НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

фізико-математичний факультет

кафедра математичного аналізу та теорії ймовірностей

«На правах рукопису»
УДК 519.237

До захисту допущено:
Завідувач кафедри
_____ Олег КЛЕСОВ
«8» червня 2022 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-науковою програмою «Страхова та фінансова математика»

зі спеціальності 111 «Математика»

на тему: « Статистичні методи знаходження оцінок латентних параметрів некомпенсаторних MIRT моделей**»**

Виконав:
студент ІІ курсу магістратури, групи ОМ-01мн
Кононенко Сергій Васильович _____

Науковий керівник:
кандидат фізико-математичних наук, доцент
Круглова Наталія Володимирівна _____

Рецензент:
кандидат фізико-математичних наук,
доцент кафедри математики та
теоретичної радіофізики
Єфіменко Світлана Володимирівна.
КНУ ім. Тараса Шевченка _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2022 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
фізико-математичний факультет
кафедра математичного аналізу та теорії ймовірностей

Рівень вищої освіти – другий (магістерський) за освітньо-науковою програмою

Спеціальність – 111 «Математика»

Освітньо-наукова програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олег КЛЕСОВ
«03» лютого 2022 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Кононенку Сергію Васильовичу

1. Тема дисертації « Статистичні методи знаходження оцінок латентних параметрів некомпенсаторних MIRT моделей », науковий керівник дисертації Круглова Наталія Володимирівна, кандидат фізико-математичних наук, доцент, затверджені наказом по університету від «25» травня 2022 №НС/372/2022

2. Термін подання студентом дисертації «7» червня 2022 року.

3. Об’єкт дослідження: тестові Модульні контрольні роботи з предмету «Теорія ймовірностей» для студентів факультету НТУУ “КПІ ім. Ігоря Сікорського”.

4. Предмет дослідження: MIRT-моделі.

5. Перелік завдань, які потрібно розробити:

1. Виконати огляд літературних джерел MIRT-моделей, методів визначення розмірностей моделей та критеріїв їх оцінки.
2. Зробити математичний опис MIRT-моделей
3. Виконати імітаційне моделювання описаних моделей та аналіз їх результатів

6. Перелік ілюстративного матеріалу: контурні лінії питань, поверхні очікуваних оцінок іспитників, 15 слайдів для проектора

7. Дата видачі завдання: 2 лютого 2022 року.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Отримання та обговорення основної задачі дипломної роботи	05.02.21 – 08.02.21	виконано
2.	Огляд основних понять сучасної теорії тестів	09.02.21- 10.03.21	виконано
3.	Огляд критеріїв оцінки MIRT-моделей	11.03.21 – 25.03.21	виконано
4.	Математичний опис MIRT-моделей	26.03.21 – 13.04.21	виконано
5.	Візуалізація даних шляхом імітаційного моделювання в середовищі RStudio	13.04.21 – 03.05.21	виконано
6.	Аналіз результатів моделювання	04.05.21 – 14.05.21	виконано
7.	Висновки	15.05.21 – 24.05.21	виконано
8.	Оформлення магістерської дисертації	25.05.21 – 05.06.21	виконано

Студент Сергій КОНОНЕНКО

Науковий керівник Наталія КРУГЛОВА

РЕФЕРАТ

Магістерська дисертація: 65 сторінок, 6 ілюстрацій, 5 таблиць, 25 формул, 43 першоджерела, 15 слайдів для проектора.

Актуальність магістерської дисертації зумовлено широким застосуванням сучасної теорії тестів до створення тестувань у різних предметних галузях (когнітивні, професійні, психологічні науки тощо). Для розв'язання цієї проблеми необхідно знайти коректну модель для тестових даних.

Метою роботи є знаходження оцінок латентних параметрів некомпенсаторних MIRT-моделей за допомогою статистичних методів.

Об'єктом дослідження є тестові модульні контрольні роботи з предмету «Теорія ймовірності» для студентів факультету ФІОТ.

Предметом дослідження є знаходження коректної MIRT-моделі відповідно до тестових даних.

Методи дослідження: аналіз (багатовимірних MIRT-моделей та критеріїв їх оцінки), формалізація (створення математичного опису MIRT-моделей), експеримент (моделювання MIRT-моделей), порівняння (аналіз результатів моделювання оцінок тестових питань, порівняння результатів застосованих MIRT-моделей).

Наукова новизна полягає у наступному: було запропоновано та застосовано MIRT-моделі до результатів тестування з предмету «Теорія ймовірності».

Практичне значення одержаних результатів роботи полягає у їх застосуванні до результатів тестування з предмету «Теорія ймовірності» та подальшому аналізі тестових питань.

Ключові слова: сучасна теорія тестів, багатовимірні MIRT-моделі, підбір моделі для даних, метод граничної максимальної правдоподібності, багатовимірна складність питань, багатовимірна диференціююча здатність питань.

ABSTRACT

Master’s thesis contains 65 pages, 6 figures, 5 tables, 25 formulas, 43 primary sources, 15 slides for projector.

The relevance of the master's thesis is justified by the wide application of modern test theory to the tests designing in various subject areas (cognitive, professional, psychological sciences etc.). To solve this problem, it is necessary to find the proper model for the test data.

The aim of the work is to find estimates of latent parameters of non-compensatory MIRT-models using statistical methods.

The object of the research is estimation of modular tests on the subject "Probability Theory" for students of the Faculty of FIOT.

The subject of the research is to find the correct MIRT model according to test data.

Research methods: analysis (multidimensional MIRT-models and criteria for their evaluation), formalization (design of MIRT-models mathematical description), experiment (modeling of MIRT-models), comparison (simulation results analysis of test questions, comparison of MIRT-models).

The scientific novelty is the following: MIRT-models have been proposed and applied to the results of testing in the subject "Probability Theory".

The practical significance of the obtained results of the work lies in their application to the test results in the subject "Probability Theory" and further analysis of test questions.

Key words: item response theory, multidimensional MIRT models, model data fit, the maximum likelihood criterion, multidimensional difficulty of the test item, multidimensional discrimination index of the test item.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ ТА ТЕРМІНІВ

IRT – сучасна теорія тестів;

MIRT – багатомірна сучасна теорія тестів;

1-PL модель – однопараметрична логістична модель;

2-PL модель – двопараметрична логістична модель;

3-PL модель – трьохпараметрична логістична модель;

M1-PL – багатовимірна однопараметрична логістична модель;

M2-PL – багатовимірна двопараметрична логістична модель;

M3-PL – багатовимірна трьохпараметрична логістична модель;

C2-PL – компенсаторна MIRT-модель;

PC2-PL – некомпенсаторна MIRT-модель;

MMLE – Оцінка граничної максимальної правдоподібності;

MLE – оцінка максимальної правдоподібності;

MAP – метод апостеріорного максимуму;

EAP – метод очікуваного апостеріорного максимуму;

BIC – баєсовий інформаційний критерій;

AIC – інформаційний критерій Акаіке;

KLIC – критерії Кульбака-Лейблера;

SABIC – BIC критерій з поправками розміру вибірки;

HQ – інформаційний критерій Ханна-Куіна;

МКР – модульна контрольна робота;

ВСТУП

На сьогоднішній день актуальною є проблема оцінки здібностей людини в різних сферах: когнітивні, професійні, психологічні тощо. Одним із поширених методів оцінки якостей людини є застосування різноманітних тестувань, що дозволяють абстрагуватися від особистісних відносин між іспитником та людиною, що виконує оцінювання. В період пандемії і війни в Україні тестування стало єдиним можливим способом оцінити здібності іспитника. Це дозволяє визначити обґрунтовану неупереджену оцінку здібностей людини. Проте, виникає проблема створення адекватних тестових пакетів, що мають вірно оцінити здібності за однією чи кількома рисами. Також для різних сфер застосування можуть виникати специфічні вимоги до створення тестів та їх властивостей.

Одним із способів, що дозволяє математично обґрунтовано побудувати тестування, є застосування сучасної теорії тестів. Ця теорія дозволяє із певною точністю гарантувати диференціюючу здатність та складність питань, що є необхідним для оцінювання іспитників у всьому спектрі їх рівня підготовки.

В залежності від кількості ознак, які оцінюються у тесті, існують одновимірні та багатовимірні MIRT-моделі.

У даній магістерській дисертації для аналізів результатів тестування з предмету «Теорія ймовірностей» було використано багатовимірні MIRT-моделі.

РОЗДІЛ 1
ОГЛЯД ОДНОВИМІРНИХ ТА БАГАТОВИМІРНИХ IRT-МОДЕЛЕЙ

1.1. Сучасна теорія тестів

Сучасна теорія тестів (IRT - Item response theory) та багатовимірна сучасна теорія тестів (MIRT - Multidimensional Item Response Theory) пов’язує латентні параметри (навички, досягнення, здібності людини тощо) із ймовірністю правильної відповіді на них [1-3].

Головною метою даних моделей є знаходження зв’язку між вказаними вище характеристиками та тестовими завданнями шляхом оцінювання осіб, що проходять тестування. Якщо враховується лише одна характеристика, то це IRT-модель, якщо декілька – MIRT-модель. Очевидно, що MIRT є розширенням IRT.

Прикладом застосування IRT-моделі є розв’язання рівняння. Це завдання містить лише математичний вираз без додаткового тексту. Щоб вирішити цю задачу, людина потребує лише навику алгебраїчних маніпуляцій із символами [1]. Прикладом використання MIRT-моделі є вирішення завдання у вигляді текстової задачі, що оцінює як алгебраїчні навички, так і розуміння прочитаного тексту.

Одним з базових припущень IRT є одновимірність, тобто, кожне завдання оцінює лише одну характеристику. На противагу, MIRT-моделі доречніші у випадку, коли завдання оцінює більш ніж одну характеристику (коли це є метою тесту апріорно, або існує проблема визначення кількості окремих характеристик) [4].

Доведено, що більшість питань в когнітивних, педагогічних і психологічних тестах є багатовимірними і цю розмірність завжди потрібно перевіряти, а не припускати [1-2, 5]. Через це існує доволі велика кількість випадків, де одновимірні IRT-моделі використовуються некоректно [6].

Неправильне використання та застосування моделі може призвести до серйозних проблем з валідністю результатів тесту. Варто зазначити, що повне дослідження всього процесу від теоретичних розробок, до їх практичного застосування, є дуже об’ємним, тож треба було спростити цей шлях. Це й стало однією з найголовніших причин, що спонукали науковців до більш глибоких досліджень [3, 7].

1.2. Багатовимірні IRT-моделі

Важливість MIRT-моделей полягає в їх практичному застосуванні, що включає наступне [2, 8, 9] :

- оцінити здібності людини і розкласти їх на складові характеристики;
- допомога в прийнятті рішень про те, який тип шкали або підшкали обрати для повної оцінки тесту;
- змога адаптувати моделі даних в залежності від обмежень інформаційних методів;
- розкрити зміст взаємодії «питання-здібність»;
- забезпечення підтримки правильності конструкції тесту;
- змога бути адаптивними до використання в різних програмах (наприклад, розробка тестів, діагностика, функціонування диференціальних елементів тощо);
- здатність точно представляти та інтерпретувати сприйняття та результати людини, яка проходить тестування;

Важливість вимірності можна проілюструвати дослідженнями Алена та Вілсона [8] та Волеми і Хойтінка [10]. Аллен і Вілсон проводили дослідження стану здоров'я людей, аналізуючи і враховуючи розмірність. Мета полягала в тому, щоб побачити, чи краще виміряти характеристику саморегуляції в одному вимірі, чи в двох (наприклад, тип регуляції та

мотивація). Інструментом аналізу була оцінка саморегуляції, яка має шість підтестів. Одновимірність оцінки досліджувалась на двох підходах: комбінованому (сума балів у всіх підтестах) і послідовному (оцінки в межах кожного субтесту підсумовувалися, а кожен субтест розглядався в окремому аналізі). Багатовимірна модель була версією компенсаторної моделі Раша MIRT (див. рівняння 1.4). Результати дослідження показали, що саморегуляція найкраще вимірюється багатовимірною структурою. Автори стверджували, що перевагами використання моделі MIRT (в порівнянні з двома іншими одновимірними підходами) є менша похибка, окремі оцінки для обох вимірів і підвищена надійність оцінок. Волема та Хойтінк [10] також досліджували вимірність шляхом аналізу психологічного опитувальника шизотипового розладу особистості (SPQ), щоб з'ясувати: дихотомічні елементи оцінюють 2 чи 3 виміри. Тестування проходили амбулаторні та стаціонарні дорослі пацієнти психіатрії. Як і в попередньому дослідженні, була використана багатовимірна модель Раша. Результати показали, що шизотопія чітко має 3 виміри, які були визначені як позитивна шизотипія, негативна шизотипія та дезорганізованість. За попередніми дослідженнями автори виявили, що тривимірна структура є найкращою серед осіб з психічними розладами і без них.

