパラメータを超パラメータと呼ぶ．能力パラメータ θ_i については，

$$\theta_i \sim N(\mu_\theta, {\sigma_\theta}^2) \tag{2.8}$$

と仮定する（$-\infty < \theta_i < \infty$）．困難度パラメータ b_j については，

$$b_j \sim N(\mu_b, {\sigma_b}^2) \tag{2.9}$$

と仮定する（$-\infty < b_j < \infty$）．識別力パラメータ a_j については，

$$a_j \sim \texttt{log-normal}(\mu_\alpha, {\sigma_\alpha}^2) \tag{2.10}$$

と仮定する（$0 < a_j < \infty$）．

項目パラメータの推定

未知の項目パラメータを推定するためには，式 (2.7) の右辺の 1 階微分をとり，イコール 0 を満たすパラメータを求めるとよい．そこで，まずは式 (2.7) の右辺を l として，その対数をとる．

$$\log l = \log L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} + \log g(a)g(b) + \log g(\theta) \tag{2.11}$$

さらに，式 (2.11) に関して，極値をとる項目パラメータを求めるために 1 階微分する．

$$\frac{\partial}{\partial a_j} \log L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} + \frac{\partial}{\partial a_j} \log g(a)g(b) + \frac{\partial}{\partial a_j} \log g(\theta) = 0 \tag{2.12}$$

$$\frac{\partial}{\partial b_j} \log L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} + \frac{\partial}{\partial b_j} \log g(a)g(b) + \frac{\partial}{\partial b_j} \log g(\theta) = 0 \tag{2.13}$$

このとき，$\log g(\boldsymbol{\theta})$ は項目パラメータを含んでいないため，0 となる．また，それぞれ互いの事前分布は微分すると 0 となるから，結局，上式は

$$\frac{\partial \log l}{\partial a_j} = \frac{\partial}{\partial a_j} \log L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} + \frac{\partial}{\partial a_j} \log g(a) = 0 \tag{2.14}$$

$$\frac{\partial \log l}{\partial b_j} = \frac{\partial}{\partial b_j} \log L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} + \frac{\partial}{\partial b_j} \log g(b) = 0 \tag{2.15}$$

となる．

■周辺尤度　このとき，式 (2.2) について

$$L = L\{\boldsymbol{\delta} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\} = \prod_{i=1}^{N} P(\boldsymbol{\delta}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{N} \int P(\boldsymbol{\delta}_i \mid \boldsymbol{a}, \boldsymbol{b}) g(\theta) d\theta \tag{2.16}$$

と周辺化する．これは能力パラメータが積分消去された形になっており，項目パラメータに関する周辺尤度になる．

■一階微分　このとき，式 (2.14) の左辺第 1 項は，式 (2.16) から

$$\begin{aligned} \frac{\partial}{\partial a_j} \log L &= \frac{\partial}{\partial a_j} \sum_{i=1}^{N} \log \left\{ \int P(\boldsymbol{\delta}_i \mid \boldsymbol{a}, \boldsymbol{b}) g(\theta) d\theta \right\} \\ &= D \sum_{i=1}^{N} \int (\delta_{i,j} - P(\theta_i, a_j, b_j))(\theta_i - b_j) \\ &\quad \times \left[\frac{P(\boldsymbol{\delta} \mid a_j, b_j) g(\theta)}{\int P(\boldsymbol{\delta} \mid a_j, b_j) g(\theta) d\theta} \right] d\theta \end{aligned} \tag{2.17}$$

同じく，式 (2.15) の左辺第 1 項は，式 (2.16) から

$$\begin{aligned} \frac{\partial}{\partial b_j} \log L &= -D a_j \sum_{i=1}^{N} \int (\delta_{i,j} - P(\theta_i, a_j, b_j)) \\ &\quad \times \left[\frac{P(\boldsymbol{\delta} \mid a_j, b_j) g(\theta)}{\int P(\boldsymbol{\delta} \mid a_j, b_j) g(\theta) d\theta} \right] d\theta \end{aligned} \tag{2.18}$$

となる．

また，式 (2.14) と式 (2.15) の左辺第 2 項は，式 (2.9), 式 (2.10) より

$$g(a) = \frac{1}{\sqrt{2\pi}\sigma_a a_j} \exp\left[-\frac{1}{2} \left(\frac{\log a_j - \mu_a}{\sigma_a} \right)^2 \right] \tag{2.19}$$

$$g(b) = \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left[-\frac{1}{2} \left(\frac{b_j - \mu_b}{\sigma_b} \right)^2 \right] \tag{2.20}$$

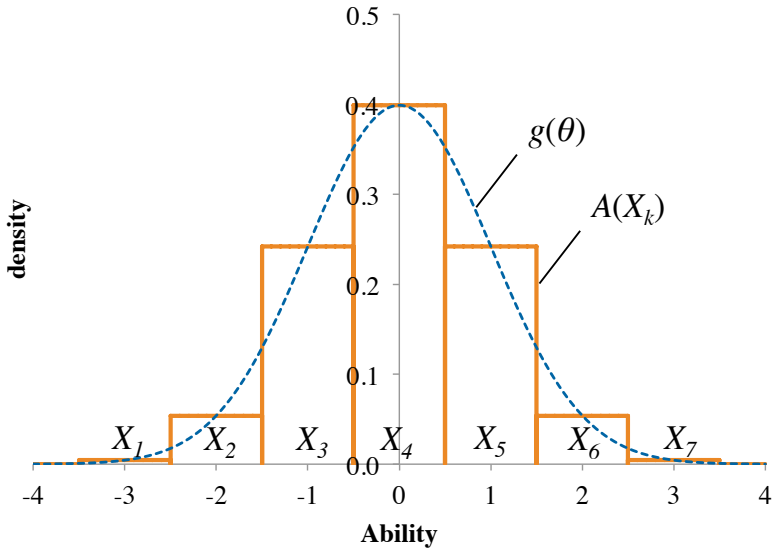


図 2.3: 区分求積法の概念図

であるから，

$$\frac{\partial}{\partial a_j}\log g(a) = -\frac{1}{a_j} - \frac{\log a_j - \mu_a}{a_j \sigma_a^2} \tag{2.21}$$

$$\frac{\partial}{\partial b_j}\log g(b) = -\frac{b_j - \mu_b}{\sigma_b^2} \tag{2.22}$$

となる．

■区分求積法　式 (2.17) と式 (2.18) に含まれる積分を直接解析的に求めることが難しいため，区分求積法を用いる．つまり，能力尺度 θ をいくつかの離散的な求積点 X_k，$(k = 1, 2, \ldots, q)$ に分け，そこでの観測頻度 $A(X_k)$ で重みづけ合計することによって，近似的に周辺尤度を求める．このとき，θ を X_k に，$g(\theta)$ を $A(X_k)$ に近似する．図 2.3 はその概念図を表している．たとえば，求積点を等間隔に $q = 15$ の点でとったとき，表 2.3 を得る．

このとき，式 (2.1) は

$$P(X_k, a_j, b_j) = \frac{1}{1 + \exp\{-Da_j(X_k - b_j)\}}$$

表 2.3: 等間隔にとった求積点と重みの例（$q = 15$ のとき）

k	求積点 X_k	重み $A(X_k)$
1	-4.00	7.65×10^{-5}
2	-3.43	6.39×10^{-4}
3	-2.86	3.85×10^{-3}
4	-2.29	1.67×10^{-2}
5	-1.71	5.24×10^{-2}
6	-1.14	1.19×10^{-1}
7	-0.57	1.94×10^{-1}
8	0.00	2.28×10^{-1}
9	0.57	1.94×10^{-1}
10	1.14	1.19×10^{-1}
11	1.71	5.24×10^{-2}
12	2.29	1.67×10^{-2}
13	2.86	3.85×10^{-3}
14	3.43	6.39×10^{-4}
15	4.00	7.65×10^{-5}

と表される．すると，周辺尤度式 (2.16) は

$$\log L = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{q} A(X_k) \prod_{j=1}^{n} P(X_k, a_j, b_j)^{\delta_{i,j}} (1 - P(X_k, a_j, b_j))^{1-\delta_{i,j}} \right] \tag{2.23}$$

となる．このとき，式 (2.17)，式 (2.18) は

$$\begin{aligned} \frac{\partial \log L}{\partial a_j} = D \sum_{i=1}^{N} \sum_{k=1}^{q} [\delta_{i,j} - P(X_k, a_j, b_j)](X_k - b_j) \\ \times \left[\frac{A(X_k) \prod_{j=1}^{n} P(X_k, a_j, b_j)^{\delta_{i,j}} (1 - P(X_k, a_j, b_j))^{1-\delta_{i,j}}}{\sum_{k=1}^{q} A(X_k) \prod_{j=1}^{n} P(X_k, a_j, b_j)^{\delta_{i,j}} (1 - P(X_k, a_j, b_j))^{1-\delta_{i,j}}} \right] \end{aligned} \tag{2.24}$$

$$\frac{\partial \log L}{\partial b_j} = -Da_j \sum_{i=1}^{N}\sum_{k=1}^{q}[\delta_{i,j} - P(X_k, a_j, b_j)] \times \left[\frac{A(X_k)\prod_{j=1}^{n} P(X_k, a_j, b_j)^{\delta_{i,j}}(1 - P(X_k, a_j, b_j))^{1-\delta_{i,j}}}{\sum_{k=1}^{q} A(X_k)\prod_{j=1}^{n} P(X_k, a_j, b_j)^{\delta_{i,j}}(1 - P(X_k, a_j, b_j))^{1-\delta_{i,j}}} \right] \tag{2.25}$$

となる．

■EM アルゴリズム

E-step　$\theta = X_k$ 付近の項目反応パターンに対応する尤度は

$$L(X_k) = \prod_{j=1}^{n} P(X_k, a_j, b_j)^{\delta_{i,j}}(1 - P(X_k, a_j, b_j))^{1-\delta_{i,j}} \tag{2.26}$$

と表される．ここで，求積点 X_k という能力を持つことが期待される受験者の数 $\bar{f}_{jk}$ は

$$\bar{f}_{jk} = \sum_{i=1}^{N} \left[\frac{L(X_k)A(X_k)}{\sum_{k=1}^{q} L(X_k)A(X_k)} \right] \tag{2.27}$$

で与えられる．また，その中で期待される正答者の数 $\bar{r}_{jk}$ は

$$\bar{r}_{jk} = \sum_{i=1}^{N} \left[\frac{\delta_{i,j}L(X_k)A(X_k)}{\sum_{k=1}^{q} L(X_k)A(X_k)} \right] \tag{2.28}$$

で与えられる．

M-step　尤度を最大にする項目パラメータを，Newton-Raphson-Fisher 法によって求める．すなわち，

$$\begin{bmatrix} a_j \\ b_j \end{bmatrix}_{t+1} = \begin{bmatrix} a_j \\ b_j \end{bmatrix}_t + \begin{bmatrix} -E\left(\frac{\partial^2 \log l}{\partial a_j^2}\right) & -E\left(\frac{\partial^2 \log l}{\partial a_j \partial b_j}\right) \\ -E\left(\frac{\partial^2 \log l}{\partial b_j \partial a_j}\right) & -E\left(\frac{\partial^2 \log l}{\partial b_j^2}\right) \end{bmatrix}_t^{-1} \cdot \begin{bmatrix} \frac{\partial \log l}{\partial a_j} \\ \frac{\partial \log l}{\partial b_j} \end{bmatrix}_t$$

として解くことができる．対数事後分布 $\log l$ の 1 階微分は，式 (2.21)，(2.22)，(2.24)，(2.25)，(2.27)，(2.28) より，

$$\frac{\partial \log l}{\partial a_j} = D\sum_k (X_k - b_j)[\bar{r}_{jk} - \bar{f}_{jk}P(X_k, a_j, b_j)] - \frac{1}{a_j} - \frac{\log a_j - \mu_a}{a_j \sigma_a^2}$$
$$\frac{\partial \log l}{\partial b_j} = -Da_j \sum_k [\bar{r}_{jk} - \bar{f}_{jk}P(X_k, a_j, b_j)] - \frac{b_j - \mu_b}{\sigma_b^2}$$

となる．対数事後分布の 2 階微分の期待値は，$E[\bar{r}_{jk}] = \bar{f}_{jk}P_j(X_k)$ を利用して，

$$E\left(\frac{\partial^2 \log l}{\partial a_j^2}\right) = -D^2 \sum_k (X_k - b_j)^2 \bar{f}_{jk} P(X_k, a_j, b_j)(1 - P(X_k, a_j, b_j)) + \frac{1}{a_j^2} - \frac{1 - \log a_j + \mu_a}{a_j^2 \sigma_a^2}$$
$$E\left(\frac{\partial^2 \log l}{\partial b_j^2}\right) = -D^2 a_j^2 \sum_k \bar{f}_{jk} P(X_k, a_j, b_j)(1 - P(X_k, a_j, b_j)) - \frac{1}{\sigma_b^2}$$
$$E\left(\frac{\partial^2 \log l}{\partial b_j \partial a_j}\right) = D^2 a_j \sum_k (X_k - b_j) \bar{f}_{jk} P(X_k, a_j, b_j)(1 - P(X_k, a_j, b_j))$$

となる．項目パラメータが安定しない場合は，E-step へ戻る．パラメータが安定するまで，E-step，M-step を繰り返す．

能力パラメータの推定

能力パラメータの推定では，前述の方法で項目パラメータが推定された状態をもとに行う．推定方法としては，ベイズ理論をもとに，ベイズ最大事後推定法とベイズ期待事後推定法が考案されている．

■ベイズ最大事後推定法　ベイズ最大事後（BMAP: bayesian maximum a posteriori）推定法は，次のベイズの定理

$$g\{\theta_i \mid \boldsymbol{\delta}_i, a, b\} \propto L\{\boldsymbol{\delta}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}\} g(\theta) \tag{2.29}$$

にもとづいている．ただし，$g(\theta) \sim N(\mu_\theta, {\sigma_\theta}^2)$ である．式 (2.29) の両辺の対数をとると，

$$\log g\{\boldsymbol{\delta}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}\} \propto \log L\{\boldsymbol{\delta}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}\} + \log g(\theta) \tag{2.30}$$

ここで，

$$L\{\boldsymbol{\delta}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}\} = \prod_{j=1}^{n} P(\theta_i, a_j, b_j)^{\delta_{i,j}} (1 - P(\theta_i, a_j, b_j))^{1-\delta_{i,j}} \tag{2.31}$$

である．式 (2.30) の右辺を $\log l$ として，能力パラメータ θ_i で微分すると，

$$\frac{\partial \log l}{\partial \theta_i} = D \sum_{j=1}^{n} \{a_j(\delta_{i,j} - P(\theta_i, a_j, b_j))\} - \frac{\theta_i - \mu_\theta}{{\sigma_\theta}^2} \tag{2.32}$$

となる．また，2 階微分の期待値を取ると，

$$E\left(\frac{\partial^2 \log l}{\partial {\theta_i}^2}\right) = -D^2 \sum_{j=1}^{n} \left\{{a_j}^2 P(\theta_i, a_j, b_j)(1 - P(\theta_i, a_j, b_j))\right\} - \frac{1}{{\sigma_\theta}^2} \tag{2.33}$$

となる．このとき，以下の Newton-Raphson-Fisher 法による更新式で $\hat{\theta}_i$ を求めることができる．

$$[\hat{\theta}_i]_{t+1} = [\hat{\theta}_i]_t + \left[-E\left(\frac{\partial^2 \log l}{\partial {\theta_i}^2}\right)\right]_t^{-1} \cdot \left[\frac{\partial \log l}{\partial \theta_i}\right]_t \tag{2.34}$$

■ベイズ期待事後推定法　ベイズ期待事後（BEAP: bayesian expect a posteriori）推定法は，計算コストが小さいにも関わらず，精度が良い推定手法である．その推定手法は次のようなものである．ベイズの定理から，

$$g\{\theta_i \mid \boldsymbol{\delta}_i, \boldsymbol{a}, \boldsymbol{b}\} = \frac{g(\theta) \prod_{j=1}^{n} P(\theta_i, a_j, b_j)^{\delta_{i,j}} (1 - P(\theta_i, a_j, b_j))^{1-\delta_{i,j}}}{\int g(\theta) \prod_{j=1}^{n} P(\theta_i, a_j, b_j)^{\delta_{i,j}} (1 - P(\theta_i, a_j, b_j))^{1-\delta_{i,j}} d\theta} \tag{2.35}$$

とできる．その期待値は，区分求積法により，

$$E\{\theta_i \mid \boldsymbol{\delta}_i, \boldsymbol{a}, \boldsymbol{b}\} = \hat{\theta}_i = \frac{\sum_{k=1}^{q} X_k L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)} \tag{2.36}$$

となり，θ_i の平均値を推定値として得ることができるので，期待事後推定値と呼ばれる．このとき，推定値の分散は，

$$Var\{\theta_i \mid \boldsymbol{\delta}_i, \boldsymbol{a}, \boldsymbol{b}\} = \frac{\sum_{k=1}^{q}(X_k - \hat{\theta}_i)^2 L(X_k)A(X_k)}{\sum_{k=1}^{q} L(X_k)A(X_k)} \tag{2.37}$$

として求めることができる．

2.3.2 マルコフ連鎖モンテカルロ法

IRT では，その次元数の多さや積分などによって，複雑な計算を要する．これは解析的に解くことが限りなく不可能であり，そのために近似的に解く必要がある．そのアルゴリズムとして，区分求積法や周辺化が用いられる．近似計算には，測定者が設定しなければならないチューニングパラメータがあり，これは推定結果に影響を及ぼす．この問題を解決する方法として，マルコフ連鎖モンテカルロ法を用いることが考えられる．

マルコフ連鎖モンテカルロ法の IRT への適用

マルコフ連鎖モンテカルロ法（MCMC: malcov chain Monte Carlo method）は，受験者母数と項目母数の同時事後分布から推定を行う．そのため，周辺化を必要としない．推定には M-H 法や Gibbs-sampler 法が考えられる．ここでは，IRT のために考案された M-H within Gibbs 法を用いる．ここで，$\beta = [a, b]$ とする．

2 母数ロジスティックモデルに対する MCMC

この方法では，まず下の手順 1 により，能力パラメータの k 番目の値 θ^k を決定する．次に手順 2 により，項目パラメータの k 番目の値 β^k を決定する．この 2 つの手順を所定の回数 K まで繰り返す．

1. $\theta^{(k)} \sim p(\theta|\boldsymbol{\delta}, \boldsymbol{a}, \boldsymbol{b})$ を求める
   (a) 候補 $\theta_i^* \sim N(\theta_i^{(k-1)}, \sigma_\theta^2)$, $(i = 1, 2, \ldots, N)$ と，$u \sim U(0, 1)$ を生成
   (b) $u \leq \alpha(\theta_i^{(k-1)}, \theta_i^*)$ であれば，$\theta_i^{(k)} = \theta_i^*$ とおく

$\alpha(\theta_i^{(k-1)}, \theta_i^*) = \min\{R_\theta, 1\}$

$$R_\theta = \frac{p(\theta_i^*|\boldsymbol{\delta}, \boldsymbol{a}, \boldsymbol{b})}{p(\theta_i^{(k-1)}|\boldsymbol{\delta}, \boldsymbol{a}, \boldsymbol{b})}$$
$$\propto \frac{p(\boldsymbol{\delta}|\theta_i^*, \boldsymbol{a}, \boldsymbol{b})p(\theta_i^*)}{p(\boldsymbol{\delta}|\theta_i^{(k-1)}, \boldsymbol{a}, \boldsymbol{b})p(\theta_i^{(k-1)})}$$
$$p(\delta_{i,j}|\theta_i, a_j, b_j) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_j(\theta_i)^{\delta_{i,j}} Q_j(\theta_i)^{1-\delta_{i,j}}$$
$$p(\theta_i) = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left[-\frac{1}{2}\left(\frac{\theta_i - \mu_\theta}{\sigma_\theta}\right)^2\right]$$

(c) そうでなければ，$\theta_i^{(k)} = \theta_i^{(k-1)}$ とおく

2. $(a^{(k)}, b^{(k)}) \sim p(a, b|\boldsymbol{\delta}, \boldsymbol{\theta}^{(k)})$ を求める

(a) 候補 $a_j^* \sim \text{lognormal}(a_j^{(k-1)}, \sigma_a^2)$, $b_j^* \sim N(b_j^{(k-1)}, \sigma_b^2)$, $\quad (j = 1, 2, \ldots, n)$ と，
$u \sim U(0, 1)$ を生成

(b) $u \leq \alpha((a_j^{(k-1)}, b_j^{(k-1)}), (a_j^*, b_j^*))$ であれば，$(a_j^{(k)}, b_j^{(k)}) = (a_j^{(*)}, b_j^{(*)})$ とおく
$\alpha((a_j^{(k-1)}, b_j^{(k-1)}), (a_j^*, b_j^*)) = \min\{R_{a,b}, 1\}$

$$R_{a,b} = \frac{p(a_j^*, b_j^*|\boldsymbol{\delta}, \boldsymbol{\theta})}{p(a_j^{(k-1)}, b_j^{(k-1)}|\boldsymbol{\delta}, \boldsymbol{\theta})}$$
$$\propto \frac{p(\boldsymbol{\delta}|\boldsymbol{\theta}, a_i^*, b_i^*)p(a_i^*, b_i^*)}{p(\boldsymbol{\delta}|\boldsymbol{\theta}, a_i^{(k-1)}, b_i^{(k-1)})p(a_i^{(k-1)}, b_i^{(k-1)})}$$
$$p(\delta_{i,j}|\theta_i, a_j, b_j) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_j(\theta_i)^{\delta_{i,j}} Q_j(\theta_i)^{1-\delta_{i,j}}$$
$$p(a_j) = \frac{1}{\sqrt{2\pi}\sigma_a a_j} \exp\left[-\frac{1}{2}\left(\frac{\log a_j - \mu_a}{\sigma_a}\right)^2\right]$$
$$p(b_j) = \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left[-\frac{1}{2}\left(\frac{b_j - \mu_b}{\sigma_b}\right)^2\right]$$

(c) そうでなければ，$(a_j^{(k)}, b_j^{(k)}) = (a_j^{(k-1)}, b_j^{(k-1)})$ とおく

このとき，各手順における候補値は，自ら定義した提案分布から生成される乱数を用いることに注意する．最終的な推定値としては，所定の回数 K までの平均値を用いる．また，

初期段階では不安定な値をとるため，一般的な方法としては初期段階値は平均に含めないことが考えられている．初期段階とみなす回数の閾値を Burn-In と呼ぶ．

2.3.3 シミュレーションによる推定手法の比較

受験者母数，項目母数の乱数を生成する．それをもとに反応パターンを作成する．ここでは受験者 1000 人，項目 20 問を仮定し，また項目特性曲線は 2 母数ロジスティック分布に従うとする．この反応データに対して，MBE および BEAP による推定結果と MCMC による推定結果を比較，検証することで，それぞれの推定手法の特徴を見る．

MCMC による推定

ある受験者 i の能力値 θ_i とある項目 j の特性値 a_j, b_j の MCMC による推定値の遷移を図 2.4 に示す．この図は $k = 5000$ までを表している．上から順番に，ある受験者 i の能力 θ_i，ある項目 j の項目識別力 a_j，項目困難度 b_j である．図中の点線は求めているパラメータの真値を表しており，MCMC による更新によって，どのパラメータも真値の周りに分布していることが分かる．また，更新の初期段階では初期値によっては真値から離れた場所から開始するため，この図では最初の 1000 回程度の計算結果は省いたほうが良いことが分かる．ここでは更新回数 5000 回の結果を図示しているが，更新回数を増やせば増やすほど，初期段階のズレは無視できるものになることが示唆される．

ここで，MCMC の計算回数における能力値 θ についての変化を表 2.4 に示す．表の平

表 2.4: MCMC の回数による能力値 $\hat{\theta}$ [31]

回数	真値	μ	$s.d.$	$bias$	$RMSE$
500	-0.229	0.206	0.419	-0.435	0.604
1000	-0.229	-0.026	0.431	-0.204	0.477
5000	-0.229	-0.239	0.365	0.009	0.365
10000	-0.229	-0.252	0.343	0.022	0.344
30000	-0.229	-0.268	0.324	0.039	0.326