Як було зазначено раніше, моделі MIRT є розширенням одновимірних моделей IRT. Детально опишемо дихотомічні моделі IRT.

Дихотомічні моделі – це моделі, які мають лише два результати: 1 (правильна відповідь), або 0 (неправильна відповідь). Після цього обов'язково розглянемо й дихотомічні MIRT моделі.

Однією з найскладніших одновимірних моделей IRT є трипараметрична логістична модель (3PL), математична форма якої має наступний вигляд:

$$P\left(U_{ij}=1\middle|\,\theta_j,\alpha_i,b_i,c_i\right)=\ c_i+(1-\ c_i)\frac{1}{1+\ e^{-1.7\alpha_i(\theta_j-b_i)}} \qquad (1.1)$$

де $P(U_{ij} = 1 | \theta_j, \alpha_i, b_i, c_i)$ – це ймовірність того, що j-тий іспитник зі здібністю θ дасть правильну відповідь на i-те питання, Uij - це відповідь на питання, (Uij = 1 – правильно, Uij = 0 – неправильно); α_i – параметр диференціюючої здатності, який свідчить про те, наскільки ефективно i-те питання може розрізняти студентів за рівнем їхньої здібності (як правило, бажаною є висока дискримінація завдань – вона показує, що слабкі респонденти завдання не вирішують, а сильні - вирішують); b_i – параметр складності питання (вказує, наскільки легке чи складне питання); c_i – ймовірність псевдо-випадкового вгадування (вказує, наскільки ймовірно, що іспитники можуть отримати правильну відповідь випадково, обираючи її серед запропонованих опцій).

Як видно, ця модель складається з трьох латентних параметрів тестових питань (α_i, b_i, c_i) і одного параметра іспитника (θ_j). Якщо виключити параметр вгадування, то 3PL-модель зведеться до 2PL логістичної моделі (2PL), а її математичною формою буде:

$$P(U_{ij} = 1 | \theta_j, \alpha_i, b_i) = \frac{1}{1 + e^{-1.7\alpha_i(\theta_j - b_i)}} \quad (1.2)$$

де всі параметри визначено вище. Наступною моделлю в цій групі є однопараметрична логістична модель (1PL), математичний запис якої є таким:

$$P(U_{ij} = 1 | \theta_j, \alpha, b_i) = \frac{1}{1 + e^{-1.7\alpha(\theta_j - b_i)}} \quad (1.3)$$

де для всіх питань a_i однакові. І останньою в даній групі є модель Раша (як окремий випадок 1PL), де всі a = 1:

$$P(U_{ij} = 1 | \theta_j, b_i) = \frac{1}{1 + e^{-1.7(\theta_j - b_i)}} \quad (1.4)$$

Як і в IRT, існує 4 дихотомічні моделі MIRT, які включають 3PL (M3PL), 2PL (M2PL), 1PL (M1PL) і модель Раша. Кожна з них має свої відмінні якості. Наприклад, M1PL серед інших є найбільш обмеженою. Це пред'являє високі вимоги до тестових даних, оскільки всі елементи мають

однаковий параметр дискримінації. З іншого боку M3PL є складною через додатковий параметр вгадування, який збільшує складність оцінювання, що, у свою чергу, вимагає більше інформації від тесту, наприклад, великий розмір вибірки. Тож було визнано, що M2PL найбільше підходить для багатьох умов тестування [11].

MIRT-моделі поділяються на компенсаторні й некомпенсаторні. Розглянемо детальніше кожну з них. Компенсаторна модель 2PL (C2PL) має такий вигляд:

$$P\left(U_{ij}=1\middle|\,\overrightarrow{\theta_j},\overrightarrow{\alpha_i},d_i\right)=\frac{1}{1+\ e^{-1.7\left(\overrightarrow{\alpha_i}\overrightarrow{\theta_j}-d_i\right)}}\ (1.5)$$

де параметр $\overrightarrow{\theta_j}$ – це вектор здібностей розмірності 1 x m, параметр $\overrightarrow{\alpha_i}$ – вектор диференціюючої здатності (за аналогією з IRT), m – кількість вимірів, d_i – допоміжний параметр, який обчислюється наступним чином: $d_i=-\alpha_i b_i$, тоді (1.5) можна записати як

$$P\left(U_{ij}=1\middle|\,\overrightarrow{\theta_j},\overrightarrow{\alpha_i},b_i\right)=\frac{1}{1+\ e^{-1.7\overrightarrow{\alpha_i}\left(\overrightarrow{\theta_j}-b_i\right)}}\ (1.6)$$

Суть компенсаторної моделі така: щоб збільшити ймовірність правильної відповіді на питання, кращі здібності за однією рисою компенсують слабкі здібності за іншою. Наприклад, іспитнику пропонується прочитати уривок про поточну подію - місцеві вибори, після цього відповісти на запитання про це. В даному випадку маємо оцінку відразу двох здібностей: розуміння прочитаного та знання про поточні події. Якщо людина знає про вибори, то це компенсує низьку здатність усвідомлювати прочитане. З іншого боку, якщо відмінно усвідомлює - то це компенсує брак знань щодо виборів.

Тепер розглянемо некомпенсаторну MIRT-модель 2PL (PC2PL). Визначається вона так:

$$P\left(U_{ij}=1\middle|\,\overrightarrow{\theta_j},\overrightarrow{\alpha_i},b_i\right)=\prod_{\ell=1}^{m}\frac{1}{1+\ e^{-1.7\vec{\alpha}_{i\ell}\left(\vec{\theta}_{j\ell}-b_{i\ell}\right)}}\ (1.7)$$

де параметр b_i – вектор складності, ℓ - вимір, а всі інші параметри визначені вище. Формулу (1.7) можна спростити до такого вигляду: $k = \prod_{\ell=1}^{m} p_\ell$, де k– це ймовірність правильної відповіді, а p_ℓ - це дріб з (1.7). При некомпенсаторній моделі здібності не компенсують одна одну. Це означає, що іспитнику потрібен високий рівень у всіх здібностях, щоб мати високі шанси правильно відповісти на питання і «ймовірність правильної відповіді на поставлене питання, що відповідає цій моделі, ніколи не може бути більшою, ніж ймовірність для компоненти p_ℓ з найменшою ймовірністю» [12]. Прикладом некомпенсаторної моделі є традиційна текстова задача, де представлено текстовий уривок і необхідні вхідні дані. Такий тест оцінює дві здібності: розуміння прочитаного та знання з математики. Якщо людина має чудові здібності розуміння прочитаного, але низькі математичні здібності, він зможе прочитати текст, але не зможе розв'язати задачу. Інший сценарій полягає в тому, що особа має високі математичні здібності, але низьку здатність сприймати текстовий матеріал. У цьому випадку вона не зможе розв'язати задачу, тому що не в змозі зрозуміти питання, яке їй задане [13].

Незважаючи на те, що обидві моделі виглядають дуже схожими (після перевизначення параметру d), між ними існують суттєві відмінності. Давайте їх розглянемо. Нехай в деякому питанні обчислено: α_1 = 1.5, α_2 = 0.5, b = 0, θ_1 = -1, θ_2 = 2 для обох моделей. Ймовірність правильної відповіді для компенсаторної моделі дорівнює 0,299. Щоб знайти ймовірність правильної відповіді для некомпенсаторної моделі (припустимо, що $b_1 = b_2 = 0$), компоненти перемножуються, в результаті отримаємо ймовірність 0,06. Це й пояснює різницю між моделями. Цей іспитник має низькі здібності по першій ознаці ($\ell = 1$) й кращі для другої ознаки ($\ell = 2$), тому при використанні компенсаторної моделі краща здібність підвищує ймовірність правильної відповіді, ніж при використанні некомпенсаторної

моделі. Але в некомпенсаторній моделі гірша здібність для першої ознаки зменшує ймовірність правильної відповіді.

З цього прикладу зрозуміло, що важливе питання щодо правильної відповіді на питання полягає в тому, чи є взаємодія здібностей іспитника компенсаторною чи ні. Компенсаторна модель є більш цілісною за своєю гіпотезою, і людина, що проходить тестування, використовує всі свої здібності одночасно, щоб дати відповідь. На противагу, некомпенсаторна модель має концепцію розділення різних навичок і знань, при відповіді на різні частини питання. Є багато літератури, в якій розповідається про те, які процеси застосовуються та які стратегії використовуються [14].

Відмінності між моделями добре ілюструють графічні зображення характеристичних поверхонь (item characteristic surfaces - ICS). Загалом, візуальне представлення моделей повинно враховувати характеристики питання в багатовимірному просторі, і оскільки ці моделі мають 2 виміри, візуалізація зв'язку між предметом і людиною повинна підтримувати обидва виміри одночасно. Приклади ICS для теоретичних компенсаторних і некомпенсаторних моделей показані на рисунках 1.1 і 1.2 відповідно. ICS для компенсаторної моделі (рис. 1.1) показує компенсаційний характер зв'язку. Подивившись на графік, можна побачити, що навіть якщо іспитник отримує нижчий бал за θ_2 (наприклад, $\theta_2 = 0$), він все одно може мати високу ймовірність правильної відповіді на питання, якщо оцінка θ_1 більше 1,5. Дивлячись на некомпенсаторну модель (рис. 1.2), можна побачити, що навіть з високим балом за першою здібністю людина з нульовою оцінкою за другою здібністю не відповість правильно на це питання. Якщо компенсаторна модель має цілу область, де ймовірність правильної відповіді приймає великі значення, то некомпенсаторна модель має точку, де досягається велике значення ймовірності, при цьому латентні параметри обох ознак повинні бути великими.

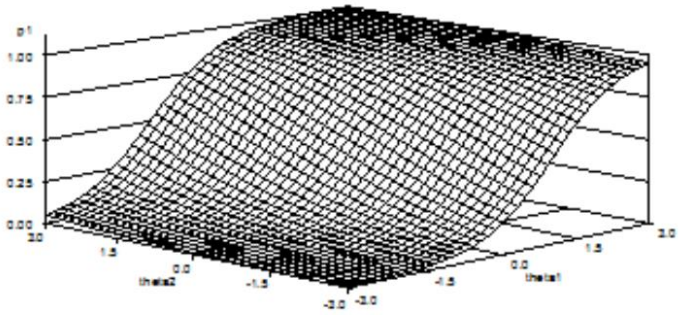


Рис. 1.1. Компенсаторна 2PL модель ICS з параметрами α_1 = 1.5, α_2 = 0.5, $d = 0$

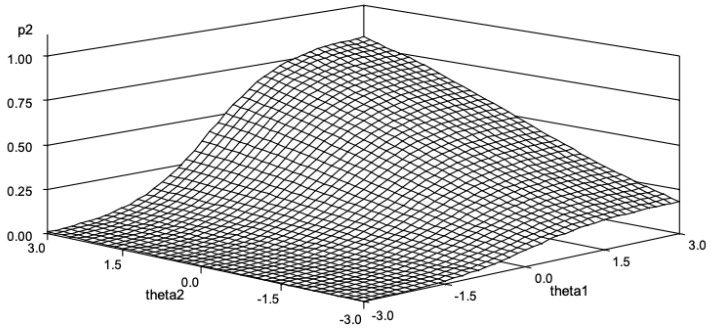


Рис. 1.2. Некомпенсаторна 2PL модель ICS з параметрами α_1 = 1.5, α_2 = 0.5, $d_1 = 0$, d_2 = 0

Компенсаторні моделі досліджувалися та застосовувалися на практиці більше, ніж некомпенсаторні моделі [1, 2, 15, 16]. Існує кілька можливих причин: теорію компенсаторності швидше прийняли [17], ці моделі були більш зручними для користувача, який використовує комп'ютерні програми, що дозволяють оцінювати параметри, і були використані з надійними та послідовними результатами. Однак бувають ситуації, коли дані насправді краще підходять до некомпенсаторної моделі. Коли це відбувається, а компенсаторні моделі все ще використовуються, результати невідповідності, у свою чергу, спричиняють серйозні наслідки, включаючи недостатню валідність, порушення припущень та проблеми з оцінками [18, 19]. До недавнього часу було важко досліджувати та використовувати

некомпенсаторні моделі, оскільки не було доступних програм для обчислення оцінок латентних параметрів. Зараз з'явилося подібне програмне забезпечення, але було дуже мало досліджень, зосереджених на даних моделях.

З огляду на це, інтерес і обсяг досліджень останніми роками зростає. Дослідження, проведене Болтом і Лаллом [20], оцінювало параметри для компенсаторних і некомпенсаторних моделей 2PL. Вибірки мали об'єми 1000 і 3000, тести складались з 25 і 50 питань і кореляція між здібностями була 0; 0,3 і 0,6. Результати показали, що параметри для обох моделей були оцінені успішно, загалом, компенсаторна модель мала менше помилок, краще оцінювалася, коли кореляція між здібностями була високою, і потрібні були менші вибірки.

Бабкок [15] досліджував некомпенсаторні моделі 2PL, використовуючи в якості прикладу дослідження Болта і Лалла [20]. Змінні мали кореляцію між ознаками 0.25, 0.50, 0.75, об'єми вибірок 1000, 2000 і 4000. Багатовимірні завдання мали загалом 50 питань, також були одновимірні компоненти, які відрізнялися за довжиною. Результати показали, що для точної оцінки необхідні надзвичайно великі розміри вибірки (наприклад, N = 4000), низька кореляція між ознаками та більше двох одновимірних елементів на ознаку. Навіть для вибірок великого об'єму було важко отримати точні оцінки з високою кореляцією між здібностями.

У ще одному дослідженні [21] метою якого було визначити, чи є різниця між моделями на рівні питання і на рівні тесту. Іншими словами, чи схожі вони, чи ні, і чи насправді вони вимірюють різні процеси? Досліджувався тест з 20 питань, з вибіркою об'єму N = 2000 та кореляцією між ознаками 0, 0.25, 0.50 та 0.75. На рівні тесту результати показали, що при більш високих кореляціях моделі майже не розрізнялися. На рівні питань було виявлено, що кількість помилок збільшується зі збільшенням кореляції між здібностями. Крім того, різниця в моделях зменшувалася зі

збільшенням кореляції між здібностями, але в цілому спостерігалася чітка відмінність між моделями.

Результати цих досліджень ілюструють відмінності між компенсаторною та некомпенсаторною моделями, а також важливість продовження їх вивчення через наступні причини:

- проблеми із коректністю моделей, якщо припущення компенсаторності є неправильним;
- у науковому товаристві продовжуються обговорення, чи слід питання вважати компенсаторними та до якої міри [1];
- недостатньо аргументів для вибору тієї чи іншої теорії та умов, за яких вони найбільш застосовні.

Однією з причин малої кількості досліджень може бути те, що навіть незважаючи на те що є програми, які запускають обидва типи моделей, існує нестаток програм, що коректно оцінюють некомпенсаторні моделі.

1.3. Структура тесту

Багатовимірна структура моделі в свою основу покладає опис зв'язків між ознаками. Для моделі M2PL структура описує, наскільки варіативність питання визначається ознакою 1 ($\ell = 1$), а наскільки — ознакою 2 ($\ell = 2$). На практиці часто зустрічаються три типи структур: проста структура, майже проста структура і складні структури [22]. Завдяки простій структурі одне питання вимірює лише одну ознаку, а декілька питань вимірюють декілька ознак. Завдяки майже простій структурі кожне питання вимірює домінуючу ознаку і одну або кілька другорядних ознак. Стоут та інші вчені в 1996 році [23] визначили питання з майже простою структурою як ті, що «можна розділити на кластери предметів, кожен із яких є відносно однорідним за ознаками і які вимірно відрізняються один від одного». Прикладом тесту, який використовує майже просту структуру, є тест на

вступ до юридичної школи (LSAT). LSAT має три типи питань, включаючи логічне міркування (LR), аналітичне міркування (AR) і розуміння прочитаного (RC). Вчені проаналізували LSAT і виявили, що він є багатовимірним і що секція LR оцінює дві ознаки додаткової інформації (AI) і прихованих припущень (HA) з чіткою майже простою структурою.

Для тесту зі складною структурою домінуюча ознака не так чітко визначена, і обидві (або всі) здібності використовуються в основному для відповіді на питання. Вагомість кожної ознаки є більш тонкою і розсіяною, що ускладнює виявлення невідповідності даних моделі порівняно з майже простою структурою. Фінч в 2011 році [24] досліджував компенсаторну модель MIRT 3PL для майже простих і складних структур. Результати показали, що стандартна помилка збільшувалася зі складнішою структурою, з ненормальним розподілом прихованих ознак, з більшими кореляціями між ознаками та з меншими об'ємами вибірки. Перевірка структури в багатовимірному просторі здійснюється шляхом розрахунку кута, який можна обчислити за допомогою наступної формули:

$$\cos \alpha_{i\ell} = \frac{\alpha_{i\ell}}{\sqrt{\sum_{k=1}^{m} \alpha_{ik}^2}} \quad (1.8)$$

Структурна різниця - це максимальна різниця кутів домінуючих вимірів що мають найбільше значення. θ_1 є домінантною здібністю, коли α_1 має більше значення, тоді як θ_2 є домінантною здібністю, коли α_2 більше значення. Для тестів з майже простою структурою розмірності можуть бути визначені кутами 0° - 15° і 75° - 90°. Для тестів зі складною структурою розміри можуть бути визначені кутами 0° - 15° і 45° - 90°. Це робить опис здібностей більш розмитим. Наприклад, якщо α_1 = 0.7, значення α_2 може приймати діапазон значень, який визначає, чи тест має майже просту чи складну структуру. При меншому значенні α_2 = 0.2 структура майже проста. При більшому значенні α_2 = 0.5, θ_1 все ще є домінантною ознакою, але значення є таким, що визначає його як складну структуру.

Вплив структур був предметом багатьох досліджень. В своїй роботі Чанг і Стон [22] розглядали застосування статистики χ^2 [25] для оцінки відповідності моделі MIRT. Дослідження помилок I роду проводили з використанням майже простих і складних структур. Майже просту структуру визначили як таку, що має кути розмірності 1 від 0° - 20° і кути розмірності 2 від 70° - 90°, а складну - такою, що має кути розмірності 1 в межах 0° - 20° і кути розмірності 2 - 45° - 90°. Було виявлено, що при помилці I роду розподіли складної структури було складніше апроксимувати порівняно з приблизно простою структурою, тоді як результати досліджень потужності були подібними до результатів, отриманих за допомогою одновимірних моделей IRT.

Однією з проблем IRT і MIRT є невизначеність, коли параметри моделей можуть мати кілька значень, які дають однакову ймовірність. Вказуючи кути та обмежуючи діапазон, усувається невизначеність.

РОЗДІЛ 2

ОГЛЯД МЕТОДІВ ОЦІНЮВАННЯ ЯКОСТІ МОДЕЛЮВАННЯ

2.1 Оцінка моделі

Оцінка граничної максимальної правдоподібності (MMLE) є найпопулярнішим методом оцінки параметрів моделей IRT [26]. Суть полягає в тому, що здібність інтегрована в оцінки параметрів тесту, щоб забезпечити точні значення оцінок параметрів питання. MMLE складається з трьох основних кроків.

Першим кроком є вибір початкових значень параметрів питання, якщо відомі наближення. Чим ближче значення до реальних значень параметрів питання, тим швидше буде збіжність.

Другим кроком є обчислення правдоподібності, але для цього окремі θ з рівняння потрібно витягти та замінити розподілом. Параметри особи вважаються випадковими ефектами, і оскільки передбачається, що калібрувальна вибірка генерується випадковим чином, то при оцінці параметрів питання можна не враховувати оцінки параметрів іспитників. Для апроксимації неперервного розподілу використовується розбиття геометричної площини на прямокутники. Варто зазначити, що при зростанні кількості прямокутників покращується якість наближення [26]. Можна або розбити на прямокутники усю площину, обмежену кривою, або використовувати метод адаптивної квадратури.

Наступним кроком власне є інтегрування. При інтегруванні для кожного прямокутника використовується процедура зважування, тобто крім правдоподібності відповіді на питання також враховується ймовірність того, що при випадковій генерації буде обрано саме цей прямокутник. Потім зважені ймовірності підсумовуються по всіх прямокутниках (наприклад, інтегруються). Тоді правдоподібність максимізується за допомогою

ітераційного процесу. Якщо збіжність не досягнута, процес повторюється, використовуючи інше початкове значення. Якщо ж збіжність досягнута, значення параметрів дуже близькі до тих, які забезпечать максимальну правдоподібність і можна зробити оцінку здібностей.

Процедура MMLE подібна для одно- та багатовимірної оцінки елемента IRT. Різниця полягає в тому, що для оцінки IRT моделей використовується правдоподібність отримання балу за питання, а у MIRT моделей - правдоподібність отримання декількох балів за питання. Правдоподібність використовується для обчислення ймовірності того, що конкретний ряд балів за питання буде знайдений в групі іспитників. Суть методу максимальної правдоподібності однакова як для IRT, так і для MIRT, так як оцінка базується на понятті «точка максимуму на поверхні — це точка, де дотична площина до поверхні має нульовий нахил у всіх напрямках», або «Точка, в якій дотична площина торкається поверхні, є максимумом цієї поверхні» [12].

Оцінка граничної максимальної правдоподібності оцінює лише параметри моделі, тому оцінки здібностей мають бути отримані з використанням оцінки максимальної правдоподібності (MLE), байєсівських методів, апостеріорного максимуму (MAP) або очікуваного апостеріорного (EAP). Байєсівські методи (EAP і MAP) використовують отримані дані (параметри питання) для розробки апостеріорного розподілу, який потім відбирається за допомогою ітераційного процесу. Перевага байєсівських методів полягає в тому, що їх використовують для отримання апостеріорного розподілу на наш тета-параметр, включаючи максимальну та мінімальну оцінку. MLE є популярним підходом до оцінки, але він не може коректно її визначити, коли іспитник відповідає на всі питання правильно або неправильно. У порівнянні з MAP є дві основні причини того, що EAP є більш ефективним. По-перше, EAP використовує квадратурні точки (дискретний випадок) для оцінки, тоді як MAP використовує континуум (неперервний випадок). Очевидно, що останній варіант програє

в затраченому часі на розв'язок. По-друге, EAP використовує середнє значення для оцінки, тоді як MAP використовує моду.

Було виявлено, що використання середнього значення робить EAP менш вимогливим до обчислень під час ітераційного процесу [26].

2.2. Підбір моделі для даних

Визначення найкращої моделі серед наявних є важливим питанням, тому що при невідповідності моделі та даних можливі порушення у припущеннях моделі чи проблеми із процедурою оцінювання [27]. Насамперед, відсутність правильного підбору моделі для даних може негативно впливати на коректність використання оцінок [18, 19]. Існують різні процедури для оцінки різних типів відповідності моделі в IRT (наприклад, вибір питань, вибір осіб, загальний підбір моделі. [18]. У поточному дослідженні інтерес викликає загальна відповідність моделі і, зокрема, відповідність даних моделі для компенсаторних і некомпенсаторних моделей 2PL MIRT.

Для визначення найкращої моделі для даних необхідно використати відповідну статистику відповідності. Одним із вирішальних факторів є взаємодія між моделями (наприклад, чи є вони вкладеними чи невкладеними). З математичної точки зору, моделі є вкладеними тоді і тільки тоді, якщо вони мають лінійне відношення, та одна з них може бути приведена до іншої шляхом лінійних звужень. Тобто, одна модель є підмножиною іншої [28, 29]. Іншою необхідною вимогою до вкладеності є теоретична схожість. На противагу, моделі є невкладеними, якщо відсутнє лінійне відношення між ними, різними є застосовувані до них розподіли та функції зв'язку, наявна теоретична відмінність [30].