均，標準偏差，バイアス，RMSE は各計算回数までの値を使って算出している．ここで

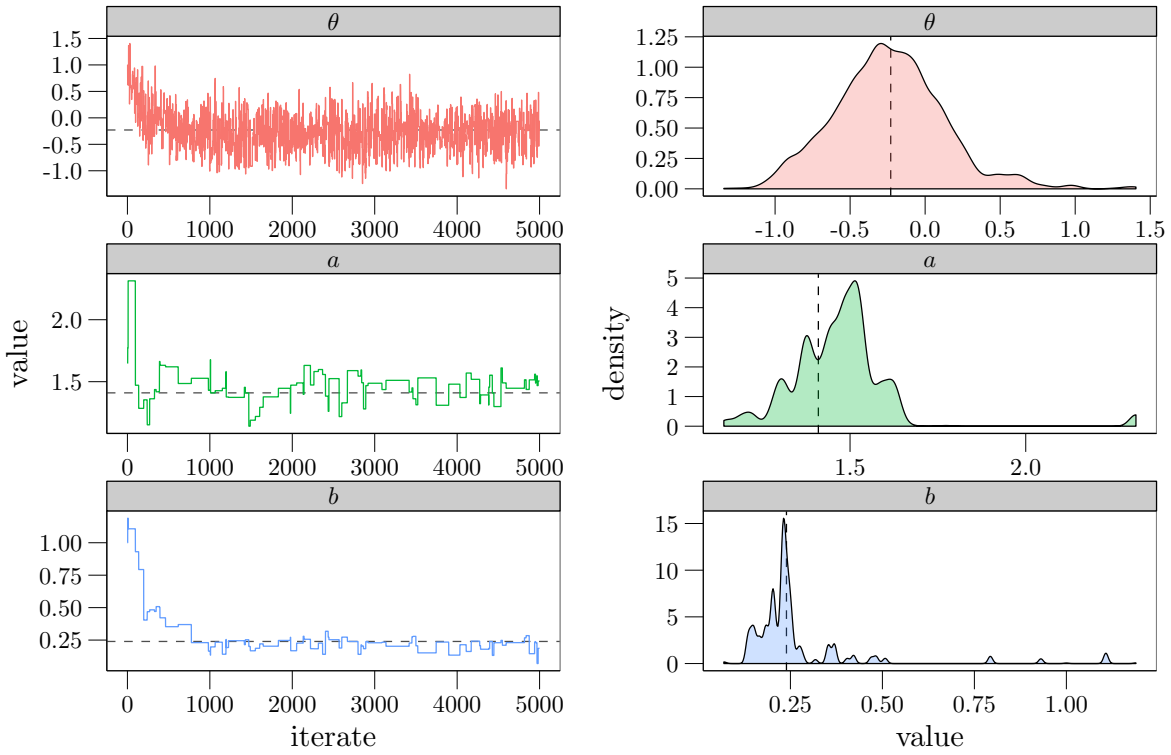


図 2.4: MCMC による推定値の軌跡と頻度（$k = 5000$）[31]

は Burn-In= 0 としている．計算回数が増すごとに，平均値は真値からわずかにずれ（バイアス）が大きくなっているが，標準偏差は小さくなっており，同時に RMSE は小さくなっていることが分かる．

また，MCMC の計算回数における項目識別力 a についての変化を表 2.5 に示す．初期

表 2.5: MCMC の回数による識別力 $\hat{a}$ [31]

回数	真値	μ	$s.d.$	$bias$	$RMSE$
500	1.410	1.589	0.362	-0.179	0.404
1000	1.410	1.561	0.260	-0.151	0.301
5000	1.410	1.470	0.152	-0.060	0.164
10000	1.410	1.498	0.138	-0.088	0.163
30000	1.410	1.497	0.128	-0.088	0.155

表 2.6: MCMC の回数による困難度 $\hat{b}$ [31]

回数	真値	μ	$s.d.$	$bias$	$RMSE$
500	0.239	0.653	0.277	-0.414	0.498
1000	0.239	0.477	0.268	-0.238	0.359
5000	0.239	0.265	0.164	-0.026	0.166
10000	0.239	0.245	0.121	-0.006	0.121
30000	0.239	0.223	0.082	0.017	0.084

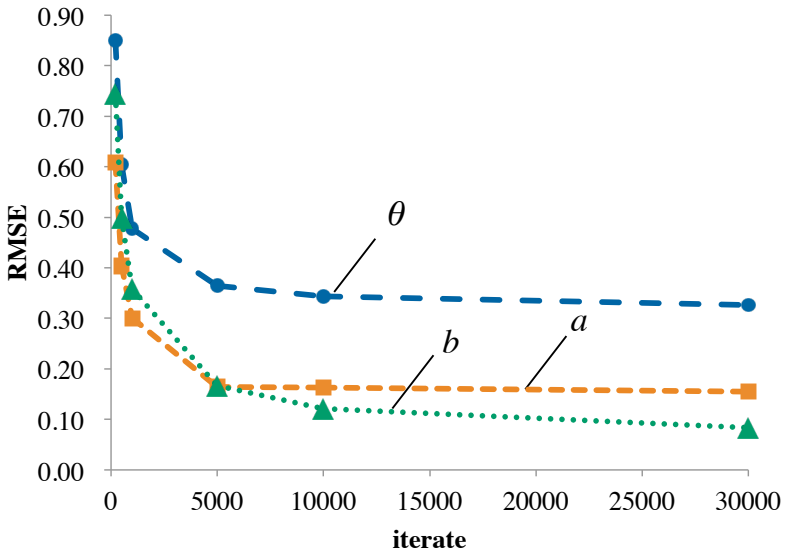


図 2.5: MCMC の回数による RMSE の比較 [31]

値がある程度真値に近いところから開始していたため，平均値は初期段階から真値に近い値を示しているが，やはり傾向としては，ここでも先ほどの能力値 θ の場合と同様，計算回数が増すごとに RMSE が小さくなっている．

また，MCMC の計算回数における項目困難度 b についての変化を表 2.6 に示す．ここでも先ほどの能力値 θ および項目識別力 a の場合と同様，計算回数が増すごとに RMSE が小さくなっていることが分かる．

これらの結果をもとに，計算回数を増すごとに RMSE が小さくなり，推定精度が良くなっている様子を図 2.5 に示す．ここでは，計算回数を 30000 回まで行ったが，10000 回

から先の変化は 10000 回までの変化に比べてわずかであることが分かる．よってここでは 30000 回も計算を行えば十分であると考えられるだろう．

MBE および BEAP と MCMC の比較

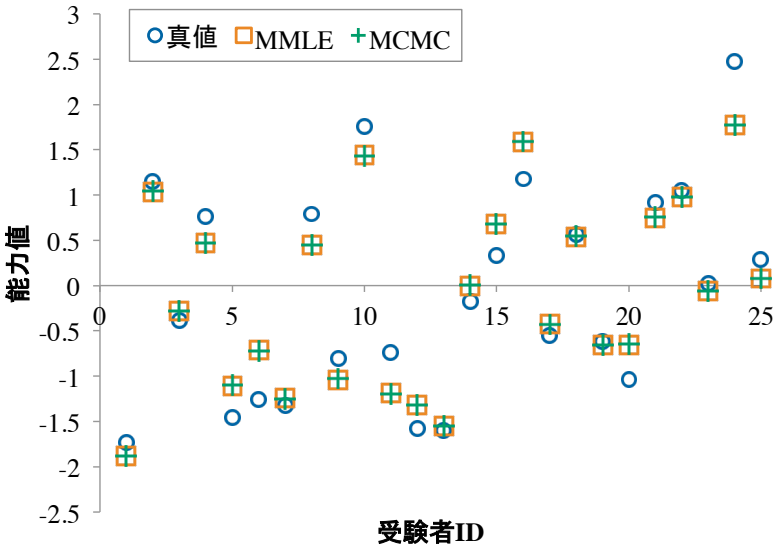
MBE および BEAP と MCMC の比較結果について，能力値 θ の比較を図 2.6 に示す．図中青丸は，シミュレーションデータを生成する上で用いた真のパラメータ値を表す．MBE および BEAP はベイズ理論を用いているものの枠組みとしては最尤推定法に属するので，ここでは MMLE として図中橙四角で表している．MCMC は図中緑十字で表している．図 2.6(a) を見ると，両手法とも真値からわずかにずれているが，そのずれかたが同程度であることが分かる．図 2.6(b) を見ると，両手法の推定値はほぼ同じ値を示しており，どちらも真値との相関が高いことが分かる．

次に，MBE および BEAP と MCMC の比較結果について，項目識別力 a の比較を図 2.7 に示す．図中青丸は，シミュレーションデータを生成する上で用いた真のパラメータ値を表す．MBE および BEAP は MMLE として図中橙四角で表している．MCMC は図中緑十字で表している．図 2.7(a) を見ると，両手法とも項目識別力が大きかった項目についてのずれが大きく，それ以外はほぼ真値と同じ値を示していることが分かる．図 2.7(b) を見ると，両手法とも真値との相関が高いことが分かる．

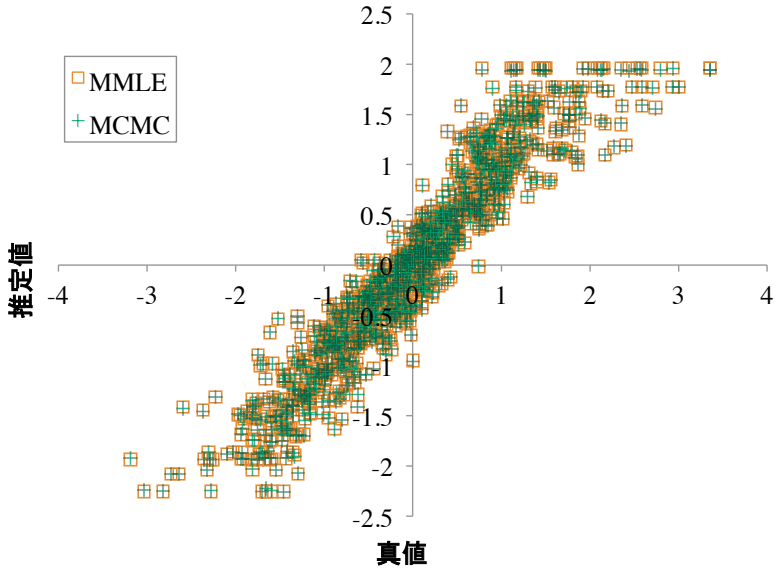
MBE および BEAP と MCMC の比較結果について，項目困難度 b の比較を図 2.8 に示す．図中青丸は，シミュレーションデータを生成する上で用いた真のパラメータ値を表す．MBE および BEAP は MMLE として図中橙四角で表している．MCMC は図中緑十字で表している．図 2.8(a) を見ると，両手法ともほぼ真値と同じ値を示していることが分かる．図 2.8(b) を見ると，両手法とも真値との相関が高いことが分かる．

また，$RMSE$ を比較した結果を表 2.7 に示す．θ と a に関しては MCMC のほうが $RMSE$ が小さく，b に関しては MBE のほうが $RMSE$ が小さいことが分かる．両手法とも $RMSE$ が十分に小さいことから，正しく推定できていることが分かる．

MCMC は単純な方法ながら精度が良く，複雑な計算が不要なことから，多次元で複雑な分布などの解析的には解けない場合に有効な可能性がある．ただし，問題点として，提案分布の取り方によって推定結果に違いが生じることや，計算コストが高いことなどが

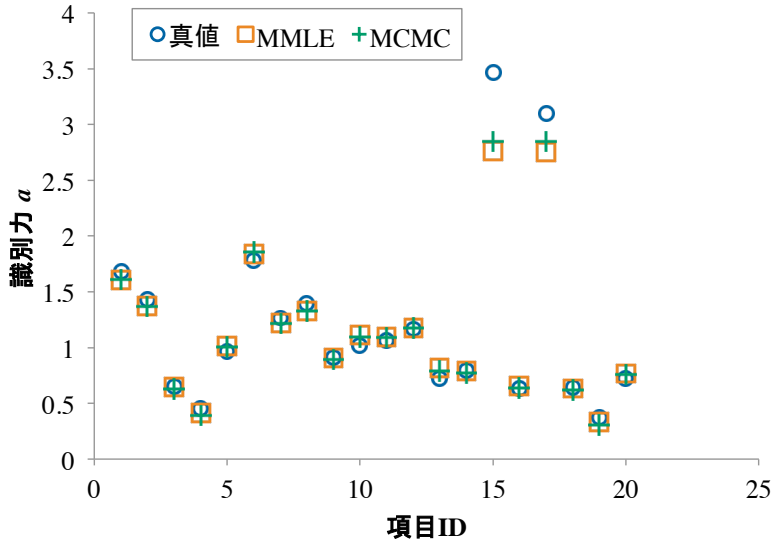


(a) 比較

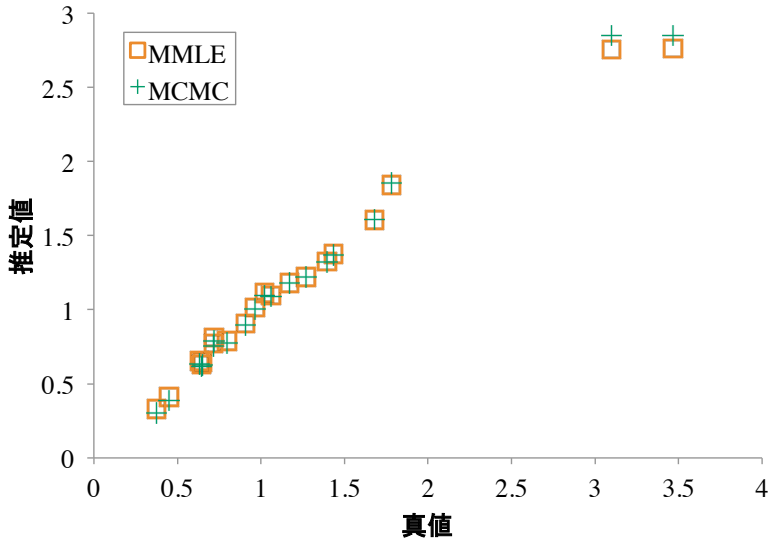


(b) 相関

図 2.6: 各推定手法による能力値 θ の比較 [31]

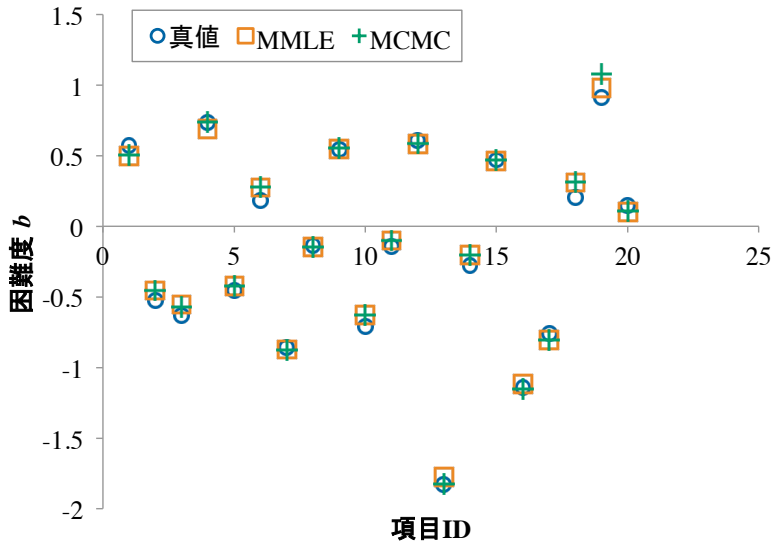


(a) 比較

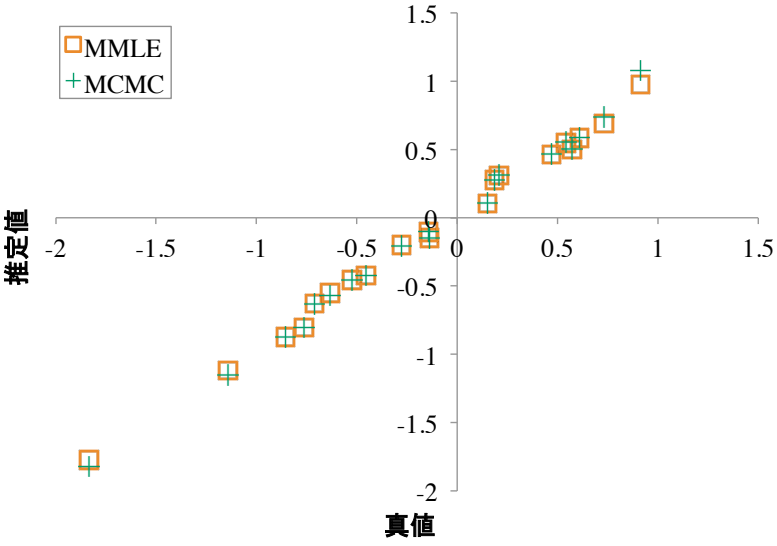


(b) 相関

図 2.7: 各推定手法による識別力 a の比較 [31]



(a) 比較



(b) 相関

図 2.8: 各推定手法による困難度 b の比較 [31]

表 2.7: MBE および BEAP と MCMC の $RMSE$ における比較 [31]

RMSE	MCMC	MBE および BEAP
θ	0.0045	0.0051
a	0.0024	0.0496
b	0.042	0.0241

あった．解決策としては，提案分布そのものも推定することや，適当な標本のみを抽出することなどが挙げられる．一方，MBE および BEAP を用いる方法では，計算が複雑であるため実装コストが高いものの，計算コストは非常に低く，また推定精度も良い．よって，事前分布に関して極端に母集団特性分布から逸脱するようなものを選ばない限り，MBE および BEAP を用いるほうが実用的であると考えられる．

2.4 本章のまとめ

本章では，いまだ多くの場面で利用されている古典的テスト理論が抱える問題点を指摘し，それを克服した現代テスト理論としての IRT の概念を説明した．また，試験結果として観測される回答結果は，従来は 0/1 の二値の値を取るが，[0,1] の有理数に拡張できることを述べた．また，IRT のパラメータ推定法としてこれまで考案されてきた 3 つの方法について説明するとともに，各手法における特徴について，シミュレーションによる比較研究を行った結果を述べた．

IRT は，受験者と項目を独立なものとして扱う．そのため，試験を受験した集団や，または時間や場所などの環境的な要因によって不公平な評価となる問題を解決する理論として有効である．そのため，より公正で公平な評価法として，TOEFL などの公的なテストに活用されている．IRT は，受験者が項目に正答する確率がロジスティック分布関数に従うと仮定して，試験結果を表す完全マトリクスから，受験者の能力と項目の特性を推定する．このとき，受験者能力と項目特性を同時に推定することによる困難さを伴うため，これを解決する推定方法として，周辺化，EM アルゴリズム，ベイズ理論，マルコフ連鎖モンテカルロ法などといった手法を援用しなければ推定値が得られない．もっとも安定した

推定値が得られるのは，マルコフ連鎖モンテカルロ法を用いた方法であるが，多くの計算時間がかかるという問題がある．現実的な方法としては，周辺ベイズ推定法と EM アルゴリズムによって項目特性を推定した後に，改めて期待事後推定法によって能力値を得る方法が推奨される．

第 3 章

項目反応理論を用いたテスト評価の Web システムの開発

これまで述べてきたように，古典的テスト法に対して IRT を用いたテスト法の優位性は，大学での中間テストや期末テストのような標準的な試験に対して明らかであるが，IRT は大学の教員らに知られていない．その理由の一つに，IRT 分析ソフトとして有名な BILOG-MG [3] の操作が難しいことが考えられる．そこで，本章ではこの新しいテスト法の使用機会をより良くするために，大学教員に対する Web を通したテスト評価システムの開発について，その概要を説明する．このシステムの特徴は，Excel ファイルに入力したテスト結果の [0,1] スコアをシステム上にドラッグ・アンド・ドロップすることで，教員らは Excel ファイルに付随された学生の能力とそれぞれの問題項目のパラメータを得ることである．テストに IRT 評価を用いることで，1) 高いレベルと低いレベルの両方を含んだ問題を同一テストに混在させられるため，学生の能力をより公正・公平に評価できる．言い換えれば，最適な問題レベルをそれぞれの学生に与えることができる．2) 学生から成績についての不平が無くなり，評価に満足するなどといった効果を期待できる．

3.1 開発の背景

これまでの研究で，大学新入生への数学テストを一斉に実施した結果を，項目反応理論（IRT: item response theory）および素点で評価し，両者を比較することで，大学基礎数学に対する IRT 評価法の有効性が確認された [33–35]．また，学内で標準的に使われている e-learning システム Moodle に，IRT 評価を組み込んだ小テストを実装し，学生の取り組み姿勢の変化を観てきた [37, 38].

古典的テスト理論に比べ，IRT を用いたテスト法は，中間試験や期末試験のような標準的な試験に対して有効であるが，大学教員にはあまり知られていない．それにはいくつかの理由が考えられる．1）`BILOG-MG` [3] のように IRT を用いたテスト評価プログラムは存在しているが，その操作は困難であること，2）EM アルゴリズムやベイズ法における未知のチューニングパラメータによって，推定結果にわずかな違いが生じること，3) 多くの教員が 0/1 による評価に慣れていないことなどが挙げられる．

そこで，これらの問題点を解決し，大学教員が IRT を使用する機会をより良くするために，Web を利用した IRT によるテスト評価システムを開発した [23].

3.2 Web を利用したテスト評価システム

3.2.1 規格

本開発システムの規格は，1）教員は正解不正解を [0,1] で表したテスト結果を Excel で作成する．2）Web 上のソフトに Excel ファイルをドラッグ・アンド・ドロップすることによって，推定結果をその Excel ファイルに返すというものである．また，`Java` 言語によって実装しており，そのため OS に依存しない．この概念図を図 3.1 に示す.

3.2.2 利用方法

利用者が Web サーバ上に配した本システムにアクセスすると，利用者のクライアント PC 上で本システムが実行される．最初に GUI ウィンドウが立ち上がる．この GUI 上

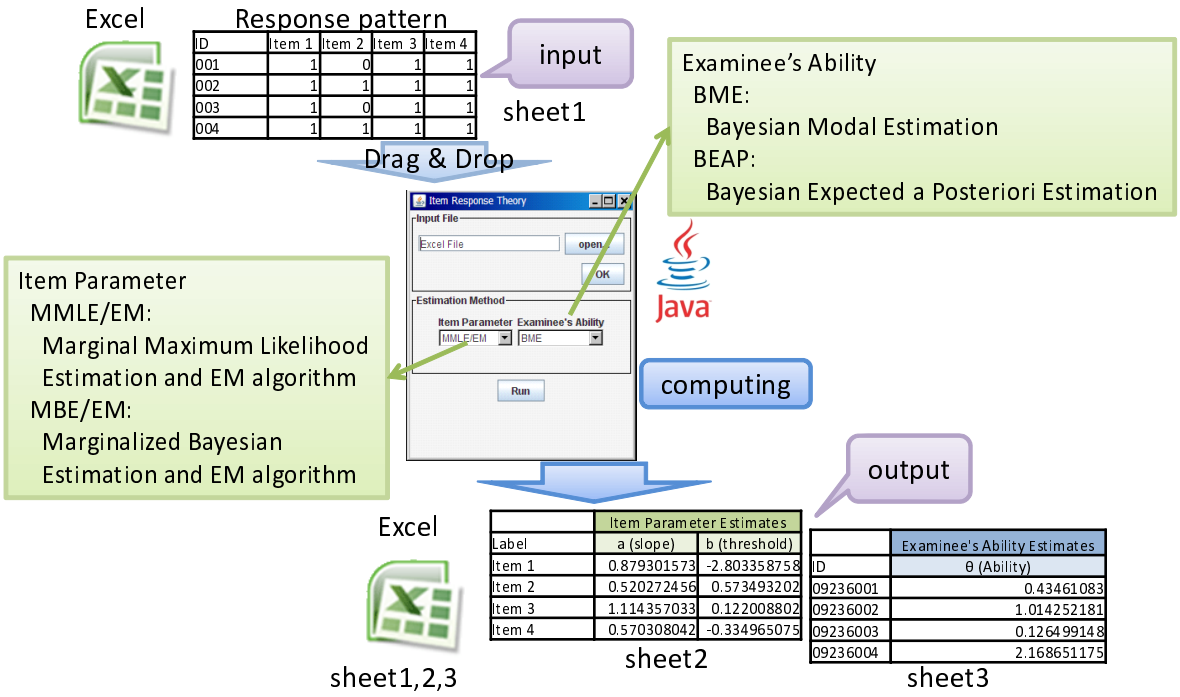


図 3.1: IRT を使った Web システムの概念図 [23]

に，あらかじめテスト結果を [0,1] で記載した Excel ファイルをドラッグ・アンド・ドロップすることで，テスト結果が入力される．次に，項目パラメータおよび能力パラメータの推定方法をリストから選択する．最後に Run ボタンを押すことで，計算が開始される．計算が終了すると，Excel ファイルに結果が出力される．

3.2.3 出力ファイル

このとき，入力および出力に利用される Excel ファイルの例を図 3.2，図 3.3 に示す．入力ファイルは，シートの名前を「pattern」として，第 1 列目に受験者の名前が記載する．第 1 行目にテストの項目名が記載する．よって，$i+1$ 行 $j+1$ 列目に，受験者 i が項目 j に回答した結果 $\delta_{i,j}$ を記載する．このときの回答結果の記載では，正答を 1，誤答を 0 とする．また，ここでは δ に対して有理数に拡張しているため，たとえば部分点のように 0.5 のような記載でも良い．

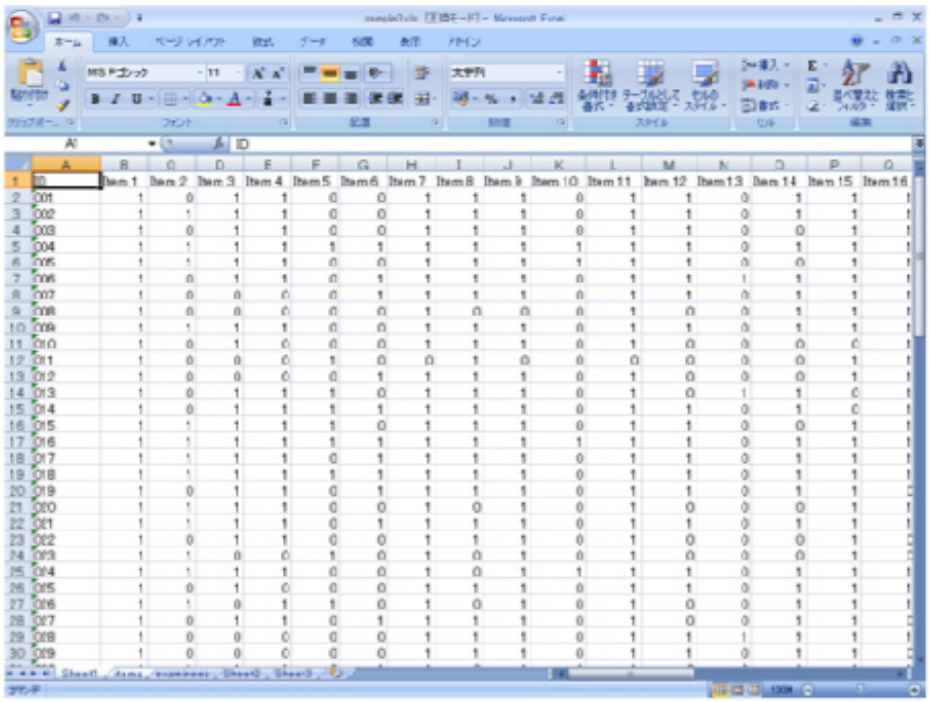


図 3.2: 入力ファイルの例 [23]

出力ファイルは，「pattern」，「examinee」，「item」，「info」，「summary」の 5 つのシートから構成される．「pattern」は，もとの入力データである回答結果 δ である．「examinee」は，各受験者の能力パラメータの推定結果が記載されている．「item」は，各項目の項目パラメータの推定結果が記載されている．「info」は，入力されたテストの，受験者能力に対する情報量が記載されている．「summary」は，テストの基本統計量が記載されている．各シートに分けて出力されるため，管理が容易であり，可読性も高い．

また，求めた項目パラメータから作成される項目応答曲線の例を図 3.4 に示す．このような図は「item」シートに添付されており，他にもテスト特性曲線，テスト情報曲線，項目情報量などの図が対応するシートに記載されている．数値としてだけではなく，図としても表示することで，テストの性質や受験者の能力推定結果を瞬時に知ることができる．

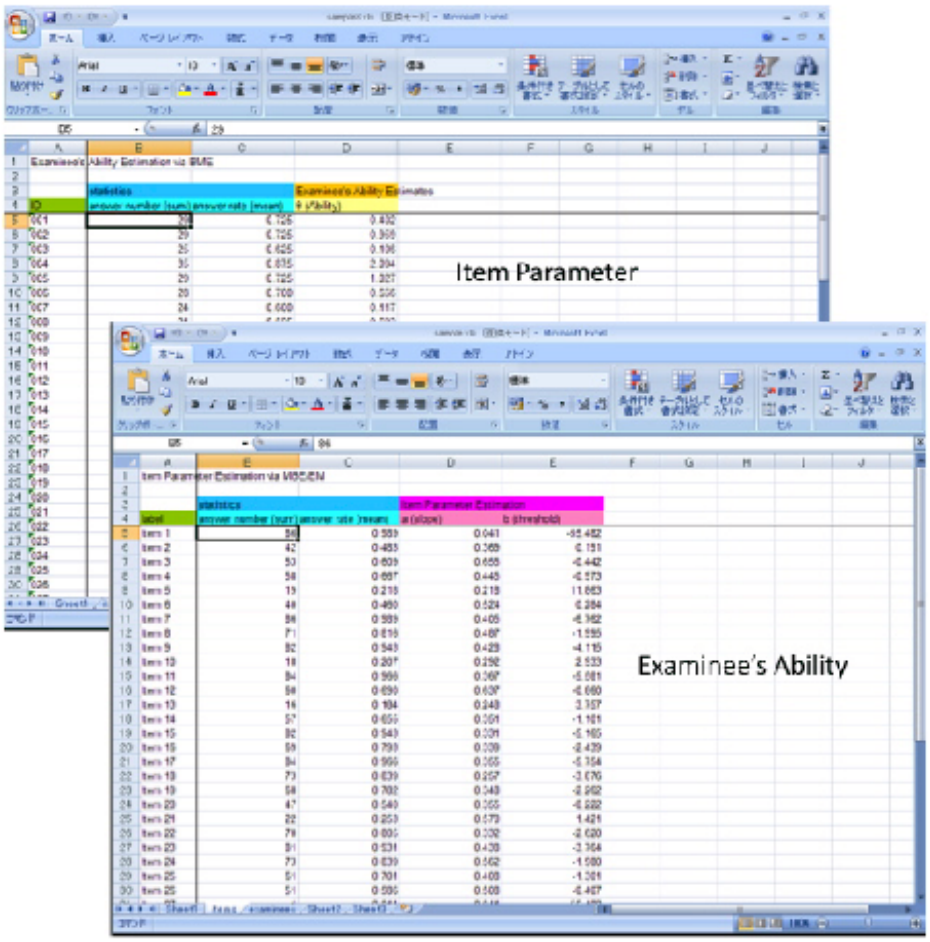


図 3.3: 出力ファイルの例 [23]

3.2.4 本開発システムの効果

本システムによる評価を実際に利用してもらい，いくつか分かったことがある．1）学生から成績についてのクレームがなくなったこと，2）低レベル，高レベル両方を含んだ問題が作れて，スキルの高い学生に対応することができたこと，3）60 点以下の学生が多いときの再試験を行うことがなくなったことなどである．

従来の古典的テスト理論では，配点方式による評価が主に行われてきたが，公正な評価のための配点の決定は困難さが伴い，また受験者のレベルに左右されやすかった．本シス

problem	1.1,	1.2,	2.1,	2.2,	2.3,	2.4,	3.1,	3.2,	3.3,	4.1,	4.2,	4.3,	4.4
correct ans.	18,	9,	83,	23,	42,	21,	8,	6,	26,	62,	5,	56,	31
slope	0.7965	1.0239	1.6765	1.0533	1.2561	1.3568	0.95502	1.1708	0.5879	0.80878	0.95188	0.9102	1.4043
threshold	1.5233	2.1374	−0.7436	1.3997	0.5224	1.2382	2.2629	2.3597	1.2343	−0.2538	2.44004	−0.07849	0.8131

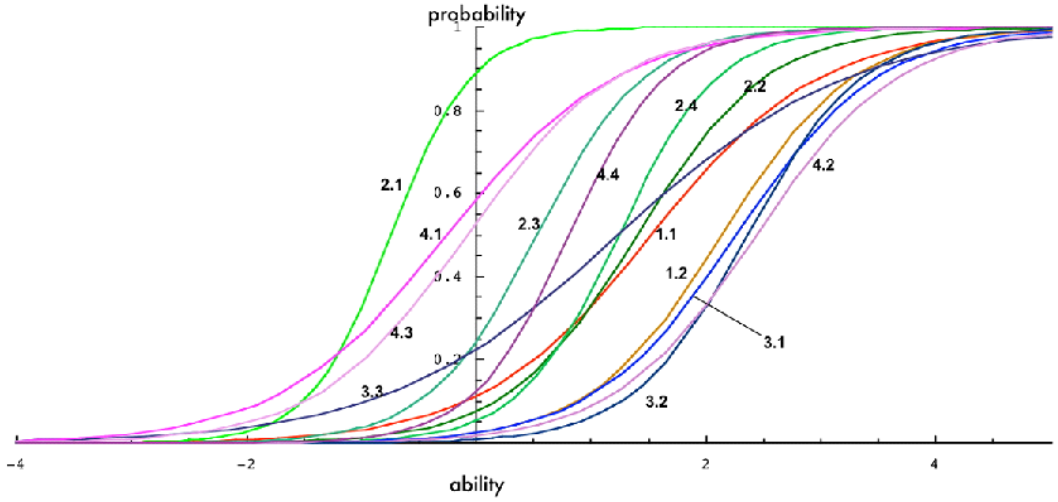


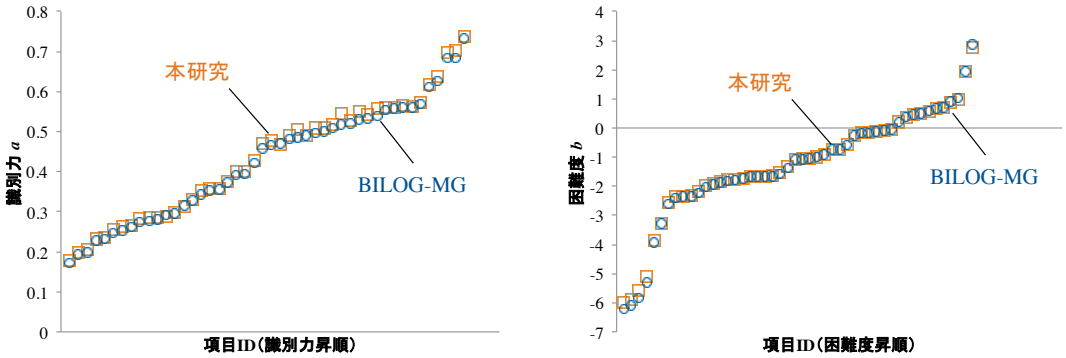
図 3.4: 項目応答曲線の例 [23]

テムによって IRT による評価を行うことで，理にかなった公正な評価を行うことが可能となる．

3.3 BILOG-MG との推定結果の比較

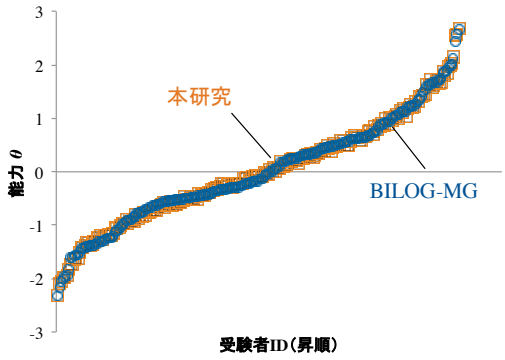
ここでは，世界で最もポピュラーな IRT 分析ソフトである BILOG-MG [3] との推定結果を比較することで，本研究で開発した Web システムの推定精度を確認する．

比較に用いた試験データは，[40] に記載されている学力テスト 1 および学力テスト 2 を利用する．学力テスト 1 は受験者数 226 人，問題項目数 50 問である．学力テスト 2 は受験者数 104 人，問題項目数 50 問である．推定に用いた手法は，項目パラメータに関しては周辺ベイズ推定法を，能力パラメータに関しては BEAP 推定を用いる．



(a) 識別力 a_j

(b) 困難度 b_j の比較

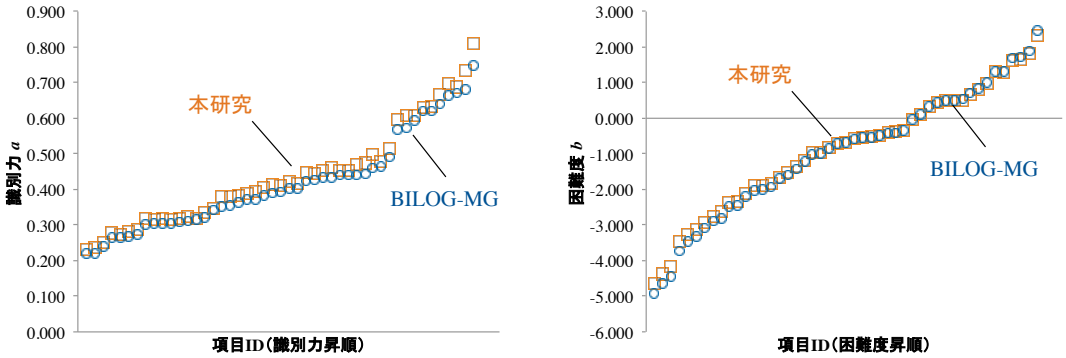


(c) 能力値 θ_i の比較

図 3.5: 学力テスト 1 による BILOG-MG と本研究で開発したツールの比較

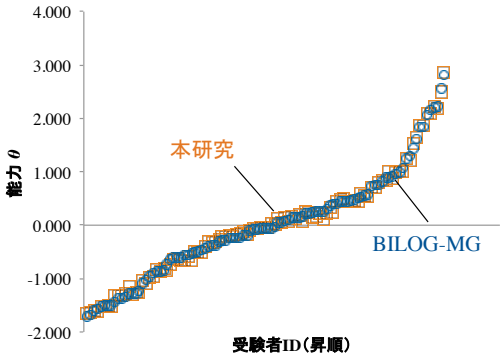
3.3.1 学力テスト 1 における比較

まず学力テスト 1 の比較結果について図 3.5 に示す．横軸に受験者および項目の ID を，縦軸に各特性値を示している．また，図中の丸が BILOG-MG の推定結果，四角が本システムの推定結果を示している．また，それぞれの特性値は BILOG-MG による推定値をもとに昇順にソートしている．図 3.5(a) を見ると，いくつかの項目で両ツールの推定値に差異が生じているものの，多くの項目では重なって表示されていることから，概ね一致していることが分かる．図 3.5(b) を見ると，どの項目もほとんど重なっており，識別力よりも一致度が高いことが分かる．図 3.5(c) を見ると，能力推定値はほぼ全体的には概ね一致していることが分かる．以上，どの推定結果も非常に近い値をとっていることが分かる．



(a) 識別力 a_j

(b) 困難度 b_j の比較



(c) 能力値 θ_i の比較

図 3.6: 学力テスト 2 による BILOG-MG と本研究で開発したツールの比較

3.3.2 学力テスト 2 における比較

次に，学力テスト 2 の比較結果について図 3.6 に示す．ここでも図 3.5 と同じく，横軸に受験者および項目の ID を，縦軸に各特性値を示している．また，図中の丸が BILOG-MG の推定結果，四角が本システムの推定結果を示している．また，それぞれの特性値を BILOG-MG による推定値をもとに昇順にソートしている．図 3.6(a) を見ると，項目全体で両ツールの推定値にわずかな差異が生じていることが分かる．図 3.6(b) を見ると，困難度が低い項目においてわずかに差異があるものの，ほとんどの項目の推定結果は重なっており，識別力よりも一致度が高いことが分かる．図 3.6(c) を見ると，能力値が 0 付近の能力推定値にわずかな差異があるが，全体的な傾向は概ね一致していることが分かる．以

表 3.1: BILOG-MG との推定値の $RMSE$

	a_j	b_j	θ_i
学力テスト 1	0.00988	0.0615	0.0564
学力テスト 2	0.0214	0.110	0.0595

上，どの推定結果も非常に近い値をとっていることが分かる．

3.3.3 $RMSE$ による評価

各パラメータ a_j, b_j, θ_i について $RMSE$（平均二乗誤差の平方根）によって定量的に評価すると，表 3.1 を得る．どのパラメータ推定値に関しても $RMSE$ は小さく，BILOG-MG との大きな違いはなかったことが分かる．したがって，本研究で開発した Web システムは十分な精度で推定できていると考えられる．

3.4 本章のまとめ

本章では，IRT に触れる機会をより良くするために，Web を利用したシステムを開発したことについて述べた．また，一般に利用される有償のソフトウェア BILOG-MG とのパラメータ精度の比較を行った．

IRT は，教育現場などでより活用されるべき評価法であるが，その理論を理解するためには，高度な確率論と統計学の知識を必要とする．BILOG-MG は世界的に有名な IRT ツールであるが，その操作には IRT の専門的な知識を必要とする．

一方，本研究で開発した Web システムは，IRT の専門的知識が無くても，利用可能なツールである．本システムは，インターネットを通じて，クライアント側の PC 上で動作する．利用者は，テスト結果である 0/1 スコアを書いた Microsoft Office Excel ファイルを，本システムにドラッグ・アンド・ドロップするだけで，テスト受験者の能力や各問題のパラメータを求めることができる．誰もがより簡単な操作で IRT に触れることができるものを目指したアプリケーションである．また，その計算精度は BILOG-MG とごく簡

単な操作のみで IRT による推定を試みることが可能であり，IRT の理解を補助するという意味で有効なものである．

第 II 部

ユーザー・アイテムの応答から構成された確率構造を持つ不完全マトリクスからの完全マトリクス推定

第 4 章

適応型試験

本章では，適応型試験について，その概要に触れる．適応型試験とは，受験者一人ひとりの能力に合わせた試験を意味する．この試験の概念そのものは古くからあったが，受験者の能力に合わせた個別試験問題を作成することが困難であることや，受験者の能力およびそれに合わせた問題レベルの定義が明確でなかったことなどを理由に，実現には至っていなかった．近年，現代テスト理論として項目反応理論（IRT; item response theory）が確立し，潜在的な受験者能力および項目特性が定量的に評価できるようになってきたことや，情報インフラが整備されてきたことなどを背景に，適応型試験が現実に利用される場面が現れてきている．適応型試験には情報端末が不可欠であり，また事前準備として項目バンクを持つことが前提条件である．そこで，実際に適応型試験を構築する際の手順および運用方法について述べる．

4.1　適応型試験の概要

4.1.1　IRT を利用したオンライン試験

項目反応理論（IRT）は，あるテストの問題を受験者が解答したときの解答パターンのマトリクスから，テストの問題項目特性および受験者能力評価を推定することができる [7,9,10,27]．通常はすべての問題に全員が解答している完全マトリクスを用いる．IRT にはいくつかのツールが考案されている．もっとも一般的なツールに [3] があるが，IRT

の専門知識が必要であり取り扱いが難しいという問題があった．そこで，[23] では，[0,1] 表記で表したテスト結果の EXCEL ファイルを単にドラッグ・アンド・ドロップすることで，IRT のパラメータ推定を可能にするシステムを開発した．また，少ない問題数を用いたときの能力評価の精度についての研究も行われている [13].

問題のパラメータがあらかじめ分かっていれば，IRT を利用した適応型オンライン能力評価システム（以下，適応型システムと呼ぶ）を構築することができる [16,24]．これは，受験者の能力に合わせたレベルの問題を自動的に出題するシステムである．出題される問題の項目群を項目バンクという．受験者の解答がシステムに取り込まれる毎にその受験者の能力を逐次評価することができる．出題する問題は受験者の能力に最適に合わせることができるため，より少ない問題で能力評価の精度を高めることができる．適応型システムには昇降法やストレス・ストレングスモデルと昇降法を組み合わせた方法 [11] もあるが，いずれも出題される問題の項目特性は分かっている必要があった．そのためには，事前に予備テストを実施するのが一般的である．予備テストとは，項目バンクに項目を登録する際に実施される試験である．

4.1.2 適応型試験の流れ

適応型試験では，受験者一人ひとりの能力に合わせた試験を行う．受験者の能力評価および回答ごとの問題項目の選出には，IRT による特性値を利用できる．たとえば，受験者が情報端末などを介して適応型試験を受験したとき，もし最初の問題が正答であれば，次の問題は，それ以前の回答結果から得られる能力評価に合わせた，より難しい問題が選出される．誤答であれば，より易しい問題が選出される．

この試験の運用には，項目バンクと呼ばれるデータベースが必須である．項目バンクとは，IRT によってすでに推定済みの項目特性値を持つ項目群を意味する．これらの項目特性値は，事前に予備テストなどを実施することで試験データを集め，そこから推定しておく必要がある．予備テストとは，項目バンクに項目を登録する際に実施される試験であり，その問題項目は，項目バンクに追加する項目群から構成される．予備テストを受験する受験者集団は，適応型試験を受験する集団とは異なる必要がある．予備テストの回答結

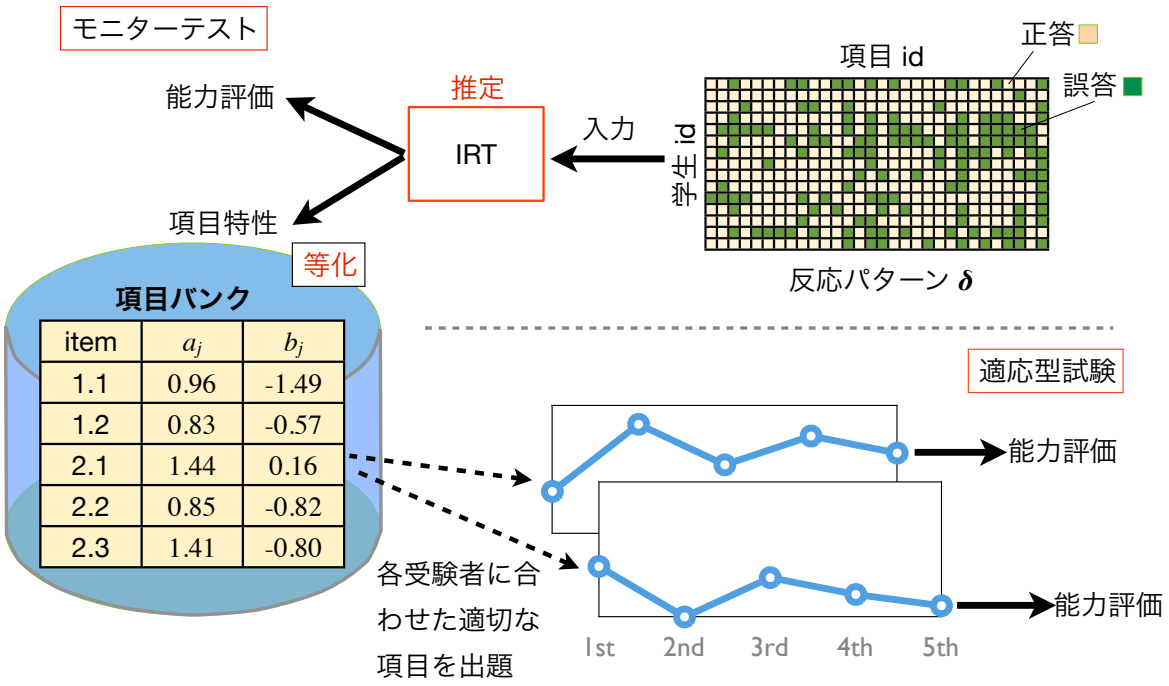


図 4.1: 適応型試験の流れ

果から問題項目特性を推定し，項目バンクに追加する．

適応型試験に既知の項目特性値を用いることで，この試験の受験者は，一問回答するごとに，式 (2.2) の尤度関数と項目特性値をもとに，能力値が求められ，そのときの受験者の能力値レベルに最も一致する項目を次の項目として提出することができる．項目バンクに登録された項目群の特性が多様であるほど，能力に適した項目が提出されやすくなり，より少ない項目数で精確な能力評価が可能となる．図 4.1 に適応型試験の流れを示す．

4.1.3 問題項目の選出方法

適応型試験における能力評価の精度を大きく左右するのは，問題項目の選出方法である．適応型試験では受験者の能力に合わせた問題項目が選出されると述べてきたが，受験者の能力に合った問題項目とは何であろうか．その一つの方法として，最大情報量選択法が挙げられる．

能力に関する情報量 $I(\theta)$ は，式 (2.2) および式 (2.6) から，フィッシャー情報量を求めることで得られる．すなわち，

$$I(\theta) = -E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = \sum_{j=1}^{n} D^2 a_j^2 P(\theta_i, a_j, b_j) Q(\theta_i, a_j, b_j) \tag{4.1}$$

である．このとき右辺は，項目一つひとつが持つ情報量の和の形で表される．すなわち，項目の持つ情報量を $I_j(\theta)$ として，

$$I_j(\theta) = D^2 a_j^2 P(\theta_i, a_j, b_j) Q(\theta_i, a_j, b_j) \tag{4.2}$$

である．これを最大とするのは $P(\theta_i, a_j, b_j) = Q(\theta_i, a_j, b_j) = 0.5$ のとき，かつ a_j が大きいときであることが分かる．式 (2.6) から，

$$P(\theta_i, a_j, b_j) = \frac{1}{1 + \exp(-Da_j(\theta_i - b_j))} = 0.5 \tag{4.3}$$

を満たすのは，$\theta_i = b_j$ のときである．よって，項目選出法として最大情報量選択法を利用する場合は，推定された能力値 θ_i に対して，$\theta_i = b_j$ かつ a_j が大きい項目が選出される．

つまり，受験者の能力に合わせた問題項目とは，項目識別力 a_j を無視して考えると，受験者の能力値に一致する困難度を持つ問題項目であることが分かる．ただし，項目バンクに含まれる問題項目は，もちろんすべての項目特性を網羅しているわけではないことが通常である．そのため，一般的には受験者の能力値 θ_i にもっとも近い困難度 b_j を持つ問題項目を選択する．一方で，$P(\theta_i, a_j, b_j)$ が 0.5 にもっとも近い値を示す問題項目を選択することもできる．後者のほうが，より情報量に則した問題項目の選出になるが，項目バンクに含まれるすべての項目について $P(\theta_i, a_j, b_j)$ を計算する必要があるため，計算コストがかかる．適応型試験では，多数の受験者に対して，リアルタイムに問題項目を選出し，回答結果から能力推定を行う操作を繰り返すことを考慮すると，出来る限り計算コストは避けたい．そのため，ここでは，受験者の能力値 θ_i にもっとも近い困難度 b_j を持つ問題項目を選出する方法を推奨する．図 4.2 に適応型試験における問題提出とそのときの能力値の推移の概要を示す．