Прикладом вкладених моделей є 2PL IRT модель та IRT модель Раша. Для опису 2PL IRT моделі використовується наступна формула:

$$P\left(U_{ij}=1\middle|\,\theta_j,\alpha_i,b_i\right)=\frac{1}{1+\ e^{-1.7\alpha_i\left(\theta_j-b_i\right)}}\ (2.1)$$

Для IRT моделі Раша (де параметр a = 1 для усіх питань):

$$P\left(U_{ij}=1\middle|\,\theta_j,b_i\right)=\frac{1}{1+\ e^{-\left(\theta_j-b_i\right)}}\ (2.2)$$

Прикладом невкладених моделей є компенсаторні та некомпенсаторні 2PL MIRT моделі. Компенсаторна 2PL MIRT визначається наступним чином:

$$P\left(U_{ij}=1\middle|\,\overrightarrow{\theta_j},\overrightarrow{\alpha_i},b_i\right)=\frac{1}{1+\ e^{-1.7\overrightarrow{\alpha_i}\left(\overrightarrow{\theta_j}-b_i\right)}}\ (2.3)$$

А для некомпенсаторної 2PL MIRT моделі Раша :

$$P\left(U_{ij}=1\middle|\,\overrightarrow{\theta_j},\overrightarrow{\alpha_i},\overrightarrow{b_i}\right)=\prod_{\ell=1}^{m}\frac{1}{1+\ e^{-1.7\alpha_{i\ell}\left(\theta_{j\ell}-b_{i\ell}\right)}}\ (2.4)$$

Для цих моделей не існує лінійної функції, що може перетворити одну MIRT модель у іншу.

Як було показано раніше, традиційні методи підбору моделі до даних, що використовуються для вкладених моделей, не можуть бути застосовані до невкладених моделей, тому перелік можливих варіантів для аналізу підборів зводиться до порівняльних критеріїв. Серед порівняльних критеріїв найбільш відомими є баєсовий інформаційний критерій та інформаційний критерій Акаіке [31].

BIC, незважаючи на назву, насправді зовсім не інформаційний критерій, оскільки він не оцінює втрачену інформацію. BIC не вимагає того, що правильна модель, яка співвідноситься із реальністю, обов’язково має існувати. Критерій застосовується до підбору регресійних моделей що відповідають оцінкам максимальної правдоподібності, та може бути використаний як для вкладених, так і до невкладених моделей. Критерій BIC оцінює коефіцієнт Баєса, який є ймовірністю того, що дані отримані з однієї моделі, а не з іншої. Якщо ймовірність даних, отриманих з Моделі 1, більша за ймовірність даних, отриманих з Моделі 2, тоді Модель 1 є кращим

вибором. Треба наголосити, що це говорить не про те, що Модель 1 є найкращим вибором, а про те, що вона є кращою ніж Модель 2. Головною метою критерію є надання можливості для відносної оцінки двох моделей [32, 33]. Найкращі моделі мають найменше значення критерію.

Іншим популярним порівняльним критерієм є інформаційний критерій Акаіке [31]. AIC побудований на інформаційному критерії Кульбака-Лейблера [34], що визначається так:

$$KLIC = E_0[\ln h_0(Y_t \mid X_t)] - E_0[\ln f(Y_t \mid X_t; \beta_*)] \quad (2.5)$$

де $h_0(Y_t \mid X_t)$ – це умовне розподілення правильної моделі при заданому значенні X_t, β_*- оцінка значення β, коли $f(Y_t \mid X_t)$ не є правильною моделлю. Найкращі моделі мінімізують значення KLIC збільшуючи значення E_0. AIC мінімізує втрати інформації шляхом максимізації логарифмічної функції правдоподібності. Головною метою є відділення інформації від шуму чи помилок. Кожній розглядуваній моделі ставиться у відповідність значення критерію AIC. Ці значення можуть бути відсортовані, та з впорядкованого масиву можна обрати модель із кращим значенням. Найкраща модель має найменше значення критерію та відповідно найменші втрати інформації [34].

У роботі Кларка розглянуто дві проблеми, пов’язані з розглянутими та схожими на них критеріями. Перша проблема полягає у тому, шо ці критерії не дозволяють зробити ймовірнісне твердження стосовно вибору моделі, тому є різниця між абсолютною (відмінності між перевіркою гіпотези) та відносною (критерії вибору моделі) дискримінацію. Друга проблема полягає в тому, що серед порівнюваних моделей неодмінно буде зроблено вибір однієї з моделей, навіть якщо вони обидві є некоректними. Існує дві статистичних характеристики, що вирішують ці проблеми. Перша – це статистичний тест Вуонга, друга – статистичний тест Кларка.

Тест Вуонга [35] – це класичний метод тестування, може бути застосований для вкладених, невкладених моделей та моделей, що перетинаються. З часу свого створення у 1989 році, цей тест широко

застосовується в економіці, політології та інших наукових областях [36, 27, 37]. Тест Вуонга, окрім критерія AIC, також використовує KLIC, в якому найкраща модель має найменше значення критерію. Це досягається шляхом максимізації значення $E_0[\ln f\,(Y_t \mid X_t; \theta_*)]$, таким чином, щоб воно стало якомога ближчим до значення правильної моделі $E_0[\ln h_0\,(Y_t \mid X_t)]$. Тест Вуонга на для нульової гіпотези має вигляд:

$$H_0 : E_0\left[\ln\frac{f(Y_t \mid X_t; \beta_*)}{g(Y_t \mid Z_t; \gamma_*)}\right] = 0 \quad (2.6)$$

де E_0 – це очікування справжньої моделі, β_* - це оцінка значення β, коли модель $f(Y_t \mid X_t)$ не є правильною, γ_* - це оцінка значення γ, коли $g(Y_t \mid X_t)$ не є правильною моделлю. Тест визначає, чи є середнє відношення правдоподібності відмінним від нуля [38]. Якщо є вірною нульова гіпотеза, тоді коефіцієнт логарифмічних ймовірностей дорівнюватиме нулю. Якщо кращою є модель f, тоді значення буде більшим від нуля; якщо модель g, тоді значення буде меншим від нуля. Навіть якщо правильне значення або очікування є невідомим, Вуонг показав, що коли припущення виконується, то очікуване значення можна оцінити як $1/n$ помножене на відношення правдоподібності [28, 35, 39]. Очікуване значення оцінки можна обрахувати так:

$$\frac{1}{n} LR_n\big(\hat{\beta}_n, \hat{\gamma}_n\big) \xrightarrow{a.s.} E_0\left[\ln\frac{f(Y_t \mid X_t;\, \beta_*)}{g(Y_t \mid Z_t;\, \gamma_*)}\right] \quad (2.7)$$

Формулювання статистичного тесту Вуонга наведено у формулі 2.8:

$$H_0 : \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{\left(\sqrt{n}\right)\widehat{\omega}_n} \xrightarrow{D} N(0,1) \quad (2.8)$$

Формулювання параметру LR_n наведено у формулі (2.9), а значення дисперсії – формулі (2.10):

$$LR_n\big(\hat{\beta}_n, \hat{\gamma}_n\big) \equiv L_n^f\big(\hat{\beta}_n\big) - L_n^g(\hat{\gamma}_n) \quad (2.9)$$

$$\widehat{\omega}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{f(Y_t \mid X_t;\ \beta_n)}{g(Y_t \mid Z_t;\ \gamma_n)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(Y_t \mid X_t;\ \beta_n)}{g(Y_t \mid Z_t;\ \gamma_n)}\right]^2 \quad (2.10)$$

Статистичний тест Вуонга є класичним тестом вибору моделі, що відноситься до типу відносної дискримінації, де базовою є гіпотеза про те, що не існує суттєвої різниці між моделями. При процесі вибору моделі запропоновані варіанти порівнюються безпосередньо за допомогою певного критерію, та потім обирається найкраща модель. Неважливо, як ведуть себе моделі у абсолютних значеннях, важливо тільки яка модель краща за модель-суперника. Тест Вуонга не є досконалим: він все ще є тестом вибору моделі, та у результаті буде обрано модель, що ближча до ідеального рішення, але може бути дуже далекою від цього ідеалу. Проте Кларк довів, що цей тест є значно кращим за критерії AIC та BIC, тому що дозволяє висунути ймовірнісне твердження [38].

Статистичний тест Вуонга використовувався, зокрема для невкладених моделей, що стосувалися досліджень міжнародних відносин [28]. У дослідженні було порівняно дві теорії, структурного реалізму і раціонального стримування, які намагалися пояснити ескалацію мілітаризованих суперечок великих держав за допомогою двох наборів моделей. Перший набір моделей мав вискоку кореляцію, і тест Вуонга не зміг їх розрізнити. Другий набір моделей мав лише незначну кореляцію. За цієї умови тест Вуонга показав хороші результати та визначив модель структурного реалізму як кращу. В інших дослідженнях було виявлено, що тест Вуонга має властивість правильно визначити некоректний підбір моделі при менших розмірах вибірки ($N = 300$) [36], та ця здатність зменшується зі збільшенням кореляції [34].

Іншим статистичним тестом для порівняння невкладених моделей є тест Кларка [38]. Цей тест було спеціально розроблено як інструмент вибору невкладених моделей. Цей тест не використовує розподіли та бере до уваги розбіжності в окремих логарифмічних функціях правдоподібності

використовуючи тест зі змінним парним знаком [40]. Парний знаковий тест – це біноміальний тест, що використовується для тестування гіпотези, що ймовірність бути більшим чи меншим за нуль є однаковою для будь-якого випадкового значення із сукупності.

Логіка тесту Кларка полягає в тому, що якщо певна модель має кращі показники вибору, то і окремі логарифмічні функції правдоподібності цієї моделі також мають кращі показники. Нульова гіпотеза тесту Кларка обчислюється за формулою (2.11):

$$H_0 : Pr_0\left[\ln\frac{f(Y_t \mid X_t;\beta_*)}{g(Y_t \mid Z_t;\gamma_*)} > 0\right] = 0.5 \quad (2.11)$$

де E_0 – це очікування правильної моделі. Нульова гіпотеза стверджує, що окремі значення логарифмічної функції правдоподібності повинні бути рівномірно розподілені навколо нуля, причому половина має буде меншою, а інша половина - більшою за нуль.

У той час як тест Вуонга визначає, чи статистично відрізняється від нуля середнє логарифмічної функції правдоподібності, тест Кларка визначає, чи статистично відрізняється від нуля медіана логарифмічної функції правдоподібності [39]. У гіпотезі Pr_0 є медіаною відношення. Значення статистичного тесту Кларка визначається формулою (2.12):

$$B = \sum_{i=1}^{n} I_{(0;+\infty)}(d_i) \quad (2.12)$$

У цій формулі, I – це індикаторна функція, $d_i = \ln f\,(Y_t \mid X_t;\hat{\beta}_n) - \ln g\,(Y_t \mid X_t;\hat{\gamma}_n)$. B – це кількість додатних розбіжностей, та береться з біноміального розподілу з параметром $p = 0.5$. Загалом, модель з більшою кількістю розбіжностей має кращу дискримінацію і, отже, є кращою моделлю.

Статистичні тести Вуонга і Кларка мають певні значні переваги перед традиційними критеріями вибору невкладених моделей такими як AIC і BIC. По-перше, як було показано вище, вони дозволяють зробити

ймовірнісні твердження та знайти p-значення, які можна порівняти з критичними значеннями. По-друге, тести Вуонга і Кларка відносно легко обчислити. По-третє, тести Вуонга та Кларка мають додаткове співвідношення, оскільки вони розподілені по-різному (мають нормальний і біноміальний розподіл) та мають різну основу для порівняння оцінюваних моделей (середнє та медіану логарифмічної функції правдоподібності).

Оскільки тест Кларка була створено для усунення деяких недоліків тесту Вуонга (низький рівень виявлення невідповідності при відсутності нормального розподілу даних), тести виявляють некоректність вибору даних за деяких спільних умов і, що більш важливо, за різних умов. Додатково слід зазначити, що оскільки різні об'єми вибірки, довжини тестів та кореляції по-різному впливають на ефективність статистичних тестів, це також може вплинути на відмінності, за яких тести Вуонга та Кларка бути кращими при виявленні некоректності вибору моделі.

Ці відмінності було розглянуто у попередніх дослідженнях [40]. Зокрема, при нормальному розподілі основних значень логарифмічної правдоподібності, тест Кларка ефективний лише на 64% від тесту Вуонга, та тест Вуонга підходить більше для виявлення некоректності вибору.

2.3 Методи визначення розмірності моделей

Однією з проблем, що виникають при застосуванні MIRT-моделей, є визначення кількості координатних осей, необхідних для коректного відображення взаємозв'язків у матриці оцінок відповідей на питання. Це пояснюється тим, що матриця залежить не лише від чутливості тестових питань до розміщень різних іспитників у θ-просторі здібностей.

Фактором, що впливає на коректне визначення розмірності є варіативність вибірки іспитників, тобто один і той самий набір тестових питань може давати дещо різні результати в залежності від набору

іспитників. Зокрема, необхідно правильно знайти розмірність моделі в залежності від різних вибірок іспитників, якщо іспитники відрізняються за декількома категоріями здібностей, або тестові питання чутливі до диференціації тільки у кількох напрямках θ-простору.

Враховуючи складність причин взаємозв'язків у матриці оцінок, доцільно думати не про кількість вимірів, необхідних для моделювання тесту, а про кількість вимірів необхідних для моделювання наявних відношень у матриці оцінок, що виникають між конкретною вибіркою іспитників та конкретною вибіркою тестових питань. Тому, кількість вимірів є залежною від конкретної вибірки, а результати моделювання не можна узагальнити на відповіді іншої вибірки іспитників, що проходять той самий набір тестових питань.

Вирішивши пояснену вище способом проблему впливу варіативності вибірок на результати моделювання далі необхідно розглянути проблему чутливості тестових питань до заданих напрямів у багатовимірному θ-просторі.

Якщо вектори напрямів відповідей на тестові питання направлені у одному напрямку, то навіть якщо для отримання правильної відповіді на тестове питання потрібно застосувати декілька когнітивних здібностей, для моделювання матриці даних необхідним є лише один вимір. При цьому той факт, що наявні у вибірці іспитники відрізняються по ряду окремих когнітивних навиків, буде мати незначний вплив на кількість вимірів, необхідних для моделювання даних.

У випадку коли тестові питання є чутливими до диференціювання у різних напрямках простору (при чому найкращим випадком є чутливість до кількості напрямів, що дорівнює кількості питань), то є можна зменшити кількість вимірів моделі. Якщо іспитники не мають розбіжностей у певному напрямку, що є напрямом чутливості тестового питання, то напрям виміру для цього питання не вливає на загальну розмірність, і її можна зменшити на одиницю.

У певному сенсі немає універсального методу пошуку розмірності моделі, оскільки мають враховуватись особливості вибірок питань та іспитників. Альтернативними підходами для розв'язання цієї проблеми є інтерпретація на рівні окремих питань та агрегація оцінок за окремі питання у підшкали.

Другий підхід є більш надійним, тому розглянемо цілі, які необхідно досягти при застосуванні цього підходу. По-перше, необхідно визначити кількість кластерів питань, що є достатньо схожими для об'єднання в підшклалу. По-друге, необхідно визначити кількість ортогональних координатних осей, необхідних для коректного моделювання взаємозв'язків здібностей іспитників та їх відповідей на питання. Для того щоб показати різницю між вказаними цілями наведемо два приклади.

Приклад 1. Учителю, що працює з дітьми початкової школи, цікаво визначити, наскільки добре учні можуть розв'язувати прості арифметичні задачі, якщо вони представлені числовим виразом або у словесній формі. Учитель стурбований тим, що рівень навичок читання учнів може вплинути на результати. Було розроблено тест, що складається з трьох наборів тестових завдань. Перший набір складається з 20 простих арифметичних питань виду «4 + 17 = ?» Другий набір складається з 20 простих арифметичних завдань виду «Чому дорівнює шість плюс п'ять?» Третій набір тестових завдань складається з десяти простих питань для читання наступної форми:

Виберіть слово для заповнення пропуску в наступному реченні. Дитина на велосипеді рухалася дуже _____ .

(a) гучно

(b) швидко

(c) високо

(d) м'яко

Ці три кластери питань пов'язані наступним чином: знання арифметичних обчислень у числовому вигляді лише незначною мірою

пов’язані з навичками читання, тож можна вважати, що ці типи питань мають вектори, які майже ортогональні один одному в двовимірному θ-просторі, арифметичні завдання у словесній формі вимагають обох навичок одночасно. Результати моделювання представлені у двовимірному θ-просторі за допомогою графіка напряму оцінок (рис. 2.1).

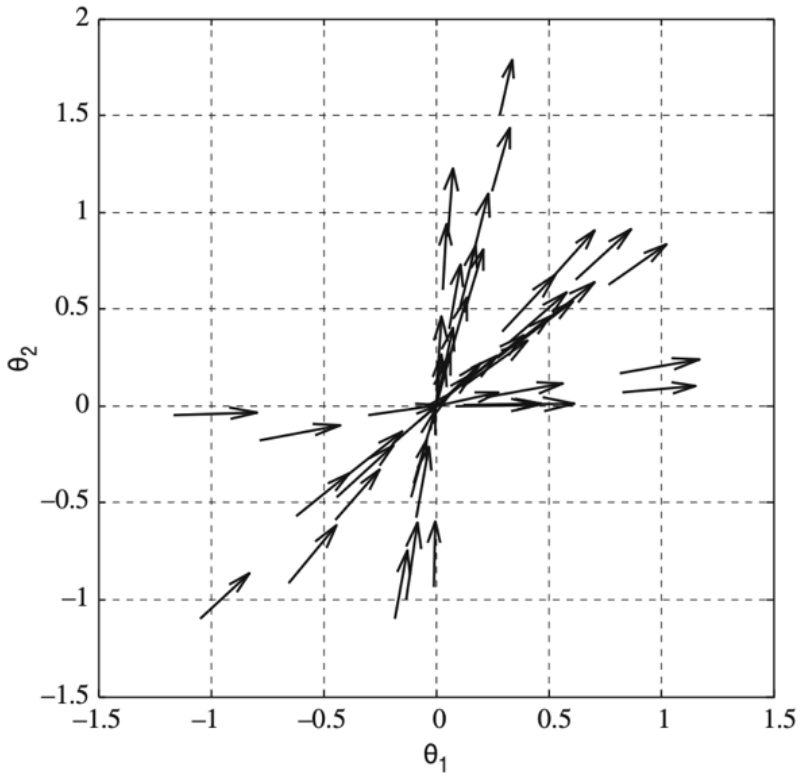


Рис. 2.1. Вектори напрямків відповідей для прикладу 1

Не складно побачити, що вектори 20 питань арифметичного блоку направлені вздовж осі θ_1, вектори 10 питань мовленнєвого блоку – вздовж осі θ_2, решта векторів питань змішаної форми мають приблизно однакові координати за двома осями. Важливо зазначити, що при моделюванні виявлено три кластери векторів, але для їх зображення вистачає двовимірного простору. Це пояснюється тим, що навички арифметики та читання є цілком незалежними одна від одної, а для вирішення завдань змішаної форми необхідно застосувати і навички арифметики, і навички читання одночасно.

Приклад 2. Учитель, який працює з дітьми шкільного віку трохи старших, ніж ті, що наведені в Прикладі 1, також зацікавлений в оцінці математичних навичок дітей. До повної вибірки включено 20 тестових питань з арифметики, 10 тестових питань на читання, та 20 завдань – на розв'язок задач, що передбачають складання певного алгоритму розв'язку до них. На рисунку 2.2 зображено результат моделювання. Навички читання вимірюються по осі θ_1, арифметичні – по θ_2, розв'язання задач – по осі θ_3. Слід зазначити, що для розв'язання задач необхідно також використати навички читання та обчислень. Вектори питань із перших двох наборів мають напрям, що є подібним до напряму осей, оскільки є незалежними від інших навичок. Вектори питань третього пакету не збігаються з жодним напрямом осей, оскільки відображають використання комбінації усіх трьох навичок.

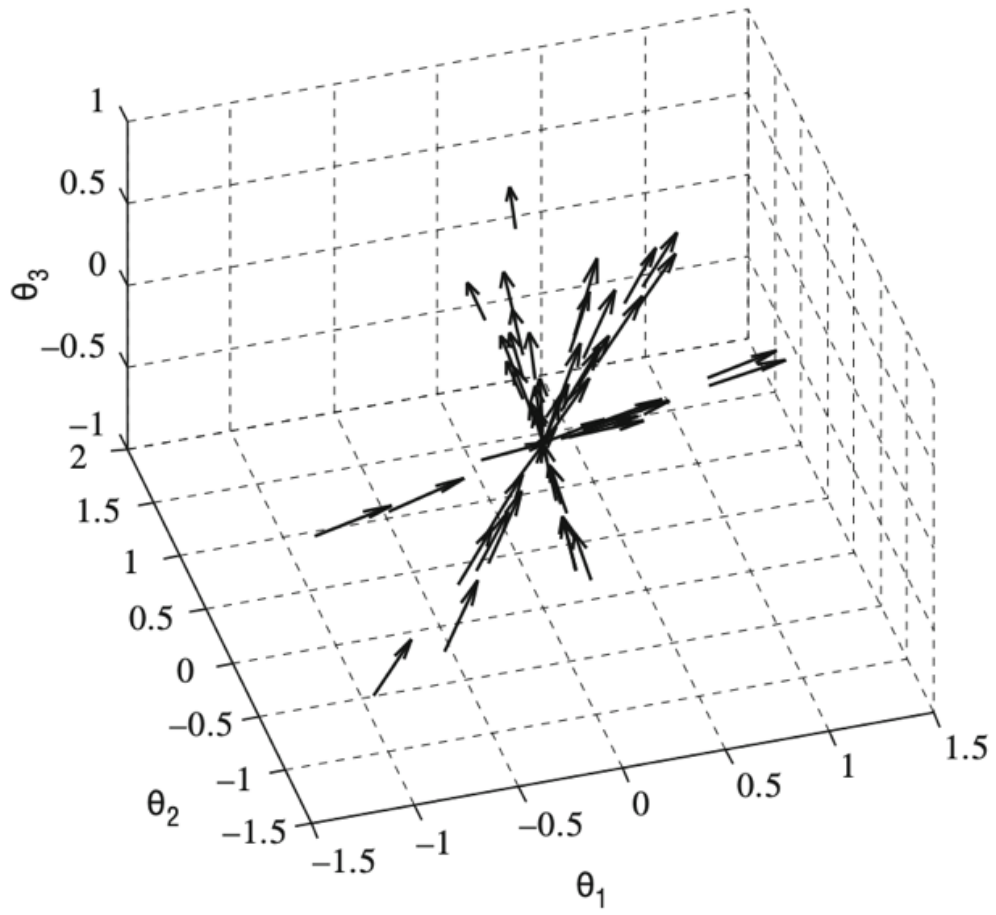


Рис. 2.2. Вектори напрямків відповідей для прикладу 2

Як бачимо, обидва приклади мають по три кластери питань. Але для другого прикладу при моделюванні необхідно використовувати тривимірний θ - простір, в той час як для першого прикладу достатньо лише двох вимірів.

Ці приклади демонструють різницю між кількістю кластерів та кількістю вимірів, необхідних для коректного відображення взаємозв'язків між тестовими питаннями.

Для знаходження необхідної кількості вимірів, що використовуються для моделювання матриці оцінок питань, застосовуються дві групи методів: параметричні та непараметричні.

Непараметричні методи побудовані на аналізі коваріацій між питаннями тесту. Ці процедури залежать від моделі, оскільки вони побудовані на метриці, що використовує лінійну залежність. Проте, використання математичної форми опису моделі, що використовується для аналізу даних відсутнє. Тому необхідна кількість вимірів визначається для конкретної моделі та вибірки питань і не є універсальною.

Одним із підходів до вирішення проблеми знаходження необхідної кількості вимірів є знаходження відповіді на питання, чи обов'язковим є використання більш ніж одного виміру. Якщо це твердження є вірним, тоді треба провести подальший аналіз для визначення необхідної кількості вимірів. Також є логічним, що необхідна кількість вимірів має залежати від моделі, що використовується для представлення даних. По цій причині непараметричні підходи використовуються тільки для початкової перевірки розмірності, а параметричні підходи дозволяють знайти остаточну кількість необхідних вимірів.

2.4. Паралельний аналіз

Одним із підходів по знаходженню розмірності моделі є Паралельний аналіз, що був запропонований Ледесмою та Валеро-Мора [41]. Цей метод складається з двох головних етапів.