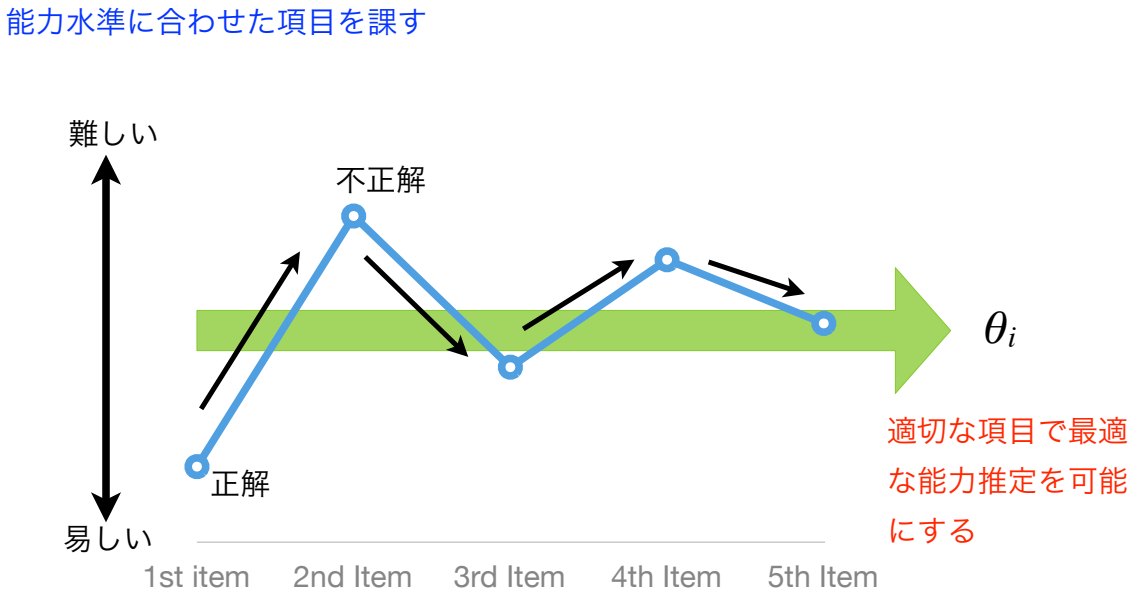


図 4.2: 適応型試験における問題選出

4.1.4 項目バンクの作成と等化

適応型試験では，受験者の能力に合わせたレベルの問題を出題するため，あらかじめ出題される問題の項目特性を知っておく必要がある．そのためには，事前に予備テスト（モニターテスト）を実施する．予備テストとは，項目バンクに項目を登録する際に実施される試験であり，その問題項目は，項目バンクに追加する項目群から構成される．予備テストを受験する受験者集団は，適応型試験を受験する集団とは異なる必要がある．予備テストの回答結果から問題項目特性を推定し，項目バンクに追加する．一般に，項目バンクの項目数が多ければ多いほど，受験者の能力の多様性に柔軟に対応することができる．

もし，すでに項目バンク上に既知の特性を持つ旧項目群が存在しており，そこに新規に項目群を追加したい場合は，予備テストの問題項目に旧項目群から数問を混ぜておくことで，旧項目群と新規項目群との特性値の尺度を統一させることができる．これを等化という．

等化による尺度変換には平均シグマ法 [6] などが利用できる．平均シグマ法による等化

図 4.3: 分冊法による等化

では，等化を行いたい項目群同士に共通項目を含めておく必要がある．

4.1.5 分冊法による予備テスト

予備テストの項目数が多すぎると，予備テスト受験者に対する負荷が大きくなり，正確な試験結果が得られにくくなるため，予備テスト項目数を少なくする工夫が必要である．

その一つに分冊法による等化法がある．分冊法とは，予備テストを複数の小冊子に分割し，それぞれに共通の項目群を付加して試験を行う等化法である．各小冊子の回答結果から得られるそれぞれの項目特性値は，共通項目の特性値をもとに共通尺度に等化される．図 4.3 に分冊法による等化法の概要を示す．

以上をまとめると，適応型試験を実施する上で必要になる事前準備は以下の流れになる．

1. 分冊法などによる予備テストを実施
2. 得られた回答結果をもとに IRT によって項目特性を推定

3. 共通尺度に等化

4. 項目バンクに登録

4.2 オンライン適応型試験システムの構築と実施

ここでは，実際に適応型試験をオンラインで動作するテストシステムとして構築する上での具体的な例を挙げる．

4.2.1 適応型オンライン能力評価システムの構築

適応型オンライン能力評価システムとは，インターネットを通して情報端末上で適応型試験を受験することで能力評価を行うシステムを意味する．適応型試験を行う上で，必要最低限の準備は，

1) IRT による能力推定プログラム，

2) 項目バンクとしてのデータベース，

3) 問題選出プログラム，

4) 情報端末上のユーザーインターフェース

の 4 つである．

1) に対しては，第 3 章で述べた Web アプリケーションシステムを援用することができる．能力推定法は，計算コストを考慮して，期待事後推定法（BEAP）を採用する．プログラミングは Java 言語を用いている．2) に対しては，MySQL を採用する．3) に対しては，PHP によって処理を行う．4) に対しては，Web ブラウザを介して HTML で記述したインターフェースとし，動的に HTML を処理するために，同じく PHP を採用する．

図 4.4 にその概要を示す．各受験者は情報端末から Web ブラウザを用いて本適応型システムのサーバーにアクセスする．各受験者はクライアント側の Web ブラウザ上で試験問題に回答する．試験問題の選出には，前述した問題項目選出法をサーバー側の PHP によって処理し，MySQL に保存された項目バンクから選択し，HTML を介して Web ブラウ

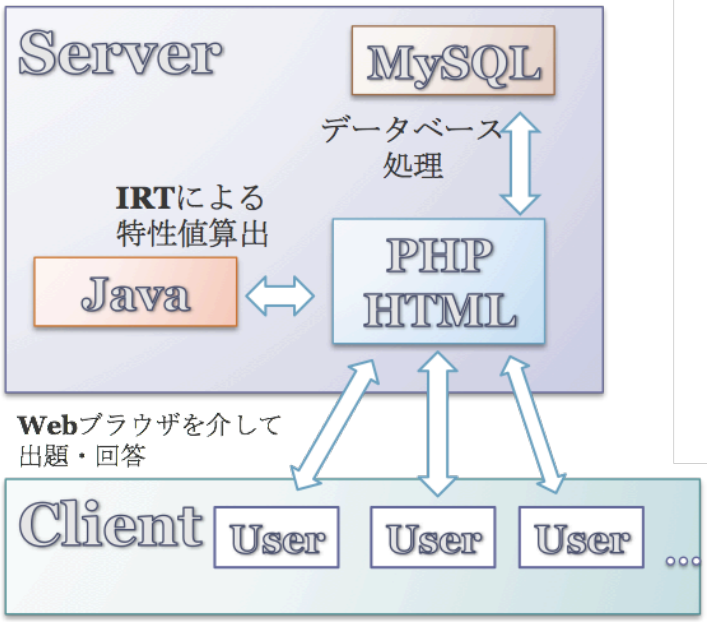


図 4.4: 適応型オンライン能力評価システムの仕様

ザ上に表示する．クライアント側の Web ブラウザ上で試験問題に回答した結果は，サーバー側に送られ，その回答結果をもとに PHP を介して Java によって能力値が推定される．能力推定値をもとに再び項目選出が行われる．各受験者はこのようなシステムを意識することなく，クライアント側の Web ブラウザ上で試験を受験することができる．

また，このとき図 4.5 に受験者が操作する Web ブラウザ上のインターフェースを示す．図中左上が本適応型システムのログイン画面を表している．各受験者はユーザー名を入力し試験を開始する．図中右上は試験問題のジャンルを選択するインターフェースを表している．ここでは 3 つのジャンルから選択する様子を表している．ジャンルを選択すると試験が開始され，試験問題が表示される（図中右下）．受験者は試験問題の回答を選択肢から選択する．回答結果をもとにサーバー側で能力値が推定され，それに合わせた項目が項目バンクから選出され，次の項目として表示される．指定回数の問題を回答すると，図下中央に示す試験終了画面に移行する．ここでは，回答結果とそれに合わせて推定されてきた能力値の遷移および最終的な能力値がランクとして表示される．また，回答結果のうち誤答した項目についての模範解答例などを確認することができる．これらの結果はすべて

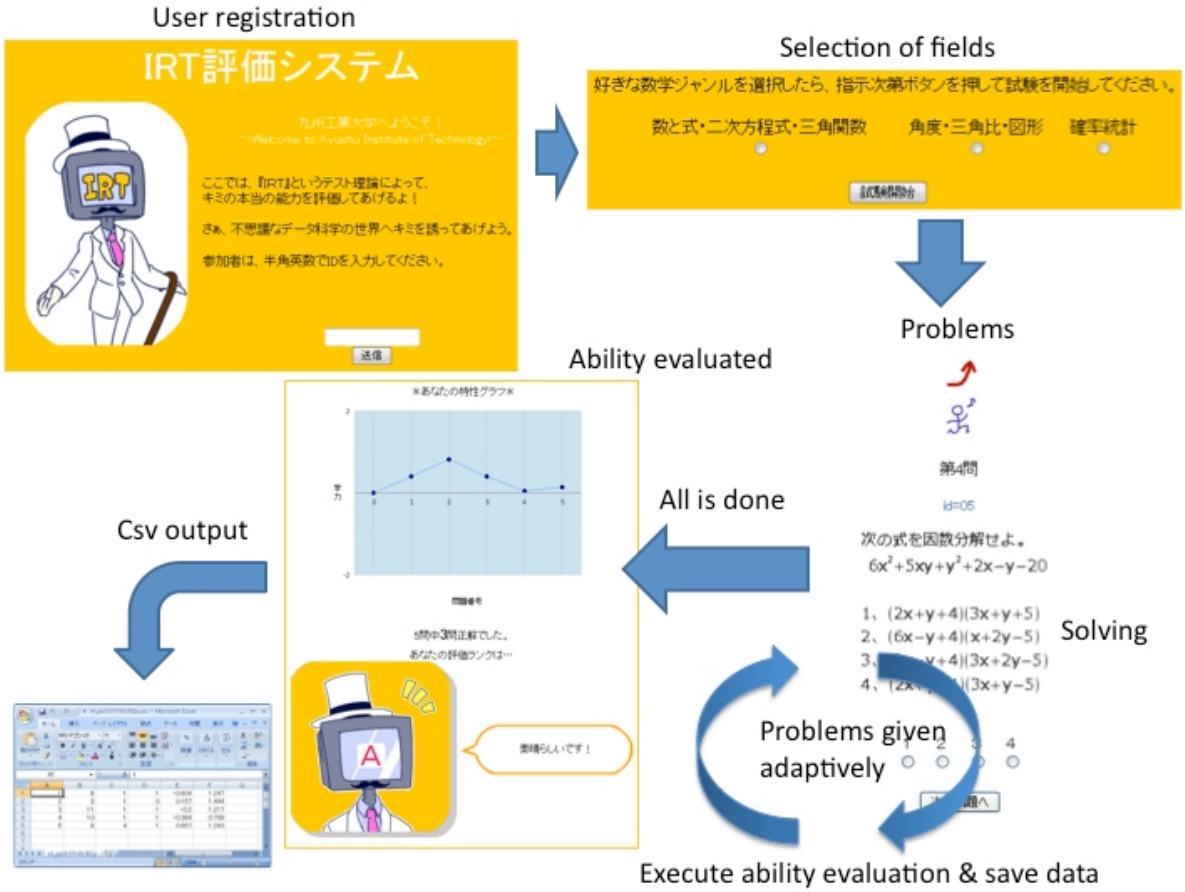


図 4.5: 適応型オンライン能力評価システムのユーザーインターフェース画面 [24]

CSV ファイルとして保管される．

4.2.2 項目バンクの拡充

ここでは，筆者らが 2011 年度に作成した項目バンクに，新たな問題項目群を追加する手順について説明する．すべて高校の初等数学レベルの問題である．2011 年度の項目バンクには 30 問の項目が登録されている．これを旧項目群と呼ぶことにする．この項目バンクを計 100 問程度とすることを目標とする．

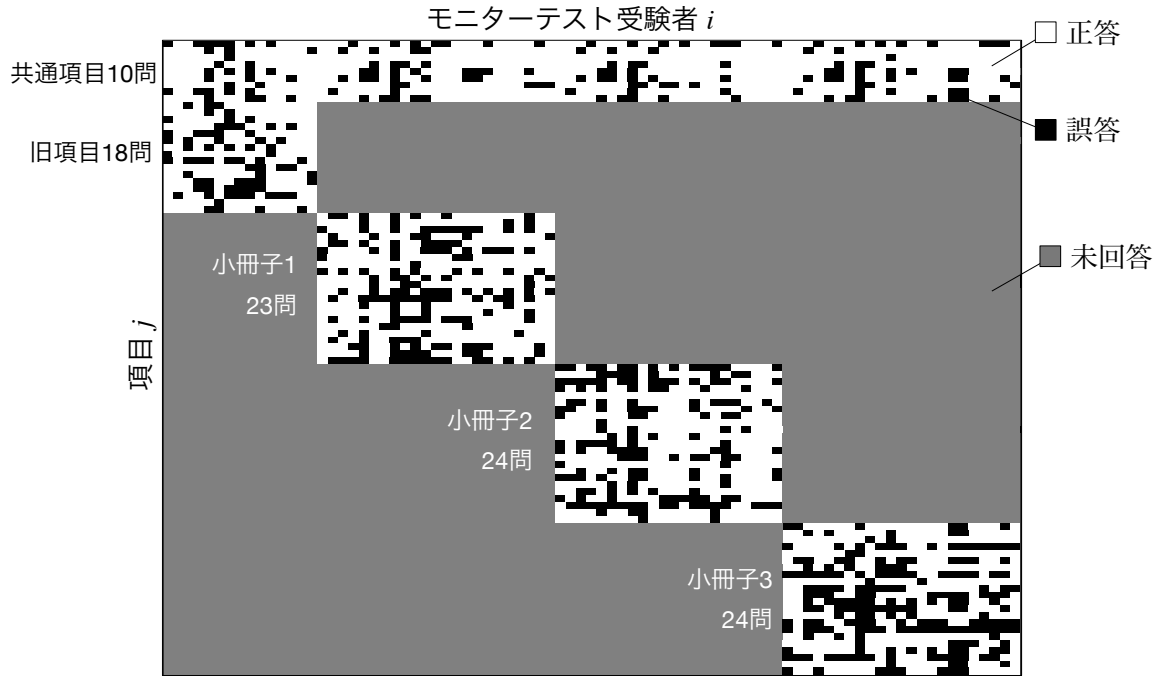


図 4.6: 分冊法による予備テストの回答結果 [30]

分冊法による予備テスト

ここでは，項目バンクへの新規項目群として 71 問を追加する．71 問の予備テストをすべて回答することは，受験者ひとりに対する負荷が大きく，得られる回答結果の精度が悪くなる恐れがある．そこで，分冊法による予備テストを行う．追加する 71 問を 3 つの小冊子に分割し，共通項目として旧項目群から 10 問選出する．各小冊子の項目数は，それぞれ 30 問程度である．受験者 68 人を対象に行い，得られる回答結果から IRT によって項目特性値を推定する．このときの予備テストの回答結果を図 4.6 に示す．マトリクスの各行が問題項目を表し，各列が受験者を表す．セルの色によって，正答であれば白色，誤答であれば黒色，未回答項目は灰色で表している．共通項目を含む各小冊子の回答結果に対して，IRT による推定を行うことで，それぞれの項目特性が与えられる．

等化

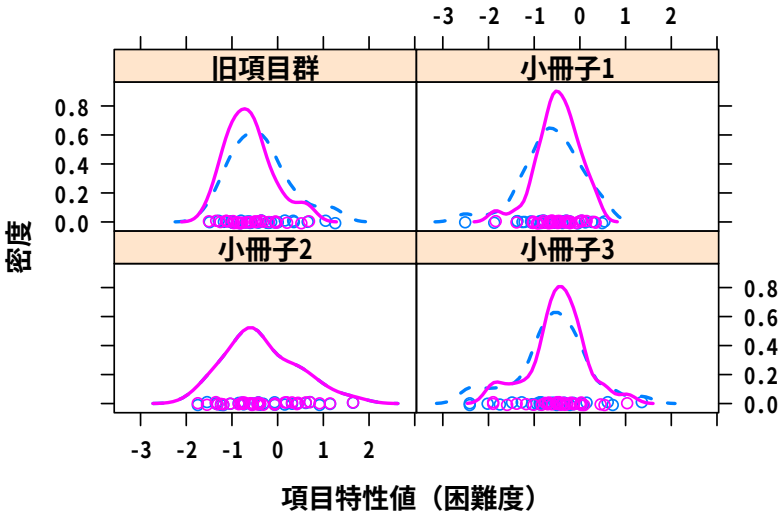
次に，小冊子ごとに求められた項目特性値に対し，共通項目の特性値にもとづいて等化を行い，共通尺度に変換する．このときの等化した結果を図 4.7 に示す．図中の点線がそれぞれの回答結果から求めた項目特性値の密度分布，実線が等化後の項目特性値の密度分布を示している．ここでは 小冊子 2 の項目特性値を基準に等化したため，小冊子 2 の項目特性値は等化前後で変化はない．図 4.7(a) から，全体的に等化前後での変化は小さいことから，小冊子間において項目困難度にそれほどの差はなかったと考えられる．図 4.7(b) から，旧項目群と小冊子 2 との間で，項目識別力に大きな違いがあったことが分かる．

このように等化を行うことによって，すべての項目特性値は共通尺度上に変換され，互いに比較可能なものとして扱うことができる．これら等化後の項目特性値を与えた項目群を項目バンクに登録する．

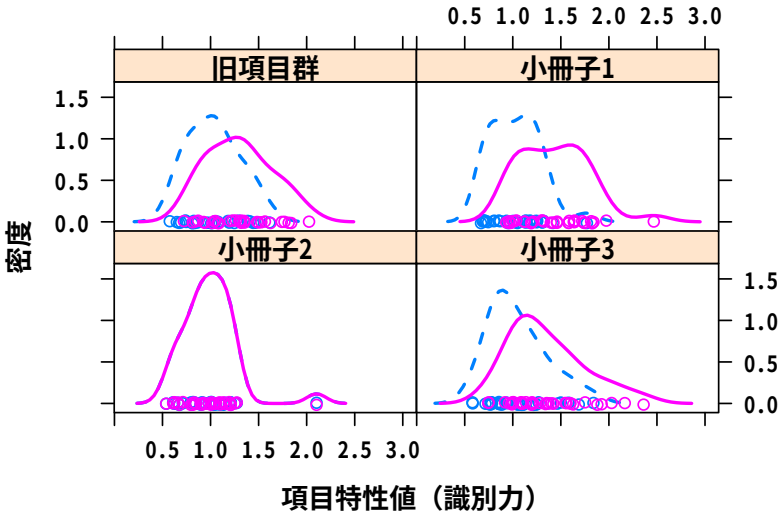
4.2.3 適応型オンライン能力評価システムの実施

作成した項目バンクに登録されている高校初等数学レベルの総問題数は，旧項目群と合わせて，101 問であったが，その内，問題にミスがあったものを 2 問除外して，合計 99 問であった．作成した項目バンクをもとに，適応型オンライン能力評価システムを高校生を対象に実施する．一人 5 問回答してもらう．受験者数は 138 人であった．このときの実際の試験の様子を図 4.8 に示す．各受験者は PC 上の Web ブラウザを介して本適応型システムを受験している．受験者は隣り合った PC で受験しているが，表示される問題項目は各受験者によって異なるため，たとえ隣の問題項目が見えてしまったとしても能力推定における影響は少ない．

このときの試験による能力値と出題される項目の項目困難度の変化の一例を図 4.9 に示す．横軸は回答順，縦軸は能力値を示している．図中の実線が能力値，点線が項目困難度を表す．各受験者に対する 1 問目は，項目バンクの中から無作為抽出している．その際，極端に難しい，あるいは易しい問題が出題されないように，平均的な問題からやや易しい項目を選出するように工夫している．能力値に近い項目困難度を持つ項目が出題されていることが分かる．



(a) 困難度の等化



(b) 識別力の等化

図 4.7: 項目特性値の等化 [32]

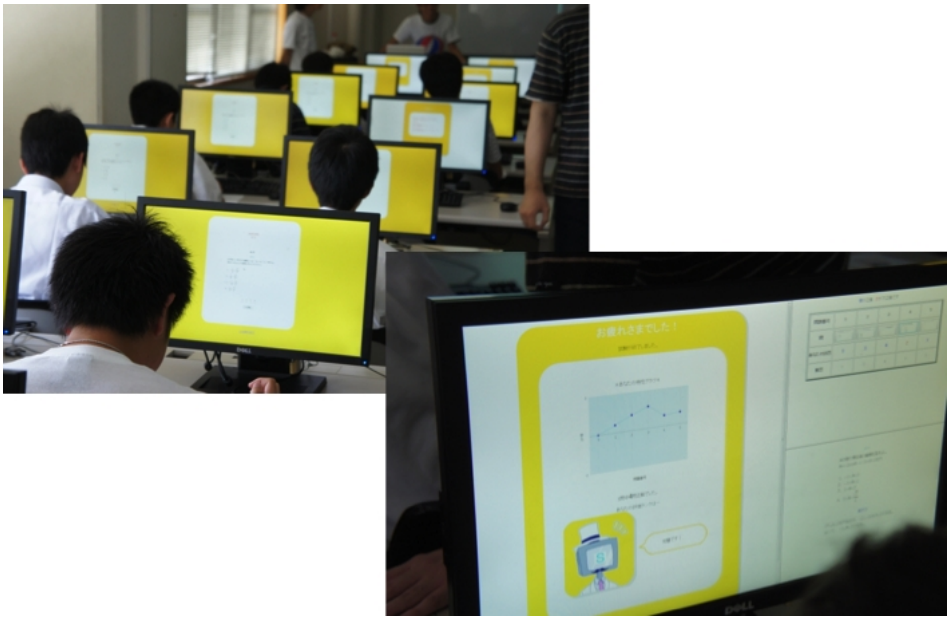


図 4.8: 適応型試験を受験する高校生の様子

図 4.10 に，適応型試験によって得られた全受験者の能力値が回答順に遷移している様子を示す．横軸は回答順，縦軸は能力値を示している．各線が受験者を表す．能力値にはさまざまな遷移の状況があることが分かる．

図 4.11 に 5 問終了時の全受験者の能力値の分布を示す．能力値のモードは 0 よりやや右側にあり，左側に裾が長い分布になっており，出題に対して能力値が高い受験者が多かったことが分かる．

各受験者は，それぞれ 5 問回答し，残り 94 問は未回答であるため，最終的に得られる項目および受験者全体のマトリクスは不完全マトリクスになる．その様子を図 5.5 に示す．図中の白色のセルが正答を表し，黒色が誤答を表す．灰色のセルは欠損していることを表す．

4.3 本章のまとめ

本章では，適応型試験の概要と運用方法を，実際に構築した経験を踏まえて説明した．適応型試験によるテスト法は，オンライン上で情報端末を利用した試験を行うことで，受

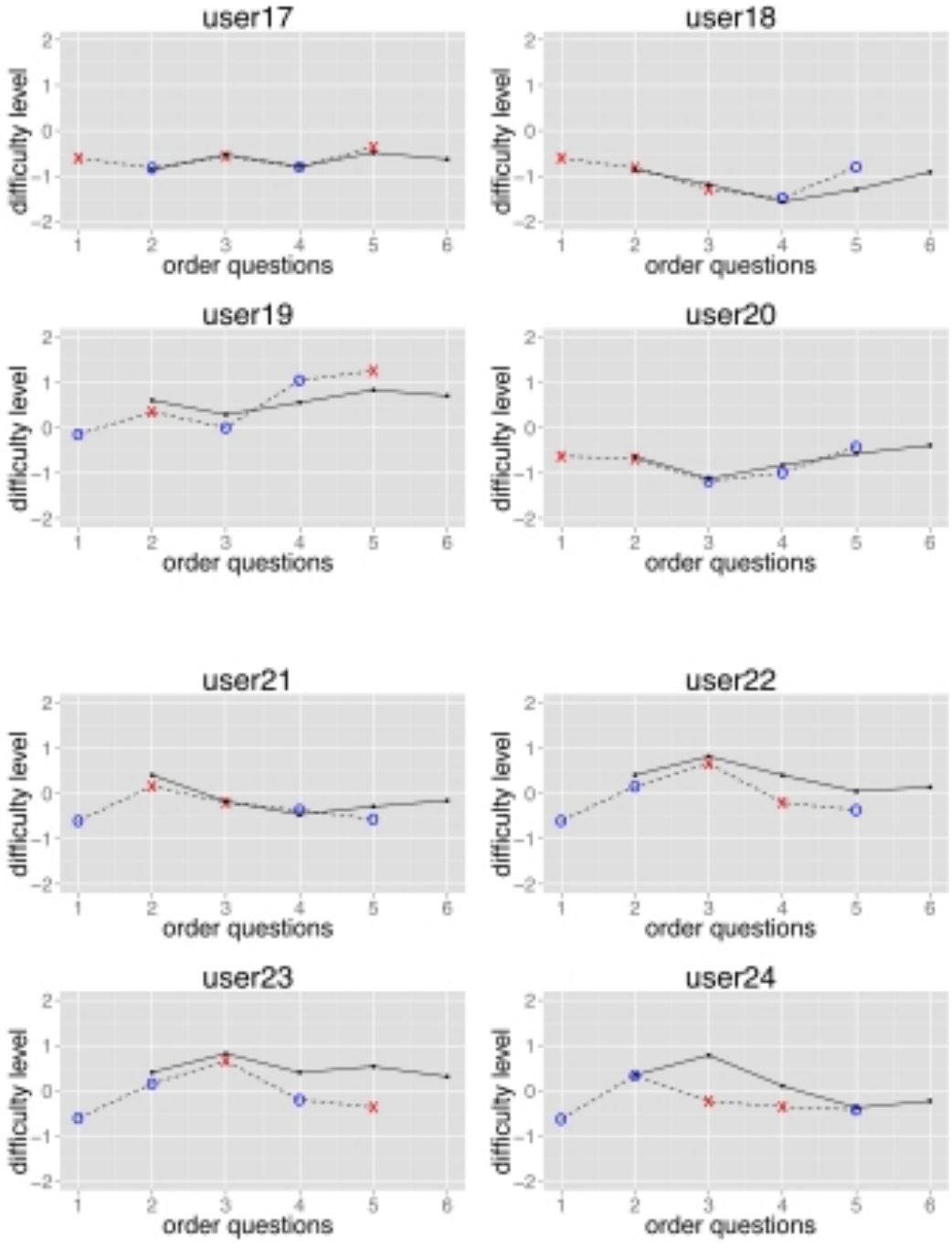


図 4.9: 各受験者の能力値と項目困難度

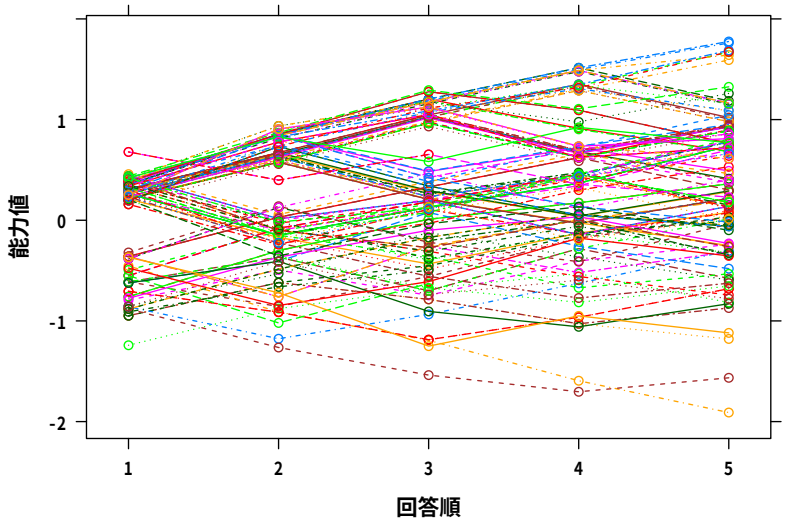


図 4.10: 能力値の遷移 [30]

験者に合わせた能力評価をリアルタイムで実行可能である．

ただし，この試験では，事前に項目特性を準備しておかなければならない．なぜならば，受験者に合わせた適切な問題項目が出題されるため，受験者集団全体の反応パターンを表すマトリクスは，未回答項目を多く含む不完全なものとなり，従来の IRT での推定法が使えないためである．あらかじめ問題の特性が与えられていれば能力値の推定は可能になる．そのため，予備テストを実施し，事前に項目特性を推定する手順が必要になる．

しかし，適応型試験の受験者が増えてくると，予備テストでの受験者特性と適応型試験での受験者特性との間に開きが出る恐れがある．従来的には，多数の受験者集団を対象に，定期的に予備テストを実施し，項目特性を推定し直すなどといった工夫が施されるが，一般に予備テストを実施するには多大な費用および労力がかかるため，現実的な方法ではない．

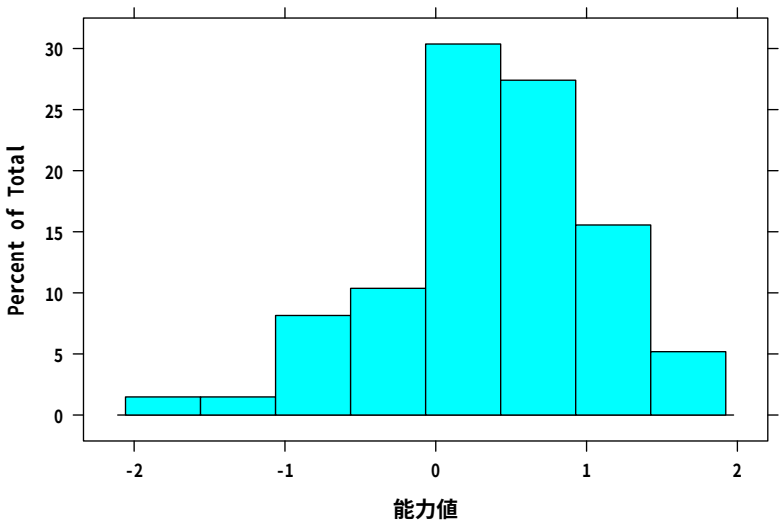


図 4.11: 5 問終了時の能力値の分布 [30]

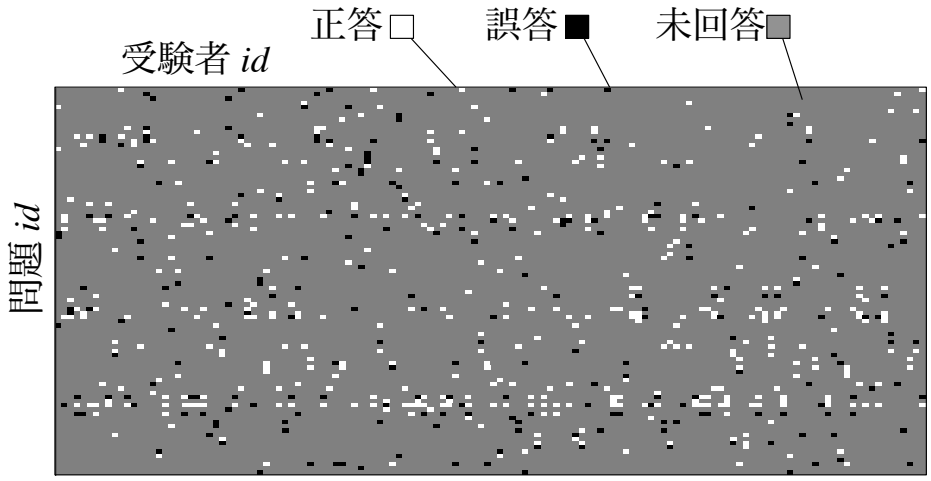


図 4.12: 2013 年度の適応型オンライン能力評価システム結果 [30]

第 5 章

EM タイプ IRT による不完全マトリクスの予測

予備テストを受験する受験者集団は適応型システムを受験する集団とは異なるため，適応型システムを受験する受験者が多くなれば，予備テストによって準備していた項目特性が適応型システムを受験する受験者から得られる項目特性とずれてくる可能性があるので，項目特性の再構成（キャリブレーション）が必要になってくる．このとき，通常の適応型テストでは受験した問題だけで受験者の能力推定が可能であったものが，項目特性まで推定しなければならなくなるため，そのままでは不完全マトリクスに対応できない推定法を変更する必要が出てくる．本章では，このように不完全マトリクスの解答パターンから項目特性を推定する新しい方法として EM タイプ IRT を提案する [14,15,24]．この方法は，データの背後にロジスティックモデルの確率構造を仮定し，不完全マトリクスでの観測された要素の値を用いて観測されていない空要素の値を確率的に予測するものである．

ここでは，まず，テストの問題項目特性および受験者能力が分かっているときに不完全マトリクスを模擬したデータを用いて，提案法が元のパラメータを再現できることを確認し，次に実際に行ったテストに対して提案法による予測を行う．推定されたパラメータに現実性を与えるため，ここでは実際に大学内で行った数学のテストから得られた解答パターンを利用する．さらに項目特性のキャリブレーションについて検証する．

5.1 EM タイプ IRT

EM タイプ IRT は，項目 j の項目特性値 a_j, b_j および受験者 i の能力値 θ_i を不完全マトリクスから推定することにより，正答確率 $P(\theta_i, a_j, b_j)$ によって不完全マトリクスの欠損部分（欠損セル）の予測値とする予測手法である．第 2 章で，回答結果を表す δ を有理数に拡張することが可能であることを示した．これを利用し，EM タイプ (expectation-maximization algorithm [8]) のパラメータ推定手順を示す．

5.1.1 欠損要素に対する予測

まず，解答パターンのマトリクスで解答されていない要素に対し，$\delta^0_{i,j} \in [0,1]$ を満たす任意の初期値を与え，$\delta_{i,j} = 0, 1$ の観測値はそのまま残す．このとき得られる初期マトリクスは，$0 \leq \delta^0_{i,j} \leq 1$ を満たす．$\delta^0_{i,j}$ の初期値としては，項目 j の平均正答率 μ_j や，受験者 i の平均正答率 μ_i などが挙げられる．各パラメータの初期値を a^0_j, b^0_j, θ^0_i とし，初期尤度 L^0 を式 (2) で定義する．

初期マトリクス $\{\delta^0_{i,j}\}$ を用いて，式 (2) の尤度 L を最大にするパラメータ a^1_j, b^1_j, θ^1_i を推定し，尤度 L^1 を得る．このときのパラメータ推定法は，2 段階アルゴリズムまたは MCMC のどちらかを用いることができる．この手順は，EM アルゴリズムの maximization ステップに対応する．

次に，得られたパラメータを用いて式 (2.6) から正答確率 $P_j(\theta_i) \in [0,1]$ が算出できる．ここで，$\hat{\delta}_{i,j} = P_j(\theta_i)$ の関係が成り立つことから，観測値および $P_j(\theta_i)$ によって，$\delta^1_{i,j}$ を得る．この手順は，EM アルゴリズムの expectation ステップに対応する．

この 2 ステップの手順を繰り返し，L^k, $\delta^k_{i,j}$, a^k_j, b^k_j, θ^k_i $(k = 0, \dots)$ を得る．$k \to \infty$ とすれば，期待される収束値 L^∞, $\delta^\infty_{i,j}$, a^∞_j, b^∞_j, θ^∞_i を得る．ただし，通常の EM アルゴリズムのような単調増加性は期待されないが，EM アルゴリズムとよく似た収束性を持つので，ここでは EM タイプ IRT と呼ぶ．この手法は，limiting IRT (LIRT) とも呼ばれる [15]．収束値が常に一意になるとは保証されない [25, 28]．しかしながら，経験的には，多くの場合で同じ値に収束することが分かっている [14]．図 5.1 に，EM タイプ IRT に

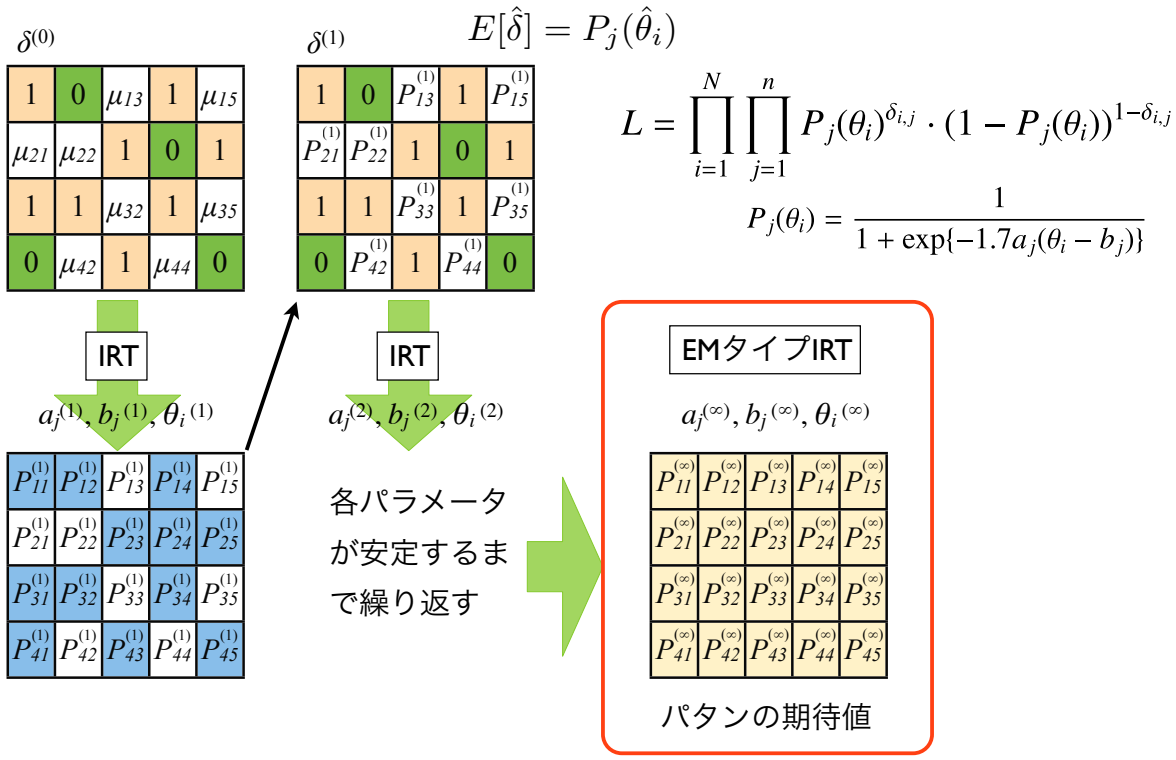


図 5.1: EM タイプ IRT による予測手順

よる予測手順の概要を示す．

得られる予測マトリクスの精度評価には下記の S^k を用いる．これは，次式で表される観測値とそれに対応する予測値 $\hat{\delta}_{i,j}^k$ の平均 2 乗誤差の平方根である．

$$S^k = \sqrt{\frac{1}{|\Delta|} \sum_{(i,j)\in\Delta} (\hat{\delta}_{i,j}^k - \delta_{i,j})^2} \tag{5.1}$$

ここで，$|\Delta|$ は観測値に対応する要素の数を表す．欠損要素の予測値は S^k に含まれないことに注意しておく．収束判定として，次式を満たした場合に収束とみなす．

$$|S^k - S^{k-1}| < 1.0 \times 10^{-8}$$

ここでは δ の大きさが $[0, 1]$ であることを配慮して，収束条件を相対誤差ではなく絶対誤差を用いてもよいと考えたからである．

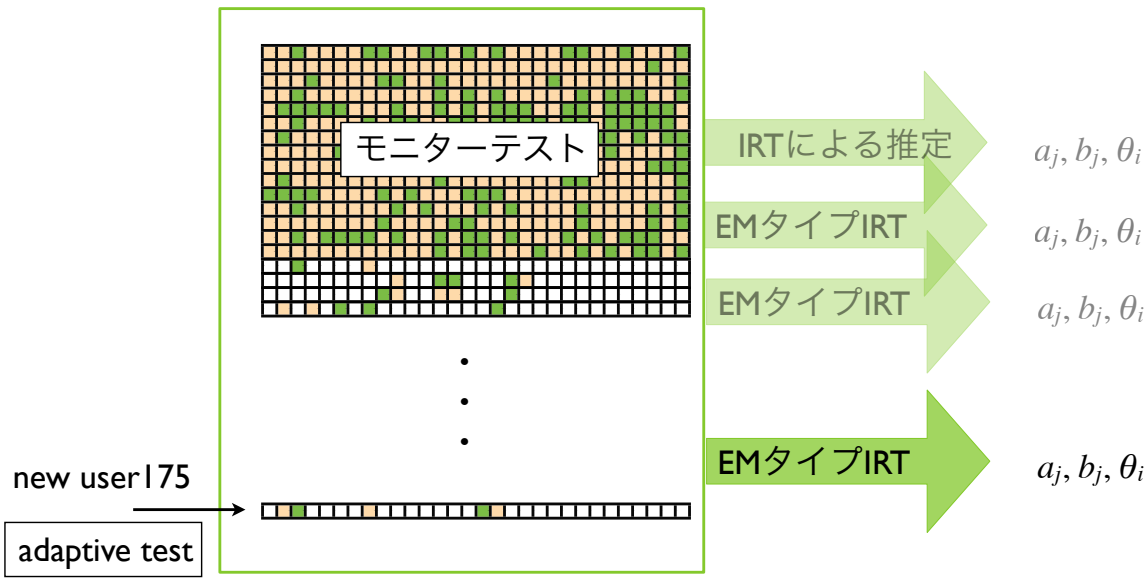


図 5.2: EM タイプ IRT によるキャリブレーション手順

5.1.2 EM タイプ IRT による項目特性のキャリブレーション

EM タイプ IRT を用いることによって，不完全なマトリクスから受験者能力と項目特性を同時に推定することが可能となる．適応型試験システムの中で，受験者が増えても問題の特性が受験者特性と開きが出ないように，受験者が増えるごとに受験者能力と同時に問題項目特性も更新していく．そのときの考え方を図 5.2 に示す．適応型試験の受験者が増えるごとに，EM タイプ IRT による推定を繰り返し適用し，問題項目特性を逐次更新していくことで，適応型試験の受験者特性に合わせた項目特性のキャリブレーションが実現できる．

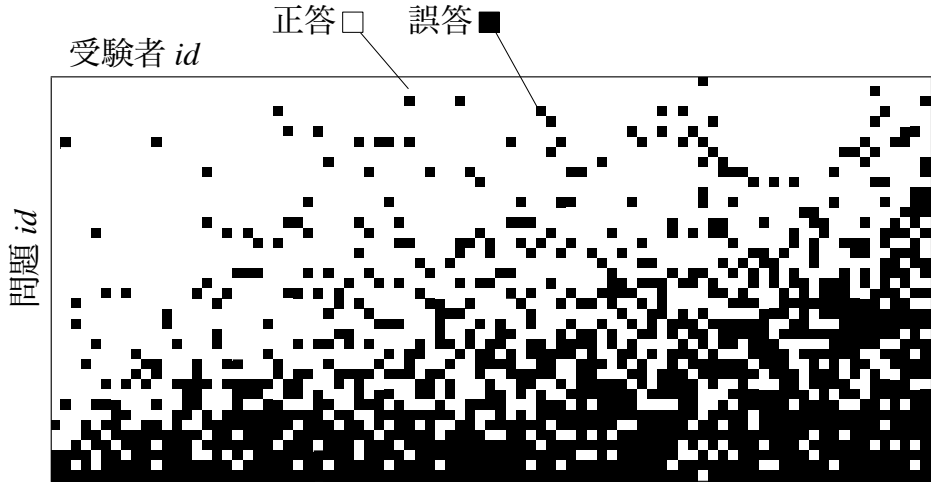


図 5.3: 数学の試験（完全データ 2009）の解答結果（87 人 ×40 問） [30]

5.2 EM タイプ IRT による不完全マトリクスの予測

5.2.1 対象データ

ここでは対象データとして，適応型試験を実際に実施して得られる不完全マトリクスデータを用いる．ただし，EM タイプ IRT の正しさを検証するために，通常の筆記試験によって得られる完全マトリクスデータをもとに，答えが分かっている不完全マトリクスデータも作成し扱う．

完全データ 2009

完全マトリクスとして，大学の学部生に実施した筆記試験による数学の試験問題とその結果（以下，完全データ 2009）を利用する [36]．このときの受験者数は 87 人，試験の問題項目数は 40 問である．このときの解答結果を図 5.3 に示す．行方向は各問題項目を表しており，列方向は各受験者を表している．また，行方向は IRT によって推定される問題項目の困難度を昇順で，列方向は同じく IRT によって推定される受験者の能力値を降順でソートしている．そのため，マトリクスの左上部は，能力の高い受験者で，かつ易し

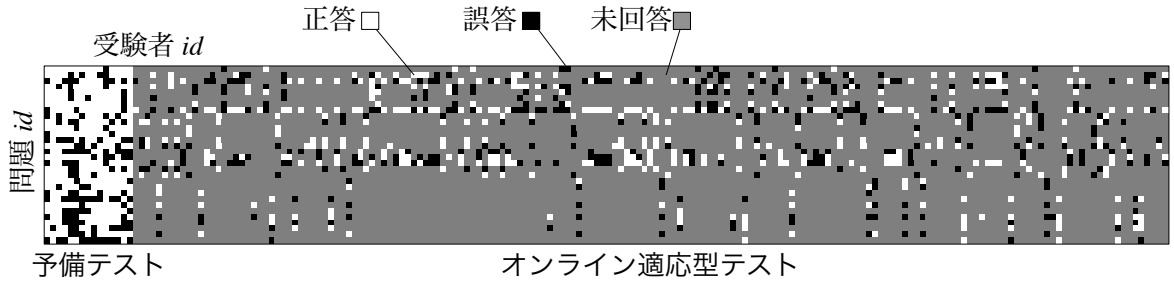


図 5.4: 予備テスト結果と適応型システム結果（不完全データ 2011）

い問題項目になるため，正答（白色）が多くなっている．逆に，マトリクスの右下部は，能力の低い受験者かつ難しい問題項目になるため，誤答（黒色）が多くなっている．

不完全データ 2011

不完全マトリクスとして，2011 年度に実施した適応型オンライン能力評価システムの結果（以下，不完全データ 2011）を用いる．このときの，項目バンクの問題項目数は 30 問であり，すべて高校数学レベルの問題である．これらの項目のパラメータを推定するために事前に実施した予備テスト受験者は 15 人である．適応型オンライン能力評価システムを受験した受験者は，175 人である．受験者一人あたりの回答数は 5 問程度である．

不完全データ 2013

不完全マトリクスとして，2013 年度に実施した適応型オンライン能力評価システムの結果（以下，不完全データ 2013）を用いる．このときの項目バンクに登録されている問題数は合計 96 問で，すべて高校数学レベルである．これらの項目のパラメータを推定するために事前に実施した予備テスト受験者は 68 人である．適応型オンライン能力評価システムを受験した受験者は，138 人である．受験者一人あたりの回答数は 5 問程度である．このとき得られる不完全マトリクスデータ（以下，不完全データ 2013）を図 5.5 に示す．図中の左側が予備テストの結果，右側が適応型システムの結果を表す．図中の白色の要素は正答を表し黒色は誤答を表す．灰色の要素は欠損していることを表す．

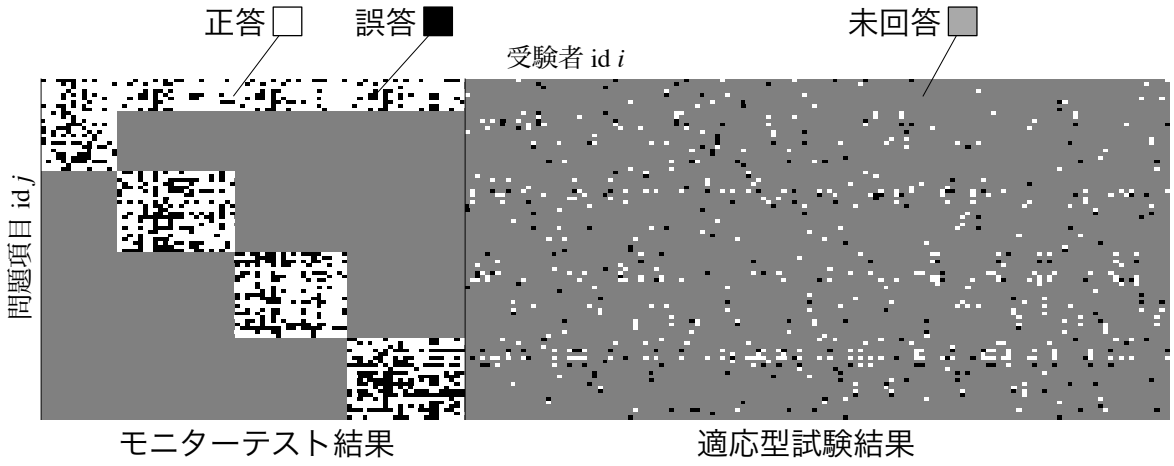


図 5.5: 予備テスト結果と適応型システム結果（不完全データ 2013） [30]

5.2.2 完全データ 2009 をもとにしたシミュレーション

検証手順

ここでは，EM タイプ IRT が正しく機能していることをコンピュータシミュレーションにより検証する．検証手順を以下に示す．

1. 完全データ 2009 から，IRT によって，能力パラメータ θ_i $(i = 1, \ldots, N)$，項目パラメータ a_j, b_j $(j = 1, \ldots, n)$，すべてのマトリクスの要素の正答確率 $p_{i,j}$ を計算する．
2. θ_i, a_j, b_j から，完全データ 2009 を模倣したマトリクス A_m $(m = 1, 2, \ldots, M)$ をモンテカルロ法によって生成する．
3. M 個のマトリクスのそれぞれについて，以下を行う．
   (a) A_m を Training（K 個の要素） と Test（k 個の要素） に無作為に分ける．
   (b) Training から，EM タイプ IRT によって，能力パラメータ $\hat{\theta}_i$，項目パラメータ $\hat{a}_j$, $\hat{b}_j$，および欠損値（Test に相当する部分）の正答確率 $t_{i,j}$ を求める．
   (c) $\hat{\theta}_i, \hat{a}_j, \hat{b}_j, t_{i,j}$ のそれぞれについて，$bias$ と mse を計算する．

$$bias = \frac{1}{n_x}\sum(\hat{x} - x), \quad mse = \frac{1}{n_x}\sum(\hat{x} - x)^2$$

$$(x, \hat{x}, n_x) \in \{(p_{i,j}, t_{i,j}, k), (\theta_i, \hat{\theta}_i, N), (a_j, \hat{a}_j, n), (b_j, \hat{b}_j, n)\}$$

モンテカルロ法によるマトリクス生成法は [39] によった．また，本研究では $M = 10$ とした．$K + k = N \times n = 87 \times 40 = 3480$ であり，$K = 2784$（全体の 80%）の場合と $K = 696$（全体の 20%）の場合で計算を行った．