По-перше, необхідно зробити аналіз розмірності за допомогою програми TESTFACT, що обчислює перші n власних значень матриці кореляцій між питаннями. Ця програма знаходить власні значення базуючись на тетрахоричних кореляціях. Кореляції можуть бути скоректовані за допомогою використання оцінки нижньої асимптоти для характеристичних поверхонь питань. Далі для побудови графіку кривих використовуються отримані власні значення в залежності від кількості знайдених вимірів.

По-друге, необхідно створити набір тестових даних, що не мають взаємозв'язку між питаннями, але мають такий же розмір вибірки і той самий розподіл правильних відповідей по кожному питанню, що і в реальних даних. Згенеровані дані потім аналізуються для отримання власних значень використанням аналізу, що був проведений для реальних даних. Оскільки згенеровані дані мають розподіл складності питань аналогічний реальним даним, результати аналізу згенерованих даних матимуть таку саму тенденцію до наявності значень складності, що і реальні дані. Власні значення згенерованих даних наносяться на той самий графік, що і відповідні власні значення реальних даних. Для визначення розмірності, запропонованої Паралельним аналізом, необхідно визначити кількість власних значень для реальних даних, яка має бути більшою за значення згенерованих даних.

Щоб отримати оцінку варіативності власних значень, обумовлених генерацією вибірки в залежності від її розміру та розподілення складності

реальних даних, необхідно декілька раз повторити генерацію імітаційних даних, що відповідають різним значенням складності реальних даних.

2.5 Критерій χ^2

Іншим популярним методом знаходження розмірності моделі є застосування критерію χ^2, який був вперше запропонований для вирішення поставленого завдання Шилінгом та Боком [42].

Метод полягає у знаходженні різниці якостей підбору моделі до даних для двох гіпотез, тобто, моделей, що мають розмірності m, та $m+1$ відповідно. Для цього необхідно знайти значення метрики якості підбору для кожної з гіпотез за допомогою програми TESTFACT, використовуючи матрицю відповідей на питання із зазначенням конкретної розмірності моделі (спочатку m, потім $m+1$). Далі різниця значень χ^2 для двох гіпотез із послідовним значенням кількості вимірів розраховується як різниця між ступенями свободи для двох значень, вираз для обчислення наведено у формулі 2.1:

$$\chi^2 = 2\sum_{\ell=1}^{s} r_\ell \ln \hat{P}_\ell - 2\sum_{\ell=1}^{s} r_\ell \ln \hat{P}'_\ell \quad (2.13)$$

У формулі використовуються наступні позначення: $\chi^2 -$ це значення різниці, $s -$ кількість наявних векторів відповідей в матриці оцінок, $r_\ell -$ це частота спостережень вектору відповідей ℓ у матриці, $\hat{P}_\ell -$це гранична ймовірність вектору відповідей ℓ при розмірності моделі m, $\hat{P}'_\ell -$це гранична ймовірність вектору відповідей ℓ при розмірності моделі $m+1$. Кількість ступенів свободи для критерію χ^2 визначається як $n-m$, де $n -$ це кількість питань у тесті.

Якщо критерій χ^2 є статистично суттєвим, це свідчить про значне покращення підбору моделі для даних при додаванні додаткового виміру в цю модель. Якщо значення критерію χ^2 не є суттєвим на заданому рівні,

додатковий вимір не призводить до значного покращення відповідності, тому для моделювання достатньо меншого значення розмірності.

2.6. Кластерний метод

Одним із підходів для визначення розмірності моделі, що є більш суб'єктивним за перелічені вище підходи, є Кластерний аналіз. Як було показано у розділі 2.3, кількість кластерів необов'язково вказує на необхідну кількість ортогональних координатних осей, необхідних для моделювання. Отримана кількість кластерів є лише верхньою межею для розмірності.

Базовим підходом знаходження розмірності моделі є порівняння результатів Кластерного аналізу виконаного для різної кількості вимірів. Якщо результати Кластерного аналізу для різної кількості вимірів мають схожі показники, то для моделювання достатньо використати менше значення кількості вимірів.

Для кластеризації питань необхідно розв'язати два завдання: знаходження ознаки, яка порівнює схожість питань , та побудова алгоритму формування кластерів. В якості ознаки застосовується кут між векторами питань, або умовна коваріація між питаннями.

Методом кластеризації, що використовує у якості ознаки кут між векторами питань та гарантує гарні показники, є метод Уорда, тому наведемо його опис. Кут між двома питаннями можна визначити наступним чином:

$$\alpha_{12} = \arccos(cos\alpha_1' cos\alpha_2) \qquad (2.14)$$

де α_1 і α_2 – це вектори двох питань. Використовуючи значення параметрів питань, формулу можна переписати так:

$$cos\alpha_{12} = \frac{\alpha_1'}{\sqrt{\sum_{\ell=1}^{m} \alpha_{1\ell}^2}} \frac{\alpha_2}{\sqrt{\sum_{\ell=1}^{m} \alpha_{2\ell}^2}} \qquad (2.15)$$

Якщо вектори питань будуть співнаправлені, то кут між ними дорівнюватиме нулю. Зі зростанням кута між векторами питань напрями їх максимальної дискримінації будуть розходитися у різні сторони θ-простору. Варто зазначити, що для питань когнітивних тестів кути між їх векторами рідко мають значення більше ніж 90 градусів. Таким чином, косинус кута між двома завданнями являє собою кореляцію між ними. Коли косинус кута дорівнює одиниці, тобто, вектори питань співнаправлені, питання мають ідеальну кореляцію. Вектори питань, що направлені ортогонально, вказують на нульову кореляцію між ними. Вхідними параметрами для алгоритму кластеризації є матриця кутів між всіма можливими парами тестових питань.

2.7. Критерій Кайзера-Гутмана

Методом знаходження розмірності моделі, що використовує обчислення власних значень і власних векторів коваріаційної матриці, є застосування критерію Кайзера-Гутмана [43].

Після обчислення перерахованих вище значень, серед них слід залишити лише ті, що мають власні значення більше 1 (критерій Кайзера-Гутмана). Це поясняється тим, що виокремлені значення мають більшу вагу для характеризування моделі ніж середнє значення загальної дисперсії.

Було досліджено, що правило Кайзера-Гутмана може залишати занадто багато параметрів, тому слід накладати додаткові обмеження. Зокрема Горсучем було рекомендовано виокремлювати кількість параметрів, що знаходиться у проміжку між $n = 6$ та $n = 10$, де n – це кількість параметрів. Також використання цього емпіричного правила є ще більш доречним при наявності менш ніж 40 параметрів та розміру вибірки більше за 100.

2.8. Складність та диференціююча здатність питань

MIRT-моделі дозволяють математично описати взаємодію іспитників і тестових питань. Незважаючи на те, що параметри MIRT-моделей містять приховані параметри питань, самих лише отриманих значень параметрів недостатньо для інтуїтивного розуміння результатів. Тому використовуються деякі інші статистичні методи оцінки значень тестових питань, які можуть більш чітко визначити розташування здібностей іспитників у багатовимірному $\theta -$ просторі. Наведені методи опису тестових питань є прямим узагальненням аналогічних характеристик для одновимірних IRT-моделей.

Оцінки складності та диференціюючої здатності одновимірних IRT-моделей безпосередньо пов'язані з параметрами характеристичної кривої питання. Параметр складності вказує значення θ, яке відповідає точці найбільш крутого нахилу характеристичної кривої. Параметр диференціюючої здатності вказує найбільше значення нахилу характеристичної кривої. Ці дві описові статистичні характеристики тестових питань можна узагальнити для MIRT-моделей, але є певні проблеми під час переходу до багатовимірного простору. Нахил поверхні залежить від напрямку руху вздовж поверхні, тому точка найбільш крутого нахилу залежить від напрямку, який розглядається.

Під багатовимірною складністю розуміється відстань від початку $\theta -$ простору до точки, яка знаходиться у місці найбільшого схилу поверхні. Знак, поруч із відстанню, вказує на відносне положення точки відносно початку координат. Значення багатовимірної складності можна обчислити за такою формулою:

$$B_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{m} \alpha_{ik}^2}} \quad (2.16)$$

У практичному трактуванні, великі додатні значення складності вказують на складні питання (тобто, для отримання ймовірності правильної відповіді більше ніж 0.5 необхідні високі значення розміщення іспитника у $\theta -$ просторі). Низькі значення складності вказують на питання із високою ймовірністю правильної відповіді для звичайних рівнів здібностей іспитників.

Під багатовимірною диференціюючою здатністю розуміють нахил характеристичної поверхні у точці найбільшої крутизни схилу у напрямі від центру $\theta -$ простору. Формула для обчислення:

$$A_i = \sqrt{\sum_{k=1}^{m} \alpha_{ik}^2} \quad (2.17)$$

РОЗДІЛ 3

МАТЕМАТИЧНИЙ ОПИС МОДЕЛЕЙ

3.1. Компенсаторна 2-PL модель

Нехай дано: $p_{ij} = \frac{e^{\alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2}+d_j}}{1+e^{\alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2}+d_j}}$, де $i = \overline{1,n}$; $j = \overline{1,m}$; n – це кількість іспитників; m – це кількість питань в тесті; α_{j1}, α_{j2} – диференціюючі здатності ; j-го завдання; θ_{i1}, θ_{i2} – рівні підготовленості i-го іспитника; d_j – рівень складності j-го завдання.

Введемо позначення: $\psi_{ij} = \alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2} + d_j$, тоді $p_{ij} = \frac{e^{\psi_{ij}}}{1+e^{\psi_{ij}}}$. Позначимо також через $q_{ij} = 1 - p_{ij} = \frac{1}{1+e^{\psi_{ij}}}$.

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} p_{ij}^{y_{ij}} \ (1-p_{ij})^{(1-y_{ij})}, y_{ij} \in \{0;1\}.$$

$$\mathcal{L} = lnL = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\ln p_{ij} + (1-y_{ij})\ln q_{ij.}$$

$$\mathcal{L} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[y_{ij}\left(\psi_{ij} - \ln\left(1+e^{\psi_{ij}}\right)\right) - \left(1-y_{ij}\right)\ln\left(1+e^{\psi_{ij}}\right)\right] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\psi_{ij} - \ln\left(1+e^{\psi_{ij}}\right)\right).$$

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\alpha_{is}} = \displaystyle\sum_{i=1}^{n}\left(y_{ij}\theta_{is} - \frac{e^{\psi_{ij}}}{1+e^{\psi_{ij}}}\theta_{is}\right) = \sum_{i=1}^{n}\left(y_{ij}-p_{ij}\right)\theta_{is} = 0; s = 1;2, \\ \dfrac{\partial\mathcal{L}}{\partial\theta_{is}} = \displaystyle\sum_{j=1}^{m}\left(y_{ij}-p_{ij}\right)\alpha_{js} = 0; s = 1;2, \\ \dfrac{\partial\mathcal{L}}{\partial d_j} = \displaystyle\sum_{i=1}^{n}\left(y_{ij}-p_{ij}\right) = 0\,. \end{cases}$$

3.2. Компенсаторна 3-PL модель

$$\psi_{ij} = \alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2} + d_j,$$

$$p_{ij} = c_j + (1 - c_j)\frac{e^{\psi_{ij}}}{1+e^{\psi_{ij}}} = \frac{c_j+e^{\psi_{ij}}}{1+e^{\psi_{ij}}};\ 1 - p_{ij} = \frac{1-c_j}{1+e^{\psi_{ij}}}.$$

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} p_{ij}^{y_{ij}}\ (1 - p_{ij})^{(1-y_{ij})}, y_{ij} \in \{0; 1\}.$$

$$\mathcal{L} = \ln L = \sum_{i=1}^{n}\sum_{j=1}^{m}[\ y_{ij}\ln(c_j + e^{\psi_{ij}}) - y_{ij}\ln(1 + e^{\psi_{ij}}) +$$

$$+(1 - y_{ij})\ln(1 - c_j) - \ln(1 + e^{\psi_{ij}}) + y_{ij}\ln(1 + e^{\psi_{ij}})\ ] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\ln(c_j + e^{\psi_{ij}}) + (1 - y_{ij})\ln(1 - c_j) - \ln(1 + e^{\psi_{ij}}).$$

Запишемо нашу систему:

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial\theta_{is}} = \sum_{j=1}^{m}\left(\frac{y_{ij}e^{\psi_{ij}}\alpha_{js}}{c_j+e^{\psi_{ij}}} - \frac{e^{\psi_{ij}}\alpha_{js}}{1+e^{\psi_{ij}}}\right) = 0,\ s \in \{1; 2\}, i = \overline{1, n} \\ \frac{\partial\mathcal{L}}{\partial\alpha_{js}} = \sum_{i=1}^{n}\left(\frac{y_{ij}e^{\psi_{ij}}\theta_{is}}{c_j+e^{\psi_{ij}}} - \frac{e^{\psi_{ij}}\theta_{is}}{1+e^{\psi_{ij}}}\right) = 0,\ j = \overline{1, m}, \\ \frac{\partial\mathcal{L}}{\partial d_j} = \sum_{i=1}^{n}\left(\frac{y_{ij}e^{\psi_{ij}}}{c_j+e^{\psi_{ij}}} - \frac{e^{\psi_{ij}}}{1+e^{\psi_{ij}}}\right) = 0\,, \\ \frac{\partial\mathcal{L}}{\partial c_j} = \sum_{i=1}^{n}\left(\frac{y_{ij}}{c_j+e^{\psi_{ij}}} - \frac{1-y_{ij}}{1-c_j}\right) = 0\,. \end{cases},$$

Позначимо $f_{ij} = \left(\frac{y_{ij}}{c_j+e^{\psi_{ij}}} - \frac{1}{1+e^{\psi_{ij}}}\right)e^{\psi_{ij}}$, тоді систему можна переписати в наступному вигляді:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_{is}} = \sum_{j=1}^{m} f_{ij}\alpha_{js} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_{js}} = \sum_{i=1}^{n} f_{ij}\theta_{is} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial d_j} = \sum_{i=1}^{n} f_{ij} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial c_j} = \sum_{i=1}^{n}\left(\dfrac{y_{ij}}{c_j + e^{\psi_{ij}}} - \dfrac{1-y_{ij}}{1-c_j}\right) = 0. \end{cases}$$

3.3. Некомпенсаторна модель PC2PL

Розглянемо PC2PL, $c_i = 0,\ \ i = \overline{1,\ m}$

$$L = \prod_{i=1}^{m}\prod_{j=1}^{n} p_{ij}^{y_{ij}}\ \left(1-p_{ij}\right)^{\left(1-y_{ij}\right)};$$

$$p_{ij} = \prod_{k=1}^{2} \frac{e^{\alpha_{ik}(\theta_{jk}-b_{ik})}}{1+e^{\alpha_{ik}(\theta_{jk}-b_{ik})}}.$$

$$\mathcal{L} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\, y_{ij}\ln p_{ij} + (1-y_{ij})\ln\left(1-p_{ij}\right)\right] =$$

$$= \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\, y_{ij}\ln p_{ij} + (1-y_{ij})\ln q_{ij}\right].$$

$$\varphi_{ijs} = \alpha_{is}\left(\theta_{js} - b_{is}\right).$$

$$\mathcal{L} = \sum_{i=1}^{m}\sum_{j=1}^{n}\Big[\, y_{ij}\left(\varphi_{ij1} + \varphi_{ij2} - \ln(1+e^{\varphi_{ij1}}) - \ln(1+e^{\varphi_{ij2}})\right)$$

$$+ \left(1-y_{ij}\right)(\ln(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}) - \ln(1+e^{\varphi_{ij1}})$$

$$- \ln(1+e^{\varphi_{ij2}}))\Big] =$$

$$\mathcal{L} = \sum_{i=1}^{m}\sum_{j=1}^{n}[y_{ij}\left(\varphi_{ij1} + \varphi_{ij2}\right) + (1-y_{ij})\ln(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}) -$$

$$- \ln(1+e^{\varphi_{ij1}}) - \ln(1+e^{\varphi_{ij2}})].$$

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_{js}} = \sum\limits_{i=1}^{m} \left[y_{ij}\alpha_{is} + \dfrac{(1-y_{ij})e^{\varphi_{ijs}}a_{is}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} - \dfrac{e^{\varphi_{ijs}}a_{is}}{1+e^{\varphi_{ijs}}} \right] = 0, \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_{is}} = \sum\limits_{j=1}^{n} \left[y_{ij}\theta_{js} + \dfrac{(1-y_{ij})e^{\varphi_{ijs}}\theta_{js}y_{ij}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} - \dfrac{e^{\varphi_{ijs}}\theta_{js}}{1+e^{\varphi_{ijs}}} \right] = 0, \\ \dfrac{\partial \mathcal{L}}{\partial b_{is}} = \sum\limits_{j=1}^{n} \left[-\alpha_{is}y_{ij} - \dfrac{(1-y_{ij})\alpha_{is}e^{\varphi_{ijs}}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} + \dfrac{\alpha_{is}e^{\varphi_{ijs}}}{1+e^{\varphi_{ijs}}} \right] = 0. \end{cases} \quad (*)$$

Позначимо через $g_{ij} = y_{ij} + \frac{(1-y_{ij})e^{\varphi_{ijs}}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} - \frac{e^{\varphi_{ijs}}}{1+e^{\varphi_{ijs}}}$. Тоді $(*) \Rightarrow$

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_{js}} = \sum\limits_{i=1}^{m} g_{ij}\alpha_{is} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_{is}} = \sum\limits_{j=1}^{n} \theta_{js}g_{ij} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial b_{is}} = \sum\limits_{j=1}^{n} -\alpha_{is}g_{ij} = 0. \end{cases}$$

3.4. Некомпенсаторна модель Сімпсона PC3PL

$$p_{ij} = c_i + (1-c_i)\frac{e^{\varphi_{ij1}+\varphi_{ij2}}}{(1+e^{\varphi_{ij1}})(1+e^{\varphi_{ij2}})} =$$

$$= \frac{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + c_i e^{\varphi_{ij1}+\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}} - c_i e^{\varphi_{ij1}+\varphi_{ij2}}}{(1+e^{\varphi_{ij1}})(1+e^{\varphi_{ij2}})} =$$

$$= \frac{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}}{(1+e^{\varphi_{ij1}})(1+e^{\varphi_{ij2}})}$$

$$q_{ij} = 1 - p_{ij} = (1-c_i) - \frac{(1-c_i)e^{\varphi_{ij1}+\varphi_{ij2}}}{(1+e^{\varphi_{ij1}})(1+e^{\varphi_{ij2}})}$$

$$= \frac{(1-c_i)(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}})}{(1+e^{\varphi_{ij1}})(1+e^{\varphi_{ij2}})}$$

$$\mathcal{L} = \sum_{j=1}^{m}\sum_{i=1}^{n} [\, y_{ij} \ln(c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}) -$$

$$- y_{ij}\ln(1+e^{\varphi_{ij1}}) - y_{ij}\ln(1+e^{\varphi_{ij2}}) + (1-y_{ij})\ln(1-c_i) +$$
$$+(1-y_{ij})\ln(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}})$$
$$- (1-y_{ij})\ln(1+e^{\varphi_{ij1}}) - (1-y_{ij})\ln(1+e^{\varphi_{ij2}})] =$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n}[\, y_{ij}\ln(c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}) + (1-y_{ij})\ln(1-c_i) +$$
$$+(1-y_{ij})\ln(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}) - \ln(1+e^{\varphi_{ij1}}) - \ln(1+e^{\varphi_{ij2}})].$$

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_{js}} = \sum\limits_{i=1}^{m}\left[\dfrac{y_{ij}(c_i e^{\varphi_{ijs}}\alpha_{is} + e^{\varphi_{ij1}+\varphi_{ij2}}\alpha_{is})}{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}} + \dfrac{(1-y_{ij})e^{\varphi_{ijs}}\alpha_{is}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} - \dfrac{e^{\varphi_{ijs}}\alpha_{is}}{1+e^{\varphi_{ijs}}}\right] = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_{is}} = \sum\limits_{j=1}^{n}\left[\dfrac{y_{ij}(c_i e^{\varphi_{ijs}}\theta_{js} + e^{\varphi_{ij1}+\varphi_{ij2}}\theta_{js})}{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}} + \dfrac{(1-y_{ij})e^{\varphi_{ijs}}\theta_{js}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} - \dfrac{e^{\varphi_{ijs}}\theta_{js}}{1+e^{\varphi_{ijs}}}\right] = 0 \\ \dfrac{\partial \mathcal{L}}{\partial b_i} = \sum\limits_{j=1}^{n}\left[-\dfrac{y_{ij}(c_i e^{\varphi_{ijs}}\alpha_{is} + e^{\varphi_{ij1}+\varphi_{ij2}}\alpha_{is})}{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}} - \dfrac{(1-y_{ij})e^{\varphi_{ijs}}\alpha_{is}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} + \dfrac{e^{\varphi_{ijs}}\alpha_{is}}{1+e^{\varphi_{ijs}}}\right] = 0 \\ \dfrac{\partial Z}{\partial c_i} = \sum\limits_{j=1}^{n}\left[\dfrac{y_{ij}(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}})}{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}} - \dfrac{(1-y_{ij})}{1-c_i}\right] = 0 \end{cases} \quad (**)$$

Позначимо:

$$\beta_{ij} = \frac{y_{ij}(c_i e^{\varphi_{ijs}} + e^{\varphi_{ij1}+\varphi_{ij2}})}{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}} + \frac{(1-y_{ij})e^{\varphi_{ijs}}}{1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}}} - \frac{e^{\varphi_{ijs}}}{1+e^{\varphi_{ijs}}}.$$

Тоді (**) ⇒

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_{js}} = \sum\limits_{i=1}^{m}\beta_{ij}\alpha_{is} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial b_i} = \sum\limits_{j=1}^{n} -\beta_{ij}\alpha_{is} = 0, \\ \dfrac{\partial \mathcal{Z}}{\partial \alpha_{is}} = \sum\limits_{j=1}^{n}\beta_{ij}\theta_{js} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial c_i} = \sum\limits_{j=1}^{n}\left[\dfrac{y_{ij}(1+e^{\varphi_{ij1}}+e^{\varphi_{ij2}})}{c_i + c_i e^{\varphi_{ij1}} + c_i e^{\varphi_{ij2}} + e^{\varphi_{ij1}+\varphi_{ij2}}} - \dfrac{(1-y_{ij})}{1-c_i}\right] = 0. \end{cases}$$

РОЗДІЛ 4

ОГЛЯД РЕЗУЛЬТАТІВ МОДЕЛЮВАННЯ ТА ЇХ АНАЛІЗ

При моделюванні було використано дані з модульної контрольної роботи (МКР) з предмету «Теорія ймовірностей». Попередньо було видалено зайві дані, матрицю відповідей було приведено до дихотомічного вигляду. Частину матриці наведено на рисунок 4.1, де стовбці відображають відповіді по питанню, а рядки є вектором оцінок певного іспитника.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1
1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	0	1	1	1	0	0	0	0	1	0	1	1	0	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1
1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1	0	0	0	0
1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0
1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	0	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1

Рисунок 4.1. Матриця відповідей іспитників на питання

Для моделювання було використано компенсаторну 2PL-модель та некомпенсаторну 2PL-модель. Лістинг коду наведений на рисунку 4.2.

Математичний опис моделей було розглянуто у розділі 3. Результати моделювання розглянуто далі у цьому розділі.

```
library(xlsx)
mydata<-read.xlsx("D:/work/tvd.xlsx",header
=F,rowIndex = c(2:61),colIndex = c(5:32),sheetIndex = 1)
colnames(mydata)<-c(1:28)
library(psych)
describe(mydata[,-1])
library(psychometric)
alph <- psychometric::alpha(mydata[,-1])
alph
library(mirt)
data_del<-mydata[,-c(19,20)]
mod_2pl<-mirt(data_del,2,"2PL")
for (i in 1:26){
print(itemplot(mod_2pl,type='tracecontour',i))
print(itemplot(mod_2pl,type='score',i))
print(itemplot(mod_2pl,i))}
residuals(mod_2pl)
summary(mod_2pl)
coef(mod_2pl)
model<- mirt.model('F1 = 1-26
              F2=1-26')
partcomp <- mirt(data_del,model,"PC2PL")
write.xlsx(coef(partcomp),"D:/work/partcomp.xlsx")
for (i in 1:26){
itemplot(partcomp,type='tracecontour',i)
itemplot(partcomp,type='score',i)
itemplot(partcomp,i)}
residuals(partcomp)
MDIFF(partcomp)
anova(mod_2pl,partcomp)
```

Рисунок 4.2. Лістинг коду

4.1. Результати моделювання компенсаторної MIRT-моделі

Під час моделювання вхідних даних, представлених на рис 4.1, за допомогою двовимірної компенсаторної MIRT-моделі, було отримано наступні результати. Значення логарифмічної правдоподібності моделі дорівнює -460.174.