欠損値の $bias$ **と** mse

ここでは，提案法の予測精度を評価するため，欠損要素に $[0, 1]$ からランダムに選んだ数値を入れた場合の結果と比較した．このときの $K = 2784$ のときの欠損値に対する $bias$ と mse の計算結果を表 5.1 に示す．表ではそれを random と表している．表 5.1 から，EM タイプ IRT による予測値は，ランダムに欠損値を埋めた場合に比べ，$bias$，mse ともに小さい値を示していることが分かる．また，$K = 696$ のときの欠損値に対する $bias$ と mse の計算結果を表 5.2 に示す．$K = 2784$ のときと比べて EM タイプ IRT による予測値は $bias$ と mse の値が大きくなっているが，ランダムに欠損値を埋めた場合に比べると，やはり $bias$，mse ともに小さい値を示していることが分かる．

表 5.1: シミュレーションによる欠損値の $bias$ と mse（$K = 2784$） [30]

	EM タイプ IRT		random	
m	$bias$	mse	$bias$	mse
1	2.42×10^{-3}	7.21×10^{-3}	-1.54×10^{-1}	1.94×10^{-1}
2	1.53×10^{-3}	8.92×10^{-3}	-1.85×10^{-1}	2.06×10^{-1}
3	-6.18×10^{-4}	9.34×10^{-3}	-1.48×10^{-1}	2.09×10^{-1}
4	8.62×10^{-5}	8.23×10^{-3}	-1.66×10^{-1}	2.12×10^{-1}
5	-3.21×10^{-5}	8.23×10^{-3}	-1.77×10^{-1}	2.01×10^{-1}
6	-1.95×10^{-3}	8.32×10^{-3}	-1.74×10^{-1}	2.13×10^{-1}
7	3.97×10^{-3}	9.97×10^{-3}	-1.73×10^{-1}	2.02×10^{-1}
8	8.83×10^{-3}	7.49×10^{-3}	-1.84×10^{-1}	2.03×10^{-1}
9	1.56×10^{-3}	6.17×10^{-3}	-1.66×10^{-1}	2.00×10^{-1}
10	1.08×10^{-3}	6.94×10^{-3}	-1.70×10^{-1}	2.01×10^{-1}
平均	1.69×10^{-3}	8.08×10^{-3}	-1.70×10^{-1}	2.04×10^{-1}

表 5.2: シミュレーションによる欠損値の $bias$ と mse（$K = 696$） [30]

	EM タイプ IRT		random	
m	$bias$	mse	$bias$	mse
1	1.37×10^{-2}	2.71×10^{-2}	-1.74×10^{-1}	2.00×10^{-1}
2	-4.16×10^{-3}	2.32×10^{-2}	-1.71×10^{-1}	1.96×10^{-1}
3	8.96×10^{-3}	2.09×10^{-2}	-1.69×10^{-1}	2.04×10^{-1}
4	-5.13×10^{-3}	1.97×10^{-2}	-1.81×10^{-1}	1.95×10^{-1}
5	1.45×10^{-2}	2.42×10^{-2}	-1.74×10^{-1}	2.04×10^{-1}
6	7.87×10^{-3}	2.28×10^{-2}	-1.71×10^{-1}	2.01×10^{-1}
7	1.47×10^{-2}	2.39×10^{-2}	-1.66×10^{-1}	1.97×10^{-1}
8	1.11×10^{-2}	2.85×10^{-2}	-1.69×10^{-1}	2.05×10^{-1}
9	-4.31×10^{-3}	2.05×10^{-2}	-1.67×10^{-1}	2.07×10^{-1}
10	3.35×10^{-2}	2.63×10^{-2}	-1.76×10^{-1}	2.06×10^{-1}
平均	9.08×10^{-3}	2.37×10^{-2}	-1.72×10^{-1}	2.01×10^{-1}

パラメータ推定値の $bias$ と mse

次に，$K = 2784$ における各推定パラメータ θ_i,a_j,b_j の $bias$ と mse を表 5.3 に示す．どのパラメータも $bias$，mse ともに小さい値を示しており，もとのパラメータを再現し

表 5.3: シミュレーションによる θ_i,a_j,b_j の $bias$ と mse（$K = 2784$） [30]

	θ		a		b	
m	$bias$	mse	$bias$	mse	$bias$	mse
1	1.42×10^{-2}	8.52×10^{-2}	5.23×10^{-2}	9.41×10^{-2}	-1.14×10^{-1}	1.14×10^{-1}
2	1.23×10^{-2}	1.31×10^{-1}	1.64×10^{-2}	5.60×10^{-2}	-1.07×10^{-2}	9.95×10^{-2}
3	1.70×10^{-2}	1.23×10^{-1}	5.37×10^{-2}	6.85×10^{-2}	-9.43×10^{-2}	9.75×10^{-2}
4	1.08×10^{-3}	1.04×10^{-1}	3.35×10^{-2}	8.08×10^{-2}	-7.09×10^{-2}	1.10×10^{-1}
5	2.51×10^{-2}	1.39×10^{-1}	3.56×10^{-2}	7.36×10^{-2}	-2.45×10^{-2}	9.78×10^{-2}
6	-7.16×10^{-3}	1.19×10^{-1}	3.21×10^{-2}	7.28×10^{-2}	-2.05×10^{-2}	8.64×10^{-2}
7	5.21×10^{-2}	1.26×10^{-1}	7.13×10^{-2}	6.67×10^{-2}	-7.06×10^{-2}	1.01×10^{-1}
8	2.97×10^{-2}	1.27×10^{-1}	2.94×10^{-2}	6.31×10^{-2}	-3.21×10^{-2}	7.35×10^{-2}
9	1.67×10^{-2}	8.22×10^{-2}	6.80×10^{-2}	6.30×10^{-2}	-3.84×10^{-2}	1.48×10^{-1}
10	8.63×10^{-3}	8.04×10^{-2}	2.03×10^{-2}	7.03×10^{-2}	-1.02×10^{-1}	1.37×10^{-1}
平均	1.70×10^{-2}	1.12×10^{-1}	4.13×10^{-2}	7.09×10^{-2}	-5.78×10^{-2}	1.06×10^{-1}

ていることが分かる．また，$K = 696$ における各推定パラメータ θ_i,a_j,b_j の $bias$ と mse を表 5.4 に示す．$K = 2784$ のときと比べて，各パラメータ推定値の mse の値が大きく

表 5.4: シミュレーションによる θ_i,a_j,b_j の $bias$ と mse（$K = 696$） [30]

	θ		a		b	
m	$bias$	mse	$bias$	mse	$bias$	mse
1	1.58×10^{-2}	2.71×10^{-1}	7.80×10^{-2}	8.45×10^{-2}	-1.01×10^{-1}	3.88×10^{-1}
2	5.87×10^{-3}	2.39×10^{-1}	3.56×10^{-2}	8.17×10^{-2}	-4.08×10^{-2}	2.49×10^{-1}
3	5.81×10^{-3}	3.41×10^{-1}	7.67×10^{-2}	8.35×10^{-2}	-1.20×10^{-1}	1.98×10^{-1}
4	1.09×10^{-2}	2.79×10^{-1}	7.80×10^{-3}	8.20×10^{-2}	-8.66×10^{-2}	2.22×10^{-1}
5	7.51×10^{-3}	3.03×10^{-1}	7.42×10^{-2}	7.30×10^{-2}	-1.36×10^{-1}	2.07×10^{-1}
6	1.72×10^{-2}	3.51×10^{-1}	8.13×10^{-2}	6.81×10^{-2}	-1.50×10^{-1}	2.00×10^{-1}
7	1.58×10^{-2}	3.07×10^{-1}	9.80×10^{-2}	9.24×10^{-2}	-1.71×10^{-1}	3.11×10^{-1}
8	2.21×10^{-2}	4.10×10^{-1}	8.17×10^{-2}	7.34×10^{-2}	-1.97×10^{-1}	3.25×10^{-1}
9	-1.23×10^{-2}	2.72×10^{-1}	5.90×10^{-2}	6.38×10^{-2}	-8.32×10^{-2}	2.22×10^{-1}
10	2.98×10^{-2}	3.61×10^{-1}	3.61×10^{-2}	6.45×10^{-2}	-2.33×10^{-1}	2.91×10^{-1}
平均	1.18×10^{-2}	3.13×10^{-1}	6.28×10^{-2}	7.67×10^{-2}	-1.32×10^{-1}	2.61×10^{-1}

なっている．しかし，その値は $bias$，mse ともに小さく，ここでももとのパラメータを再現していると考えられる．

5.2.3 不完全データ 2011 への適用

不完全マトリクスの予測結果

不完全データ 2011 に対する EM タイプ IRT による予測結果 [24] を図 5.6 に示す． 上図は不完全データ 2011 であり，下図はそれをもとに EM タイプ IRT で予測した結果を表す．各マトリクスの行は項目を，列は受験者を表す．マトリクスにおける白色は正答，黒色は誤答，灰色は未回答を表している．また，[0,1] の値を青色から赤色にかけてグラデーションで表している．図から，誤答（黒色）が多い行および列は 0 に近い予測値（青色）を示しており，正答（白色）が多い行および列は 1 に近い予測値（赤色）を示している．受験者および項目の特性を反映していると考えられる．

このときの EM タイプ IRT による繰り返しごとの予測されたマトリクスの様子を図 5.7

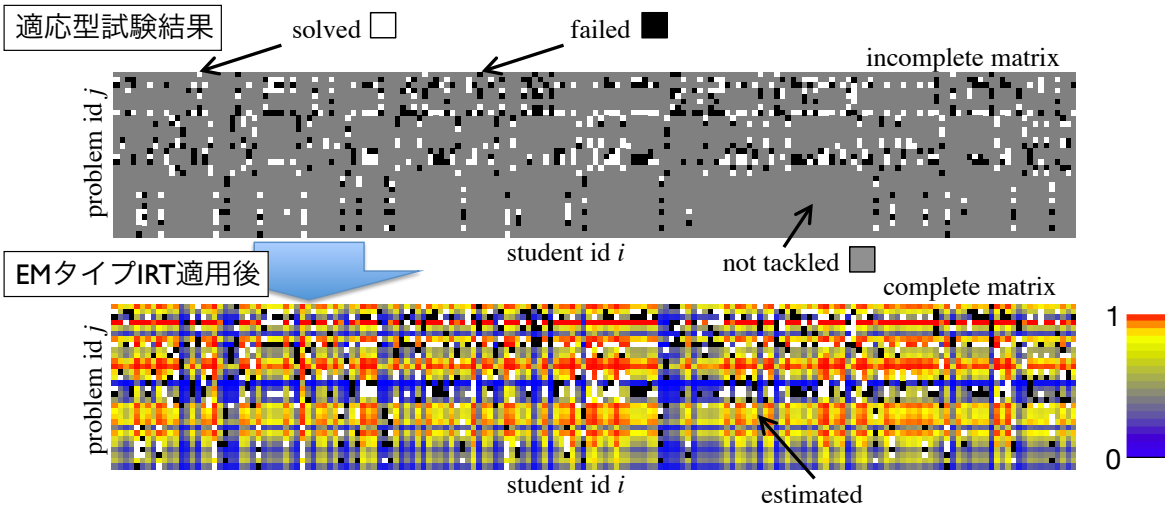


図 5.6: 不完全データ 2011 と EM タイプ IRT によって予測したマトリクス [24]

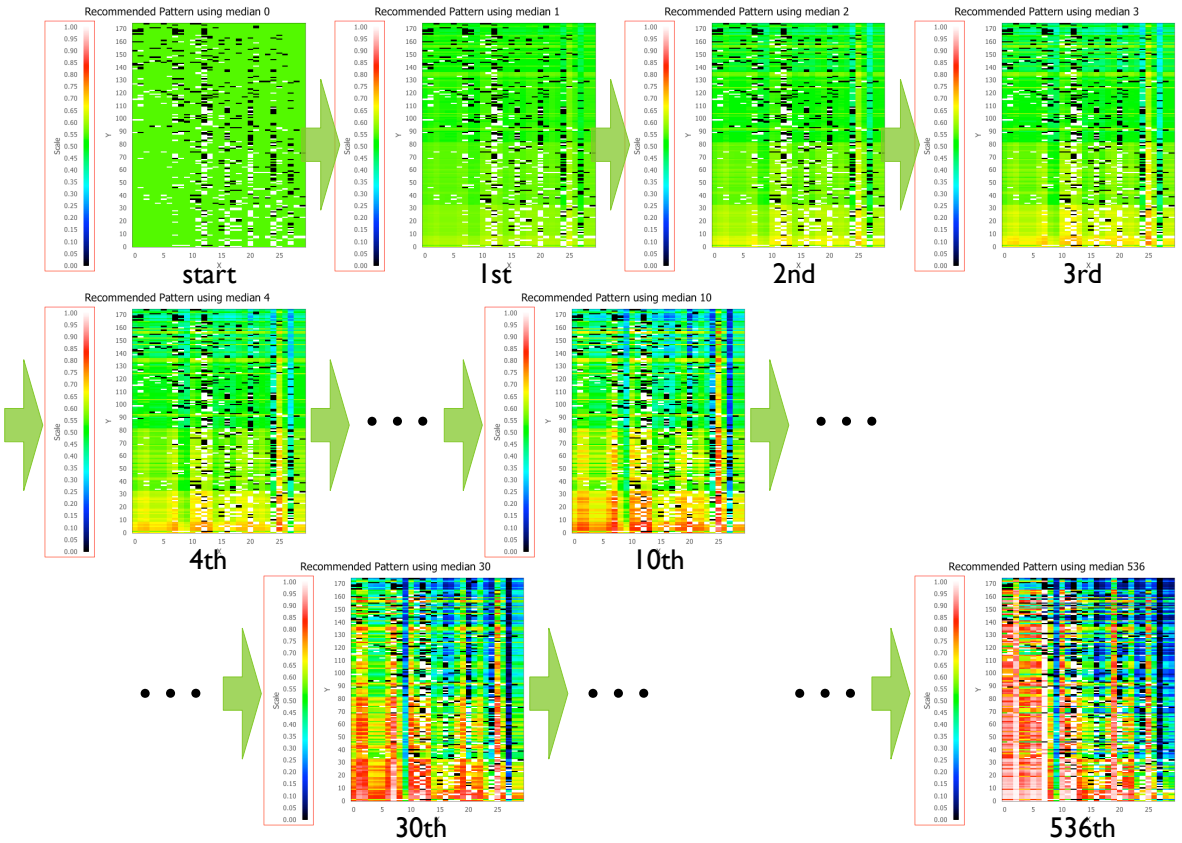


図 5.7: EM タイプ IRT による不完全データ 2011 の予測の様子

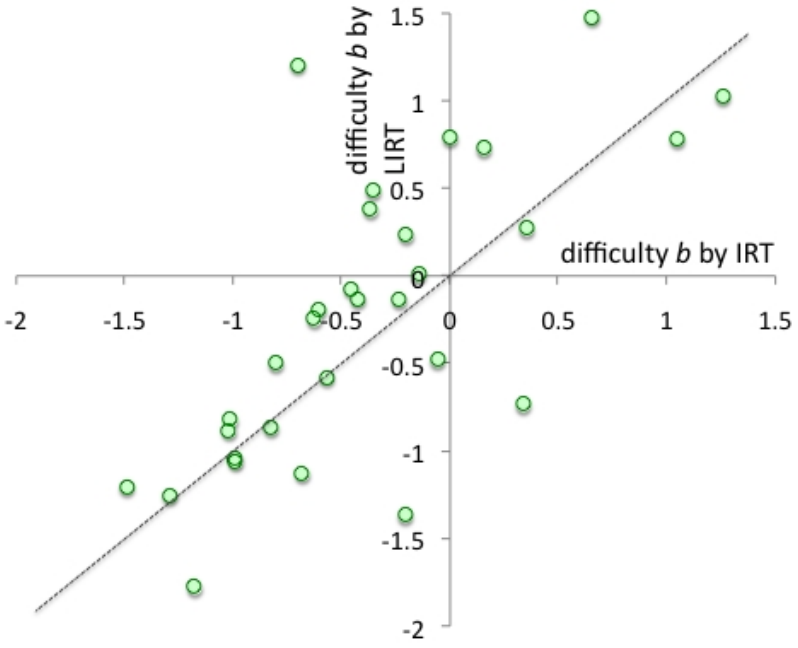


図 5.8: 項目バンクの項目困難度と EM タイプ IRT による項目困難度の違い [24]

に示す．マトリクスの横軸は問題項目，縦軸は適応型試験の受験者を表す．また，それぞれの軸について，初期における問題項目の困難度および受験者の能力値によってソートしている．図中の白色が正答，黒色が誤答を表し，青から赤までの色の変化が 0 から 1 までの値の変化に対応している．EM タイプ IRT を繰り返すごとに，予測されるマトリクスが少しずつ変化していることが色の変化で分かる．

項目特性値の比較

ここで，EM タイプ IRT によって推定された項目特性値が，項目バンクにもともと登録してあった項目特性値からどのように変化したのかを見る．まずはじめに，図 5.8 に項目困難度を示す．横軸は，項目バンクに登録してあったもとの項目困難度，縦軸は EM タイプ IRT によって推定された項目困難度を表している．両者にはある程度の相関があるように見えるが，いくつかの項目で困難度が大きく変化していることが分かる．

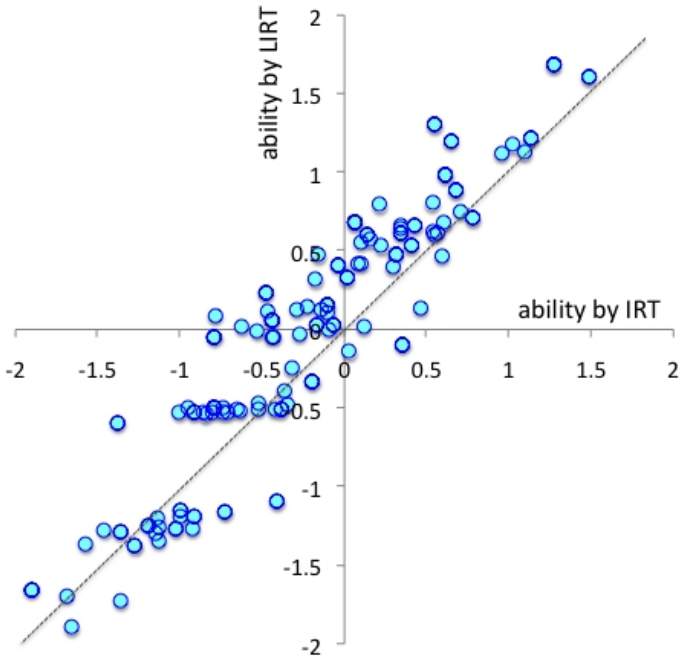


図 5.9: 適応型試験による能力推定値と EM タイプ IRT による能力推定値の違い [24]

能力値の比較

次に，能力値について見てみる．図 5.9 に，適応型試験によって 5 問の回答結果から得られた能力推定値（横軸）と，EM タイプ IRT によって得られた能力推定値（縦軸）を表す．両者の間には強い相関があることが分かる．

図 5.10 に，適応型試験と EM タイプ IRT による能力推定値（横軸）とその標準偏差（縦軸）の関係を示す．適応型試験の結果では，5 問しか回答していないため，能力値にばらつきが現れていることが分かる．一方，EM タイプ IRT の結果では，適応型試験の結果と比較して，標準偏差は小さくなっており，より信頼度の高い能力推定値を期待できることが分かる．なぜなら，EM タイプ IRT では完全なマトリクスを回答しているとみなしているためである．

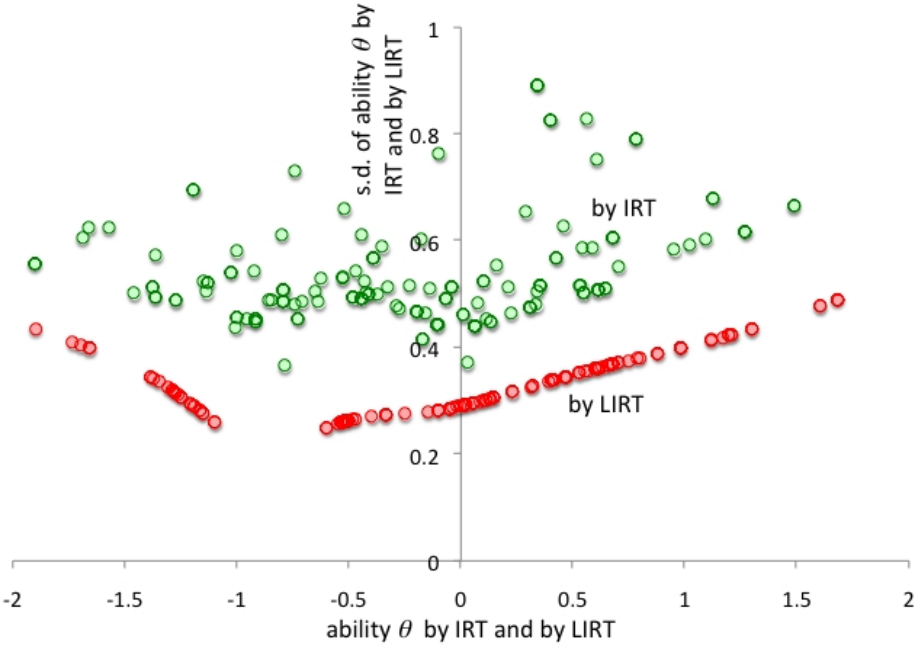


図 5.10: 適応型試験と EM タイプ IRT による能力推定値とその標準偏差 [24]

初期値による収束性

次に，EM タイプ IRT における繰り返し手順によって得られる RMSE および対数尤度の変化を図 5.11 に示す．上図が対数尤度，下図が RMSE を示している．図中の 3 本のラインは，それぞれ初期値が異なる場合の EM タイプ IRT の結果を示している．図 5.11 上図を見ると，初期値 μ_j のときの対数尤度 $\log L$ の変化は初期段階において減少した後に増大しており，単調に変化していないことが分かる．ただし，他の初期値については単調増加している．図 5.11 下図を見ると，RMSE についてはすべての初期値において単調減少している．また，どの初期値においても，収束には至っており，その収束値はほぼ一致している．以上のことから，EM タイプ IRT による推定では，その計算過程において，単調に収束に向かってはいないが，少なくとも収束には至っていることが経験的に分かる．

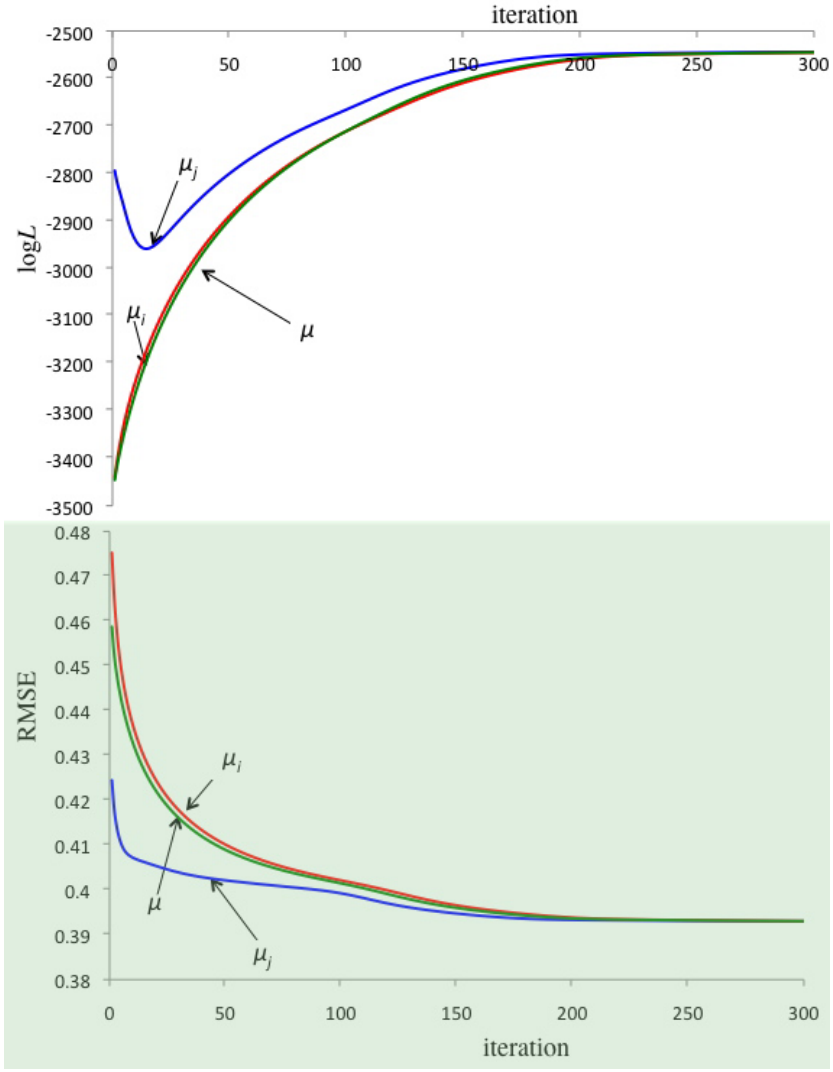


図 5.11: EM タイプ IRT における収束 [14]

項目特性のキャリブレーション結果

ここで，適応型試験の受験者が一人増すごとに EM タイプ IRT を適用した場合における，各特性値の変化を見る．先ほど示した図 5.8 や図 5.9 は，適応型試験を受験した受験者数が 175 人に達したときの EM タイプ IRT の結果を表しているのに対し，ここではその受験者数が一人増すごとに EM タイプ IRT を適用し，項目特性および能力値にどのような変化が現れているのかを考察する．つまり，図 5.2 で示したキャリブレーションを

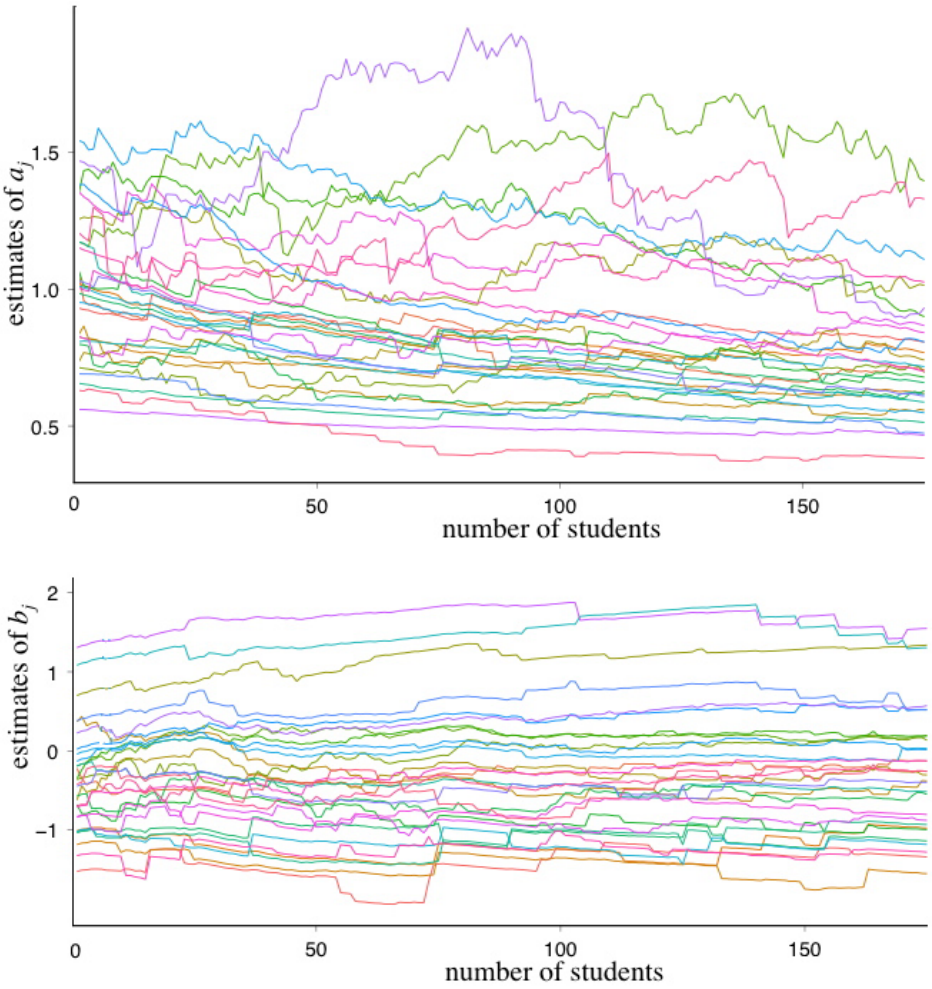


図 5.12: EM タイプ IRT によってキャリブレーションした場合の項目特性値 a_j^{∞} と b_j^{∞} の振る舞い [14]

行う．

図 5.12 に項目特性値についてのキャリブレーションの様子を示す．図 5.12 上図が項目識別力，下図が項目困難度を表している．各線は一つひとつの項目を示している．横軸は適応型試験の受験者数，縦軸は各項目特性値を表している．項目識別力については，いくつかの項目で識別力の推定値が大きく変化している．一方，項目困難度については，大きな変化は見られない．

このときの項目特性値の標準偏差を表したものが図 5.13 である．先ほどと同じく上図

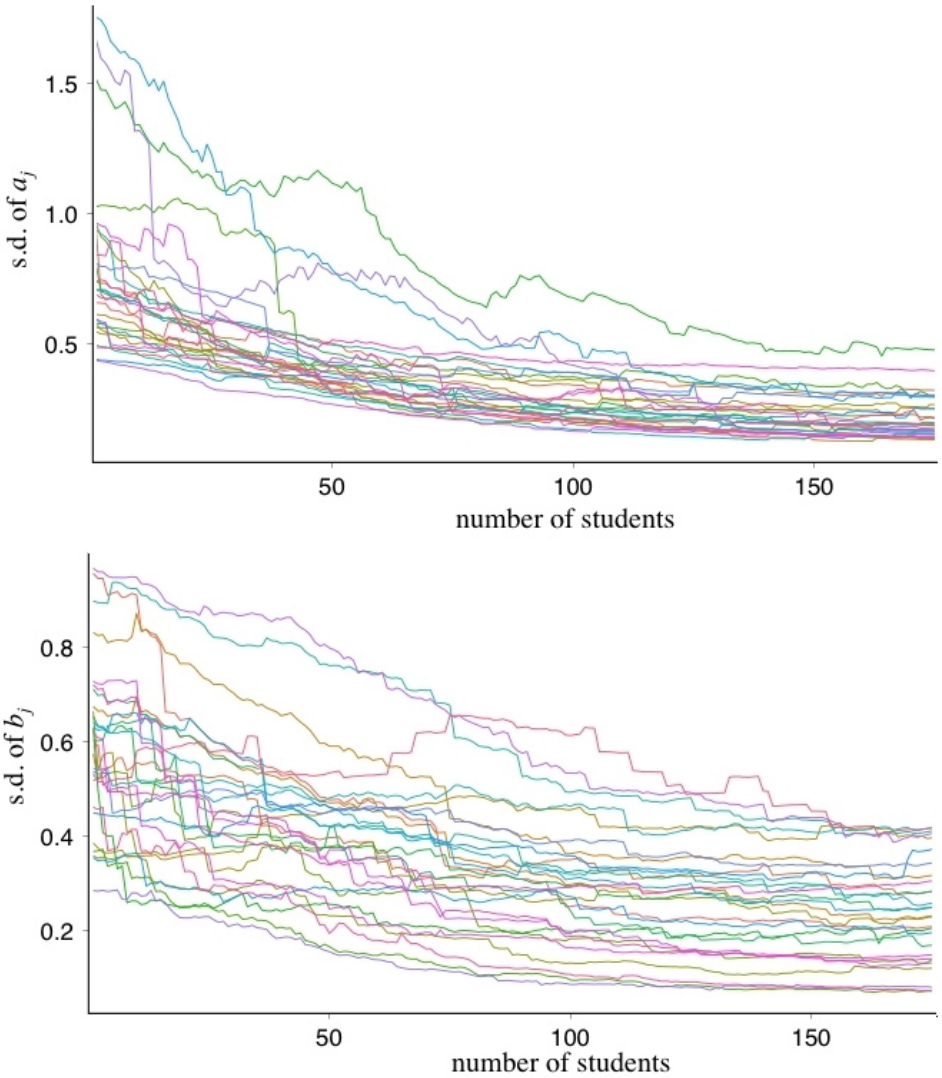


図 5.13: EM タイプ IRT によってキャリブレーションした場合の項目特性値 $sd(a_j^{\infty})$ と $sd(b_j^{\infty})$ の振る舞い [14]

が項目識別力，下図が項目困難度に対応する．横軸は適応型試験の受験者数，縦軸は各項目特性値の標準偏差を表している．どちらの場合も，適応型試験の受験者が増すごとに，標準偏差が小さくなっている．

次に，適応型試験によって得られた能力推定値が，EM タイプ IRT によってキャリブレーションされた項目特性値によってどのように変化するのかを考える．図 5.14 にそのときの能力推定値の変化を表す．各線は適応型試験受験者を表している．横軸は適応型試

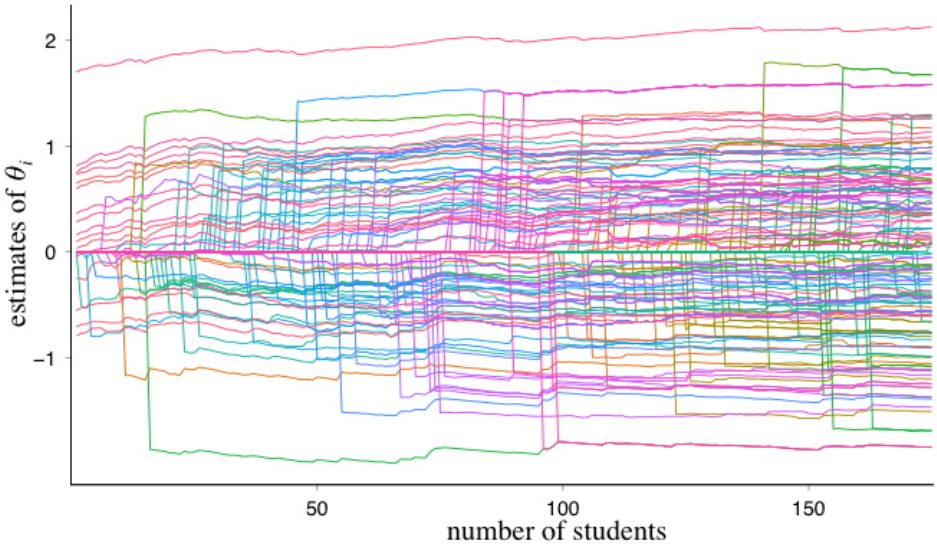


図 5.14: EM タイプ IRT によってキャリブレーションした場合の能力推定値 θ_i^{∞} の振る舞い [14]

験の受験者数，縦軸はそのときの能力推定値を表す．項目困難度のときと同様に，大きな変化はない．適応型試験の受験者数が増し，項目特性値がキャリブレーションされているが，適応型試験によって得られた能力推定値と矛盾していないことを意味する．

このときの能力推定値の標準偏差を表したものが図 5.15 である．横軸は適応型試験の受験者数，縦軸は能力推定値の標準偏差を表している．能力推定値の標準偏差については，ほとんど変化がなく一定の値を示している．これは，どの段階においても同じ数の問題項目数から推定された結果であるためであると考えられる．

5.3 本章のまとめ

本章では，オンラインにおける適応型試験の中で，受験者が増えても問題の特性が受験者特性と開きが出ないように，受験者の評価を行うと同時に問題の特性も受験者が増える毎に更新していく方法として，EM タイプ IRT を提案した．具体的には，不完全マトリクスにおける空き要素での反応を適当な初期値で与え，不完全マトリクスを一旦完全マトリクスに置き換えて，通常の IRT を拡張した枠組みでパラメータを推定し，推定された反応パターンを観測値と比較して，尤度と距離の 2 つの規準からその差が最小になるよう

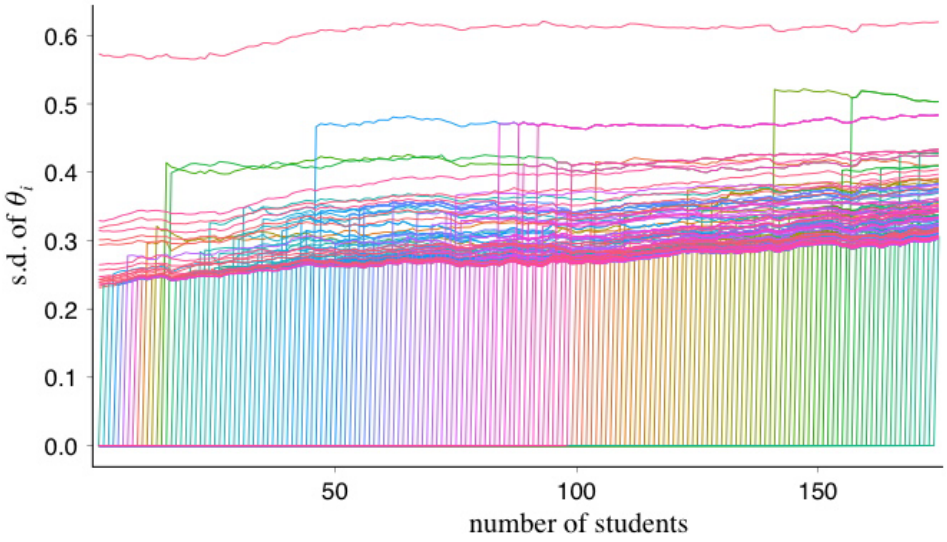


図 5.15: EM タイプ IRT によってキャリブレーションした場合の能力推定値 $sd(\theta_i^{\infty})$ の振る舞い [14]

に更新しながら繰り返しパラメータを推定していく方法である．

提案法によって能力特性と項目特性を同時に効率よく推定できているかを，まず，テストの問題項目特性および受験者能力が分かっているときに不完全マトリクスを模擬したデータを用いて確認した結果，提案法が元のパラメータを再現できることが分かった．次に，実際に得られた不完全マトリクスに試みた結果，初期に空き要素に与える反応パターンをどのように変えても推定値は同じ収束値に向かい推定がうまくいっていることが経験的に示された．

本来，適応型試験では事前準備として予備テストを行い，項目特性を調べておく必要があるが，本提案手法によって，予備テストそのものを適応型試験の中で行える可能性があることが示唆される．予備テストを必要としない適応型試験システムは，大幅な時間と労力のコストを削減できるため，本提案手法は大きな可能性を秘めたものであると考えられる．

第 6 章

EM タイプ IRT と推薦システムの比較

不完全マトリクスから完全マトリクスを推定する方法の 1 つに，マトリクス分解法（MD; matrix decomposition method）がある [12, 26]．これは，データの背後に確率構造を仮定しないノンパラメトリックな方法であり，推薦システムのようなものに用いられている．しかしながら，受験者の資質がある程度予測できて，問題の解答パターンもある程度想定される確率分布の下で変動すると仮定できるような場合，背後に確率分布の構造を仮定した方が推定精度が良くなることも考えられる．つまり，能力評価試験のような受験者の特性が強く反映される不完全マトリクスの場合，EM タイプ IRT は MD よりも有効に働く可能性がある．

ここでは，提案法を不完全マトリクスの予測法の 1 つであるマトリクス分解法と比較した．

6.1 EM タイプ IRT とマトリクス分解法の比較

6.1.1 マトリクス分解法（MD） [26]

マトリクス分解法（matrix decomposition method; MD）は，推薦システムにおいてよく利用されている方法であり [2, 19, 20]，これを受験者と項目から作られるマトリクス

にも適用できる．マトリクス分解法では，不完全マトリクス V が 2 つの未知マトリクス U と M の積 $P = U^T M$ で表されるような U, M を最小 2 乗法によって探索する [26]．最小 2 乗法では次のように 2 乗誤差に罰則項をつけた最小化を行う．

$$f(U,V) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)\left\{V(i,j) - P(i,j)\right\}^2 + \frac{k_u}{2}\sum_{i=1}^{m}||U_i||^2 + \frac{k_m}{2}\sum_{j=1}^{n}||M_j||^2 \tag{6.1}$$

ここで，$I(i,j)$ は観測値のインデックス関数，k_u と k_m は過学習を防ぐための正則化係数である．大規模マトリクスに対する最適化においては確率的勾配法が利用されるが，ここでは次式で表される通常の勾配法を利用している．

$$\frac{\partial f}{\partial U_i} = -\sum_{j=1}^{n} I(i,j)\left\{V(i,j) - P(i,j)\right\} M_j + k_u U_i \tag{6.2}$$

$$\frac{\partial f}{\partial M_i} = -\sum_{i=1}^{m} I(i,j)\left\{V(i,j) - P(i,j)\right\} U_i + k_m M_j \tag{6.3}$$

適当な値で分解後の行列 U と M を初期化し，各反復で両行列の行と列を次式で，収束するまで更新する．

$$U_i^{(t+1)} \leftarrow U_i^{(t)} - \mu\frac{\partial f}{\partial U_i} \tag{6.4}$$

$$M_j^{(t+1)} \leftarrow M_j^{(t)} - \mu\frac{\partial f}{\partial M_j} \tag{6.5}$$

ここで，μ は学習率である．

6.1.2 予測値の評価

予測値の評価を行うために，もとの不完全マトリクスデータを Training および Test に分け，Training で予測を行い，Test でその精度を評価する．評価式は，式 (5.1) に示す $RMSE$ を用いる．図 6.1 に Training と Test に分けて評価を行う手順を示す．Training と Test は T セット作成する．

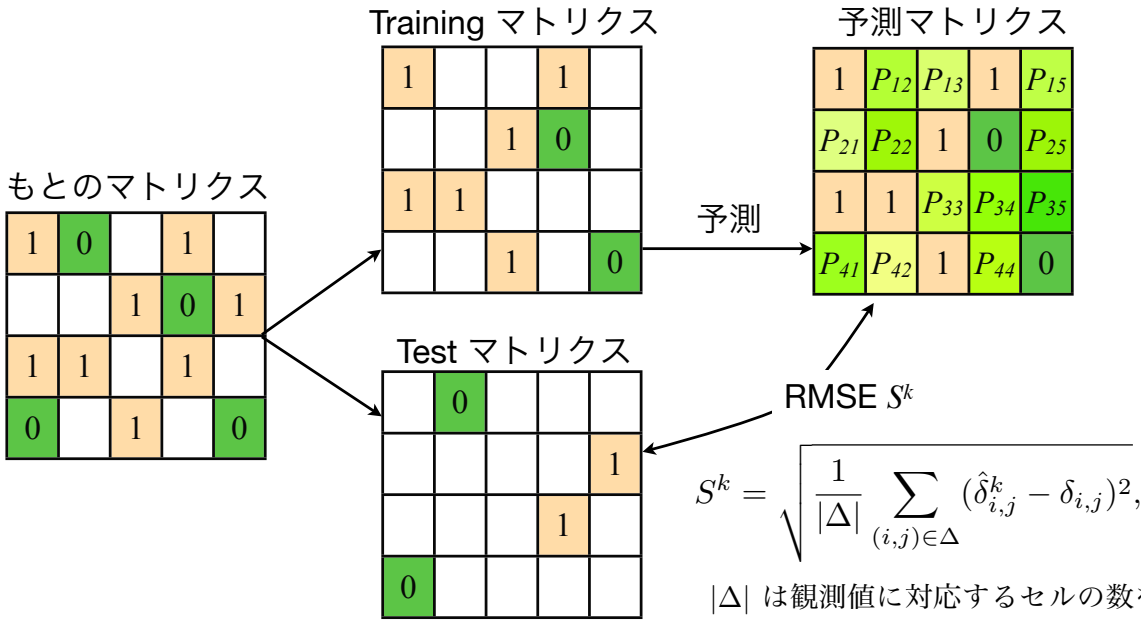


図 6.1: Training と Test に分けて推定を行う手順 [30]

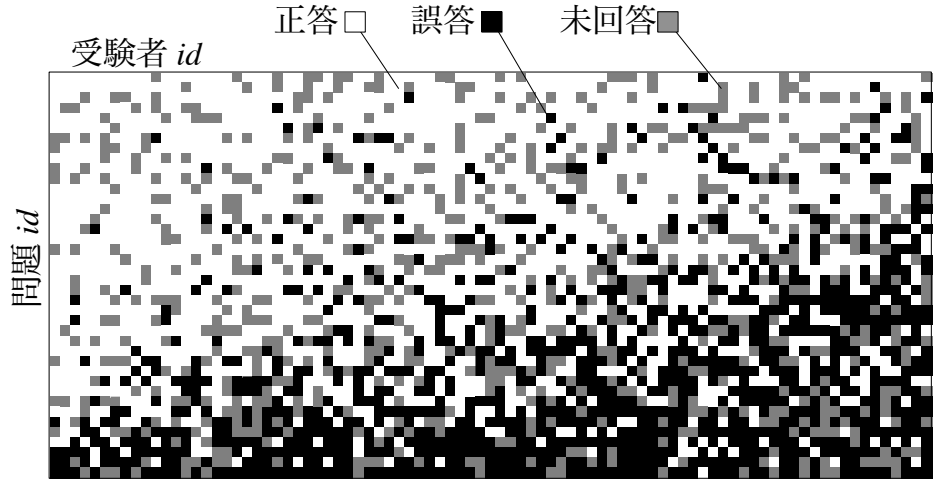
6.2 実データにおける予測精度の比較

前章の完全データ 2009 および不完全データ 2013 に対して，EM タイプ IRT および MD によるマトリクス予測手法を適用し，その予測精度を比較する．

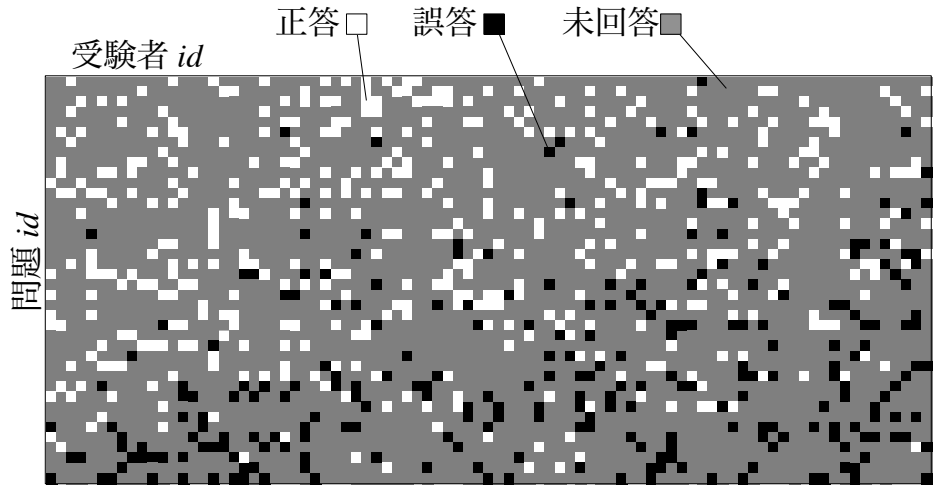
6.2.1 完全データ 2009 の予測

ここでは，完全マトリクスである完全データ 2009 から一部を欠損値として扱い，欠損部分の答えが分かっているケースとして予測精度の比較を行う．ただし，マトリクスの各行および各列について，解答結果が少なくとも 1 つは存在するようにする．欠損値の割合については，マトリクス全体の 20% および 80% が欠損している 2 つのケースを考える．欠損値の生成法については，無作為抽出による．

各ケースにおいて，もとの完全マトリクスから欠損させた部分を Test データ，不完全マトリクスとして残った部分を Training データとして扱う．Training データから欠損部



(a) 20% 欠損



(b) 80% 欠損

図 6.2: 完全データ 2009 から生成した不完全マトリクス [30]

分のマトリクス予測を行い，その答えである Test データとの $RMSE$ を式 (5.1) によって求める．これを 30 パターン行い，得られる 30 個の $RMSE$ の平均および標準偏差によって予測精度とする．

完全データ 2009 から無作為に 20% および 80% を欠損させた場合の一例を図 6.2 に示す．図中の白色部分は正答，黒色部分は誤答，灰色部分は欠損を表す．

ここで，完全データ 2009 から生成した不完全マトリクスの 30 ケースの $RMSE$ の平

表 6.1: 完全データ 2009 から生成した不完全マトリクスの 30 ケースの $RMSE$ の平均 [30]

	EM タイプ IRT		MD	
	Training	Test	Training	Test
Training 80%	0.3578	0.3818	0.2147	0.4542
Test 20%	0.0022	0.0087	0.0046	0.0138
Training 20%	0.3270	0.4037	0.0299	0.4518
Test 80%	0.0116	0.0062	0.0025	0.0073

上は平均，下は標準偏差

均を表 6.1 に示す．Test の $RMSE$ の平均値を見ると，EM タイプ IRT の $RMSE$ が小さく予測精度が良いことを示している．

6.2.2 適応型システム結果（不完全データ 2013）の予測

不完全データ 2013 に対し，EM タイプ IRT および MD による不完全マトリクスの予測を行う．ここでは，もとのデータを Training と Test に 9 対 1 に分けて評価を行う．分け方は，無作為抽出による．完全データ 2009 の場合と同様に，Training と Test を 30 セット作成する．また，不完全データ 2013 には，予備テストの試験結果が含まれているため，適応型システムの結果のみを不完全データ 2013A，予備テストも合わせたものを不完全データ 2013B と再定義する．

表 6.2 にこのときの $RMSE$ を示す．表 6.2 を見ると，Test の $RMSE$ に対しては，不完全データ 20131，B2 ともに EM タイプ IRT の方が小さいことが分かる．特に，不完全データ 2013B の結果は EM タイプ IRT の方がより小さく，予測精度が良いことを示している．また，不完全データ 2013A と不完全データ 2013B の Test の $RMSE$ を比較すると，MD に比べて EM タイプ IRT の Test の $RMSE$ の方が変化が大きい．不完全データ 2013A と不完全データ 2013B の欠損率（マトリクス全体の要素数に対する欠損要素の割合）はそれぞれ 95.1% と 85.7% であり，EM タイプ IRT は欠損率が小さくなることによって精度が良くなっておりデータ数の影響を受けていると考えられる．

表 6.2: 不完全データ 2013 を 9:1 の Training と Test に分けた 30 ケースの $RMSE$ の平均 [30]

	EM タイプ IRT		MD	
	Training	Test	Training	Test
不完全データ 2013A	0.340	0.521	0.0133	0.523
	0.00440	0.0299	0.00116	0.0295
不完全データ 2013B	0.380	0.443	0.145	0.501
	0.00162	0.0154	0.00410	0.0178

上は平均，下は標準偏差

6.3 収束性の検討

EM タイプ IRT は，その予測手順の性質から，予測値が収束に至るまでの $RMSE$ の挙動は単調ではないと考えられる．そこで本節では，不完全データ 2013 における 30 ケースの Training に対する予測値の収束を見ることで，その収束性について検討する．

図 6.3(a) に不完全データ 2013A を用いた場合について，図 6.3(b) に不完全データ 2013B を用いた場合について，30 ケースの Training による $RMSE$ の収束の様子を示す．どちらの場合も，30 ケースすべてが単調に減少している．

次に，30 ケースの Training による対数尤度 $\log L$ の収束の様子を，不完全データ 2013A については図 6.4(a) に，不完全データ 2013B については図 6.4(b) に示す．ここで言う対数尤度とは，観測された値に対応した観測値による尤度である．欠損値を予測した値は含まれない．どちらの場合の $\log L$ も，30 ケースすべてが単調に増加している．

EM タイプ IRT では，計算の過程で，欠損要素を予測値で置き換える操作を繰り返し，そこで得られる観測値と予測値を組み合わせたマトリクスを次の初期値としている．つまり，計算の更新ごとに，扱うデータが異なっていることになる．そのため，計算の更新ごとに得られる $RMSE$ および $\log L$ の値は，単調に減少または増加するとは限らない．実際に [14] の場合には単調性は見られなかった．しかし，今回のように経験的には多くの場合で同じ値に収束している．

表 6.3: 制約がある場合とない場合の MD における $RMSE$ の平均 [30]

	制約付き MD	制約なし MD
不完全データ 2013A	0.523	0.557
	0.0295	0.0287
不完全データ 2013B	0.501	0.575
	0.0178	0.0178

上は平均，下は標準偏差

6.4 考察

6.4.1 制約付き MD について

本研究では，不完全マトリクスに対する予測手法として，EM タイプ IRT と MD の比較を $RMSE$ による評価で行った．EM タイプ IRT では，予測値 T は確率値として得られるため，$0 \leq T \leq 1$ を満たす．一方，MD では，予測値は 1 以上の値や 0 未満の値も含まれる．そこで，MD では，$0 \leq T \leq 1$ を満たすように，予測値 T が 0 未満の場合は 0 に，1 以上の場合は 1 に制約を付ける操作が行われる．ここで，制約なしの場合を考える．表 6.3 を得る．制約なしの場合，明らかに予測精度は悪くなっている．不完全マトリクスの欠損部分を予測する場合，0/1 の範囲に制約をつける操作は自然である．

6.4.2 誤分類率による評価

もとのマトリクスデータは 0/1 の 2 値データであるため，予測マトリクスもまた 2 値に分類する問題として捉えることができる．そこで，得られる予測マトリクスを 2 値データに分類したときの誤分類率について考察する．簡単のため，観測値 $x_{i,j}$ に対する予測値 $\hat{x}_{i,j}$ を改めて

- $(x_{i,j} = 0) \cap (0 \leq \hat{x}_{i,j} < 0.5) \ \Rightarrow \ T = 0$
- $(x_{i,j} = 0) \cap (0.5 \leq \hat{x}_{i,j} \leq 1) \ \Rightarrow \ T = 1$
- $(x_{i,j} = 1) \cap (0 \leq \hat{x}_{i,j} < 0.5) \ \Rightarrow \ T = 1$

表 6.4: Test データの誤分類率の平均 [30]

	EM タイプ IRT	MD
不完全データ 2013A	0.390	0.337
	0.0608	0.0491
不完全データ 2013B	0.287	0.259
	0.0247	0.0239

上は平均，下は標準偏差

- $(x_{i,j} = 1) \cap (0.5 \leq \hat{x}_{i,j} \leq 1) \ \Rightarrow \ T = 0$

と分類した場合を考える．T は 2 値関数であり，$T = 1$ のとき誤分類，$T = 0$ のとき正分類を表す．このときの 30 ケースの Test に対する誤分類率の平均 $\sum_{i,j} T/\#T$ は，表 6.4 となる．この結果を見ると，MD のほうが誤分類率が低いことが分かる．両手法とも，データ数の増加によって誤分類率は低下しており，データによる誤分類率のばらつきも小さいことが分かる．

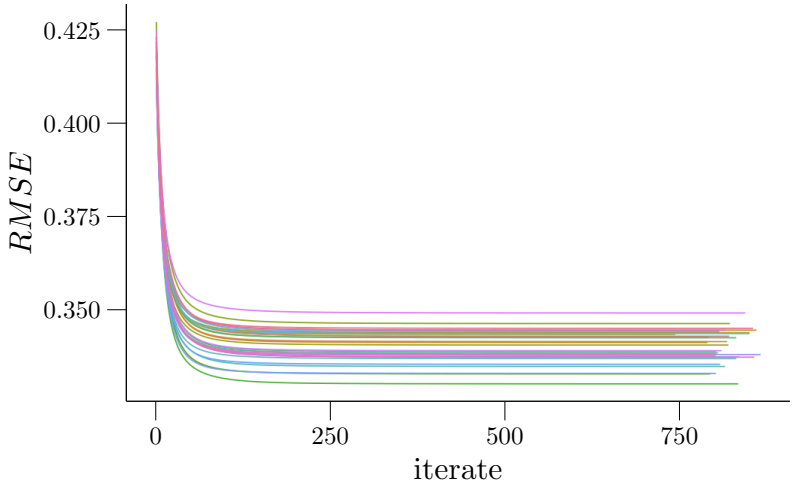
以上のことから，両手法の予測値を連続値として評価すると EM タイプ IRT のほうが優れているが，0/1 の離散値にすると MD が優れていることが分かった．

6.5 本章のまとめ

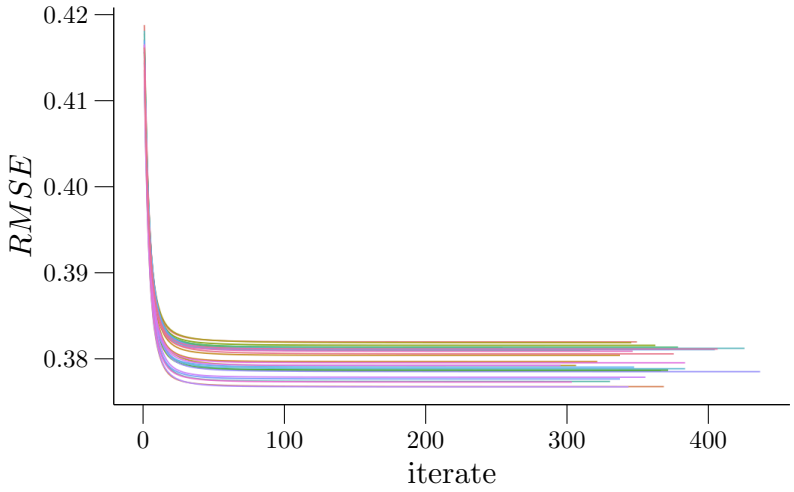
項目反応理論（IRT）は，あるテストの問題を受験者が解答したときの解答パターンのマトリクスから，テストの問題項目特性値および受験者能力評価値を推定することができる．IRT は通常，すべての問題に全員が解答している完全マトリクスを用いるが，ここでは未解答部分を含む不完全マトリクスに IRT が適用できる EM タイプ IRT について，不完全マトリクスの予測法の 1 つであるマトリクス分解法（MD; matrix decomposition method）[12, 26] と比較した．EM タイプ IRT は，データの背後にロジスティックモデルの確率構造を仮定し，不完全マトリクスでの観測された要素の値を用いて観測されていない空要素の値を確率的に予測するものである．一方で，マトリクス分解法は，データの背後に確率構造を仮定しないノンパラメトリックな方法であり，推薦システムなどに用い

られている．しかしながら，受験者の資質がある程度予測できて，問題の解答パターンもある程度想定される確率分布の下で変動すると仮定できるような場合，背後に確率分布の構造を仮定した方が推定精度が良くなることも考えられた．つまり，能力評価試験のような受験者の特性が強く反映される不完全マトリクスの場合，EM タイプ IRT は MD よりも有効に働く可能性があった．

本章では，欠損部分の答えが分かっているケースと，分かっていないケースにおいて両手法の比較を試みた．前者については，筆記試験によって実際に得られた完全マトリクスから不完全マトリクスを生成した．後者については，適応型システムによって実際に得られた試験データを用いた．その結果，最小二乗誤差からは，どちらのケースにおいても，EM タイプ IRT はマトリクス分解法よりも良い予測精度を出すことが分かった．実際のテストではロジスティックモデルが適切なのかノンパラメトリックが適切なのかは不明であるが，本提案法である EM タイプ IRT の実際問題への適用の可能性を示しており，実用的な新しい教育支援システムとして有用であることが期待できる．

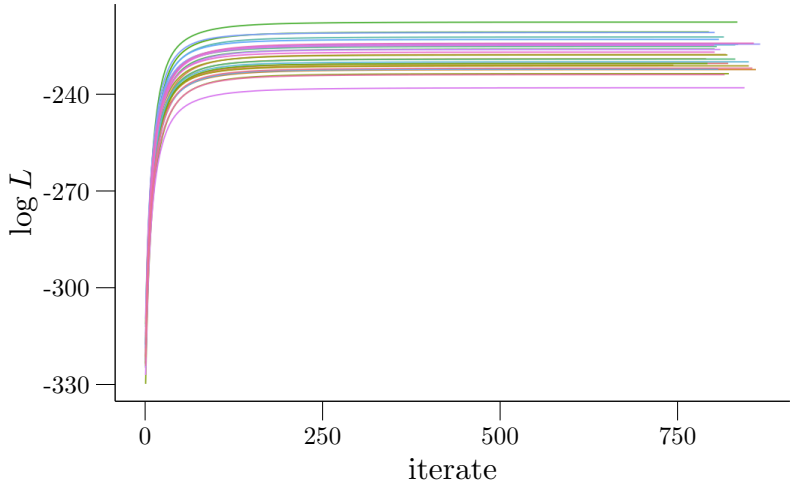


(a) 不完全データ 2013A

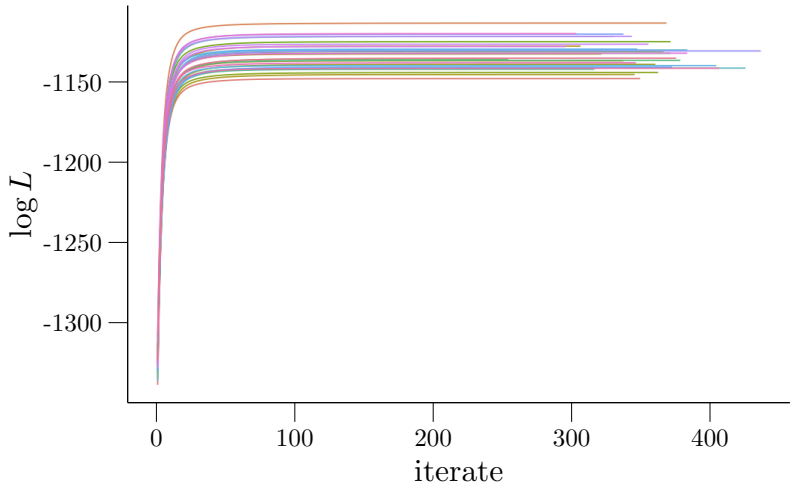


(b) 不完全データ 2013B

図 6.3: 9:1 の Training と Test に分けた 30 ケースに EM タイプ IRT を適用したときの Training に対する $RMSE$ の変化 [30]



(a) 不完全データ 2013A



(b) 不完全データ 2013B

図 6.4: 9:1 の Training と Test に分けた 30 ケースに EM タイプ IRT を適用したときの Training に対する $\log L$ の推移 [30]

第 7 章

おわりに

本論文では，まず，第 I 部で，項目反応理論（IRT）の基本的な概念と，これまでに考案されてきたパラメータの推定方法について紹介するとともに，IRT に触れる機会を改善する試みとして，容易に利用可能な Web アプリケーションの開発を行った．次に，第 II 部では，IRT を利用した適応型試験について議論した．ここでは，適応型試験による能力評価は，一部の大規模な検定試験などでのみ利用されており，大学などの教育機関ではあまり利用されていない理由について述べるとともに，その解決策として，新たな手法を提案し，それにもとづいた適応型試験の運用法について，実例を交えて述べた．本論文での主要の提案および得られた結果を以下に示す．

第 2 章では，IRT の基本的な概念と，データとして扱われる試験結果は [0,1] の有理数に拡張可能であることを述べた．また，IRT において主に用いられるパラメータ推定法について紹介し，それらの推定法について考察を行った．テストによる評価法として現在でも主に用いられている古典的な方法は問題毎にあらかじめ与えられた素点の得点を合計して受験者全体の中の相対的な位置を決めるという方法である．この方法では素点の与え方により評価結果が変わるため，より公平で公正な評価法として項目反応理論（IRT）が考案され，公的なテスト（例えば TOEFL）に活用されている．IRT は，受験者 i（ユーザー）が問題 j（アイテム）を解く確率 $P(\theta_i, \phi_j)$ がロジスティック分布関数に従うと仮定して，$(i,\ j)$ 要素に 2 値反応パターンを持つ完全マトリクスからパラメータ θ_i（受験者の能力）と ϕ_j（問題の特性）を推定することによって受験者 i の能力 θ_i を求める方法で

ある．このモデルでは θ_i と ϕ_j を同時に推定することによる困難さを伴うため，ベイズ，EM（expectation-maximization），MCMC（マルコフチェーンモンテカルロ）といった手法を援用しなければ推定値が得られない．これらの推定法をシミュレーションデータに対して適用し，それぞれの推定精度に大きな違いはないことを示した．MCMC はシンプルなアルゴリズムであり，容易に実装可能である反面，計算に多くの時間を要する．一方で，ベイズ，EM などを組み合わせた最尤推定では，チューニングパラメータの設定などに注意が必要だが，計算は高速であり，適応型試験のようなリアルタイム処理では実用的な方法である．

第 3 章では，IRT に触れる機会を改善する試みとして，容易に利用可能な Web アプリケーションシステムの開発を行った．これは，インターネットを通じて，クライアント側の PC 上で動作するもので，データの入出力には一般的に普及している Microsoft Office Excel を利用する．誰もがより簡単な操作で IRT に触れることができるものを目指したアプリケーションである．世界的に著名な BILOG-MG による推定結果との比較を行い，BILOG-MG と本システムとで矛盾しない推定結果が得られることを確認した．

第 4 章では，IRT を利用した e-learning システムとして，オンライン適応型試験システムについて説明する．このシステムを運用する上で必要な事前準備について述べるとともに，実際に運用して得られる不完全マトリクスについて説明した．通常の IRT では全受験者に同じ問題が課されているため，受験者によっては適切でない問題も与えられることがある．情報量の観点からは，受験者の能力を最も精度良く推定するには受験者の能力に合致させた問題を集中的に与えることが好ましい．こうすることで，速く正確に受験者の能力を測定することが可能になる．そこで，IRT にこのような仕組みを持たせる adaptive なテスト法が開発されてきた．これはコンピュータ支援テスト法によく馴染むためオンラインテストに用いられる．ただし，受験者に適切な問題が選択されて与えられるため，2 値反応パターンマトリクスは不完全になり，従来の IRT での推定法は使えない．あらかじめ問題の特性 ϕ_j が与えられていれば θ_i の推定は可能になる．このため，一定程度（4-500 人とも言われている）の受験者集団に予備テストを受けてもらって特性 ϕ_j を準備する必要がある．しかし，受験者が増えてくると予備テストでの受験者特性と本テストでの受験者特性の間に開きが出る恐れがあった．

第 5 章では，adaptive なオンラインテストの中で，受験者が増えても問題の特性 ϕ_j が受験者特性と開きが出ないように，受験者の評価を行うと同時に問題の特性 ϕ_j も受験者が増える毎に更新していく方法を提示した．具体的には，不完全マトリクスにおける空き要素（ある受験者には出題されていない問題）での反応を適当に与え，不完全マトリクスを一旦完全マトリクス（ただし 2 値ではなく [0,1] の有理数に拡張）に変え，通常の IRT を（反応に有理数を許すように）拡張した枠組みでパラメータを推定し，推定された反応パターンを観測値と比較してその差（尤度と距離の 2 つの規準から）が最小になるように更新しながら繰り返しパラメータを推定していく方法である．提案法によって θ_i と ϕ_j を同時に効率よく推定できているかを，実際の試験データを模擬したシミュレーションによって試みた結果，もとのパラメータを正しく再現できていることを確認した．また，適応型システムによって実際に得られた問題数 30 問程度，受験者数 200 人程度の不完全マトリクスに試みた結果，初期に空き要素に与える反応パターンをどのように変えても推定値は同じ収束値に向かい推定がうまくいっていることが経験的に示された．

第 6 章では，不完全マトリクスの予測法の 1 つとして一般に有名な推薦システムとしてのマトリクス分解法と本提案法とを比較し，その有効性を述べた．提案の問題解決内容は，ユーザーとアイテムから構成される評価パターンの不完全マトリクスからユーザーの嗜好を予測して商品を推薦する推薦システムとよく似ている．ただし，推薦システムではユーザーとアイテムの間の確率的構造は仮定していないため，ノンパラメトリックな枠組みとなる．そこで，実際のテスト結果をもとに提案法による推定結果と推薦システムによる結果の比較を行った．実際のテストではロジスティックモデルが適切なのかノンパラメトリックが適切なのかは不明であるが，比較の結果，最小 2 乗誤差からは提案法がわずかではあるが良い結果を与えた．

以上の検討から，本提案法は，実際問題への適用の可能性を示しており，実用的な新しい教育支援システムとして有用であることが期待できる．また，本来，適応型試験では事前準備として予備テストを行い，項目特性を調べておく必要があるが，本提案手法によって，予備テストそのものを適応型試験の中で行える可能性があることが示唆される．予備テストを必要としない適応型試験システムは，大幅な時間と労力のコストを削減できるため，本提案手法は大きな可能性を秘めたものであると考えられる．

謝辞

本論文は九州工業大学大学院情報工学府において筆者が行ってきたユーザー・アイテムの応答から構成された確率構造を持つ不完全マトリクスからのユーザーとアイテムの評価法に関する研究をまとめたものです．

本研究を遂行するにあたり，終始懇切丁寧なご指導とご鞭撻を賜りました九州工業大学大学院情報工学研究院 廣瀬 英雄 教授 に心から感謝申し上げます．

また，本論文をまとめるにあたり，有意義なご助言を賜りました九州工業大学大学院情報工学研究院 野田 秀樹 教授，九州工業大学大学院情報工学研究院 岡本 卓 教授，九州工業大学大学院情報工学研究院 宮野 英次 教授，九州工業大学大学院情報工学研究院 竹内 章 教授 に誠意を申し上げます．

また，終始ご激励を頂いた九州工業大学大学院情報工学研究院 廣瀬研究室の方々に厚く御礼申し上げます．

参考文献

[1] Score Evaluation Service using IRT. http://ume98.ces.kyutech.ac.jp/score-service/.

[2] R. Bell, J. Bennett, Y. Koren, and C. Volinsky. The million dollar programming prize. *Spectrum, IEEE*, Vol. 46, No. 5, pp. 28–33, 2009.

[3] Bilog-MG. http://www.ssicentral.com/irt/index.html, 2005.

[4] R.D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 1981.

[5] R.D. Bock and M. Lieberman. Fitting a response model for n dichotomously scored items. *Psychometrika*, 1970.

[6] L. L. Cook and D. R. Eignor. Irt equating methods. *Educational Measurement*, Vol. 10, No. 3, pp. 37–45, 1991.

[7] R.J. De Ayala. *The theory and practice of item response theory.* Guilford Press, 2009.

[8] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 1–38, 1977.

[9] R.K. Hambleton. *Fundamentals of item response theory*, Vol. 2. Sage Publications, Incorporated, 1991.

[10] R.K. Hambleton and H. Swaminathan. *Item response theory: Principles and applications*, Vol. 7. Springer, 1984.

[11] H. Hirose. An optimal test design to evaluate the ability of an examinee by using the stress–strength model. *Journal of Statistical Computation And Simulation*, Vol. 81, No. 1, pp. 79–87, January 2011. first published 12/09/2009 (iFirst).

[12] H. Hirose, T. Nakazono, M. Tokunaga, T. Sakumura, S.M. Sumi, and J. Sulaiman. Seasonal infectious disease spread prediction using matrix decomposition method. In *4th International Conference on Intelligent Systems, Modelling and Simulation, ISMS 2013.*, pp. 152–156, Bangkok, Thailand., Jar 2013. The Royal Society.

[13] H. Hirose and T. Sakumura. An accurate ability evaluation method for every student with small problem items using the item response theory. In *Computers and Advanced Technology in Education, CATE 2010.*, pp. 152–158. ACTA Press, 2010.

[14] H. Hirose and T. Sakumura. Item response prediction for incomplete response matrix using the em-type item response theory with application to adaptive online ability evaluation system. In *Teaching, Assessment and Learning for Engineering (TALE), 2012 IEEE International Conference on*, pp. T1A–6–T1A–10, Aug. 2012.

[15] H. Hirose, T. Sakumura, and S. Ichii. A recommendation algorithm that assumes a probabilistic structure and its application to questionnaire data. In *in IPSJ SIG Technical Report.*, pp. 1–7, Fukuoka, Japan., Mar 2011.

[16] C.N. Mills, M.T. Potenza, J.J. Fremer, and W.C. Ward. *Computer-based testing: Building the foundation for future assessments.* Lawrence Erlbaum, 2002.

[17] R.J. Mislevy. Estimating latent distributions. *Psychometrika*, Vol. 49, No. 3, pp. 359–381, 1984.

[18] R.J. Mislevy. Bayes modal estimation in item response models. *Psychometrika*, Vol. 51, No. 2, pp. 177–195, 1986.

[19] Netflix. Netflix prize. `http://www.netflixprize.com/`.

[20] Netflix. Netflix update: Try this at home. `http://sifter.org/si-mon/`

journal/20061211.html.

[21] J. Neyman and E.L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, Vol. 16, No. 1, pp. 1–32, 1948.

[22] R.J. Patz and B.W. Junker. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 2, p. 146, 1999.

[23] T. Sakumura and H. Hirose. Test evaluation system via the web using the item response theory. *Information*, Vol. 13, No. 3, pp. 647–656, May 2010.

[24] T. Sakumura, T. Kuwahata, and H. Hirose. An adaptive online ability evaluation system using the item response theory. In *in Education and e-Learning (EeL2011).*, pp. 51–54. Global Science and Technology Forum (GSTF), 2011.

[25] H.K. Suen and P.S.C. Lee. *Constraint optimization: An alternative perspective of IRT parameter estimation*, chapter 17, pp. 289–300. Norwood, NJ: Ablex., 1994.

[26] S. Takimoto and H. Hirose. Recommendation systems and their preference prediction algorithms in a large-scale database. *Information*, Vol. 12, No. 5, pp. 1165–1182, 2009.

[27] W.J. van der Linden and R.K. Hambleton. *Handbook of modern item response theory*. Springer, 1996.

[28] W.M. Yen, G.R. Burket, and R.C. Sykes. Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika*, Vol. 56, No. 1, pp. 39–54, 1991.

[29] 池田央. 現代テスト理論. 朝倉書店, 1994.

[30] 作村建紀, 徳永正和, 廣瀬英雄. EM タイプ IRT による不完全マトリクスの完全化とその応用. 情報処理学会論文誌. 数理モデル化と応用, Vol. 7, No. 2, pp. 17–26, Nov 2014.

[31] 作村建紀, 廣瀬英雄. 項目反応理論を使った最適能力判定システムについて. 日本

オペレーションズ・リサーチ学会九州地区における若手 OR 研究交流会 2010, Oct 2010.

[32] 作村建紀, 廣瀬英雄. IRT を用いた adaptive online system の実装と評価. 日本計算機統計学会第 27 回シンポジウム講演論文集, pp. 27–30, Nov 2013.

[33] 鈴木敬一, 月原ゆき, 廣瀬英雄. IRT を用いた数学テストの一評価法. 日本行動計量学会大会第 35 回大会, pp. 99–100, September 2007.

[34] 鈴木敬一, 月原由紀, 廣瀬英雄. IRT を用いた数学テストの評価. 2007 年度統計関連学会連合大会, p. 280, September 2007.

[35] 月原由紀, 作村建紀, 廣瀬英雄. IRT における項目レベル最適選定法. 2008 年度統計関連学会連合大会, p. 106, Sep 2008.

[36] 月原由紀, 作村建紀, 廣瀬英雄. IRT を用いた解析学試験評価と e-learning 支援の試み. PC Conference 論文集, pp. 123–124, Aug. 2010.

[37] 月原由紀, 鈴木敬一, 廣瀬英雄. IRT を援用した e-learning システムへの試み : 大学数学の基礎教育. 電気関係学会九州支部第 60 回連合大会, Vol. 10-2A-06, p. 371, September 2007.

[38] 月原由紀, 鈴木敬一, 廣瀬英雄. IRT を用いた数学テストの e-learning システムへの実装 : 分野別問題への適用. 電子情報通信学会総合大会 2008, Vol. D-15-43, p. 237, March 2008.

[39] 月原由紀, 鈴木敬一, 廣瀬英雄. 項目反応理論による評価を加味した数学テストと e-learning システムへの実装の試み. コンピュータ＆エデュケーション（CIEC）, Vol. 24, pp. 70–76, Jun. 2008.

[40] 豊田秀樹. 項目反応理論 入門編—テストと測定の科学—. 朝倉書店, 2002.

[41] 文部科学省. http://www.mext.go.jp/.