Аналітично отримані значення параметрів диференціюючої здатності a_1, a_2 та складності питань d наведено у таблиці 4.1 для усіх питань.

Таблиця 4.1. *Результати моделювання параметрів питань за допомогою компенсаторної MIRT-моделі*

N питання	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{d}$
1	-3.34	-4.455	8.78
2	-44.567	-66.688	102.375
3	-36.574	-65.432	86.167
4	-3.445	-2.749	9.003
5	-3.518	-1.039	5.913
6	-1.206	-0.541	2.2
7	-1.037	-0.391	1.807
8	-0.796	-0.222	1.76
9	-1.544	0.07	2.06
10	-3.982	0.413	3.844
11	-29.672	0.147	26.03
12	-4.161	-0.856	4.221
13	-5.466	-1.465	5.586
14	-1.449	-1.041	2.663

15	-1.419	-2.131	3.722
16	-1.439	0.627	3.254
17	-1.37	-0.625	1.679
18	-1.665	-0.698	1.39
21	-130.761	-52.126	120.249
22	-147.523	-54.837	129.587
23	-0.924	-0.365	1.031
24	-1.683	-0.992	2.642
25	-74.83	18.695	109.6
26	-3.16	1.349	4.702
27	-47.211	24.169	60.824
28	-2.503	0	3.547

У таблиці відсутні рядки для питань 19 та 20, тому що їх треба було вилучити із вхідних даних для моделі через наявність проблеми зі збіжністю.

Для квадратурних методів це пояснюється тим, що вичерпно-інформаційний факторний аналіз питань без накладених обмежень має проблеми зі збіжністю, і деякі питання доводиться обмежувати або повністю видаляти, щоб забезпечити збіжність. За загальним правилом дихотомічні питання із середніми значеннями більше ніж 0,95 або питання, які лише на 0,05 перевищують параметр вгадування, слід видалити з аналізу або зробити попередню обробку розподілу параметрів. Такі ж міркування застосовуються і при включенні параметрів верхньої межі. В політомічних питаннях це спричинить аналогічні проблеми. Зазначимо, що використання методу квазі-Монте-Карло може допомогти стабілізувати процес оцінки у більших вимірах. Рішення, які не чітко визначені, також будуть мати

труднощі зі збіжністю і можуть вказувати на те, що модель була неправильно визначена.

Для алгоритму MH-RM, коли кількість ітерацій стає дуже високою (наприклад, більше 1500) або коли Max Change = 0.25, то значення друкуються на консолі занадто часто (вказуючи на те, що параметри були обмежені, оскільки вони рухались із кроком, більшим за 0,25. В такому випадку модель може бути погано визначеною або мати дуже плоску інформаційну поверхню, а справжні оцінки параметрів максимальної правдоподібності може бути важко знайти. Стандартні помилки обчислюються після збіжності моделі шляхом передачі SE = TRUE, щоб виконати додатковий етап MH-RM, але розглядаючи оцінки максимальної правдоподібності як фіксовані точки.

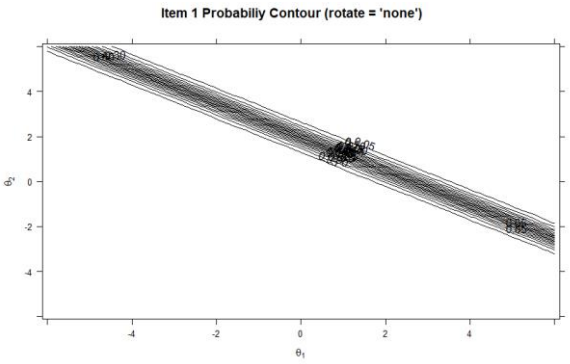
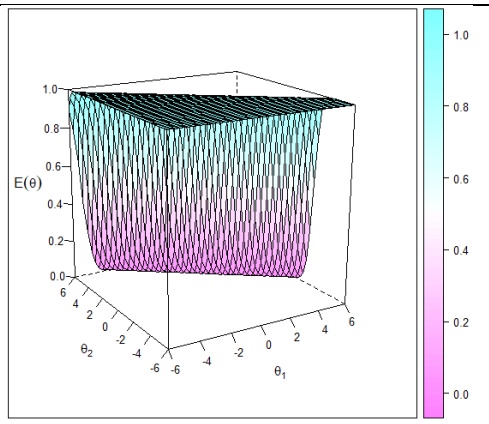
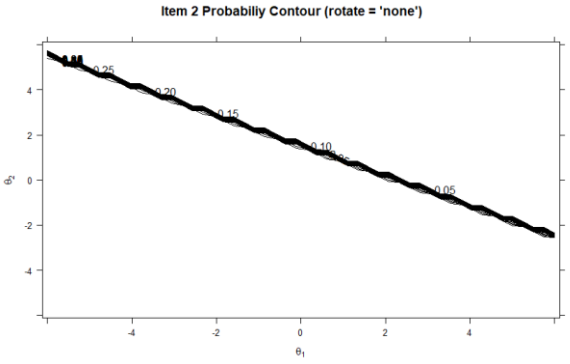
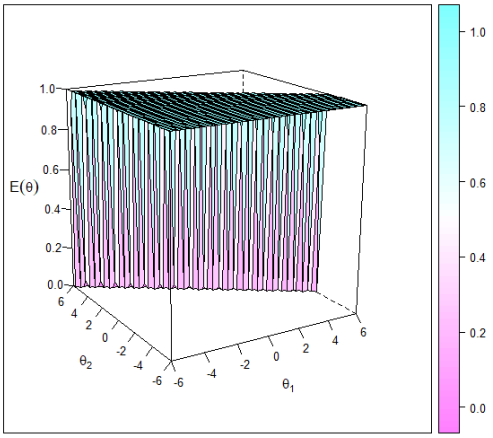
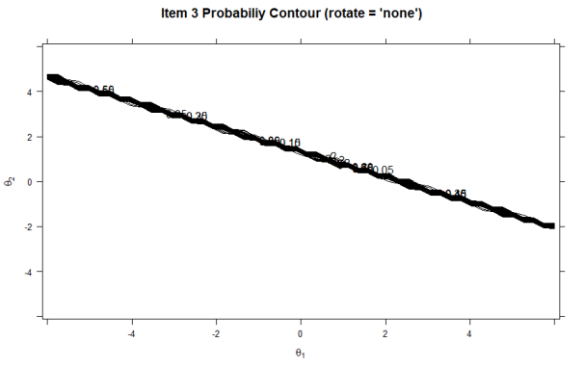
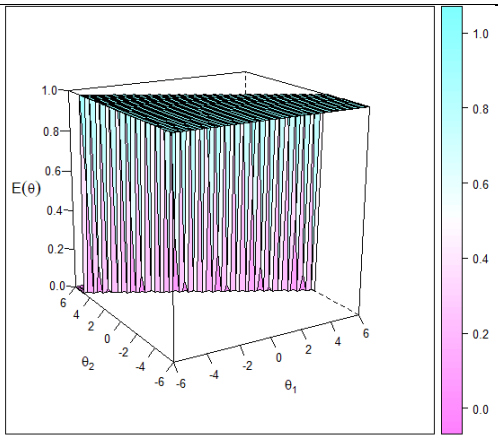
Аналізуючи результати моделювання отримані аналітично, можна побачити, що завдання 2, 3, 21, 22 мають високе значення параметру диференціюючої здатності за двома здібностями, 11, 25, 27 – за першою, але 25 та 27 мають велике значення диференціюючої здатності за другою здібністю. Питання 2, 3, 21, 22, 27 є дуже складними, 11 має помірну складність.

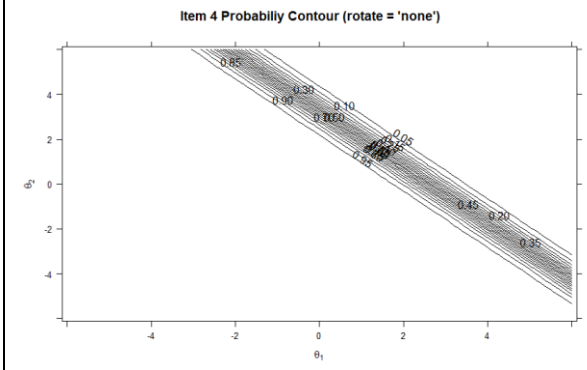
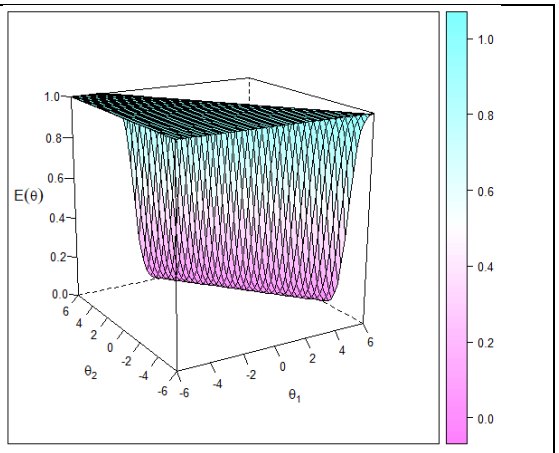
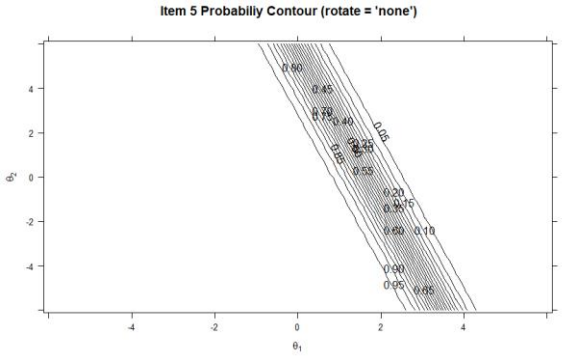
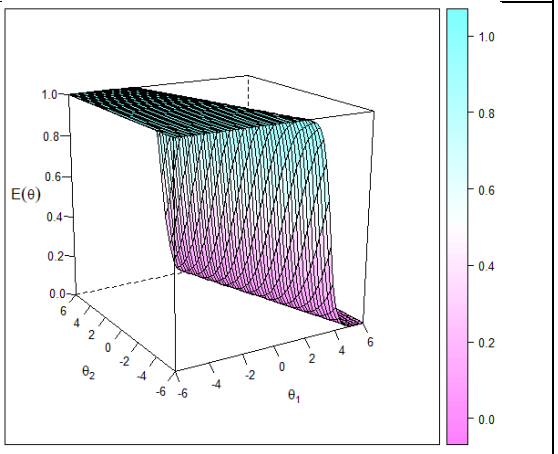
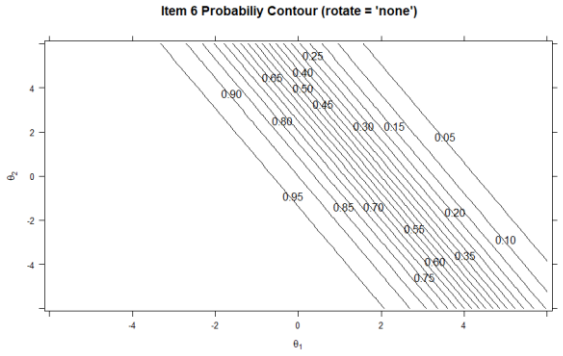
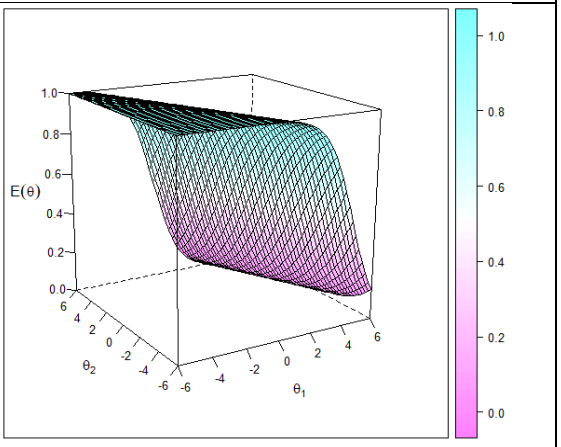
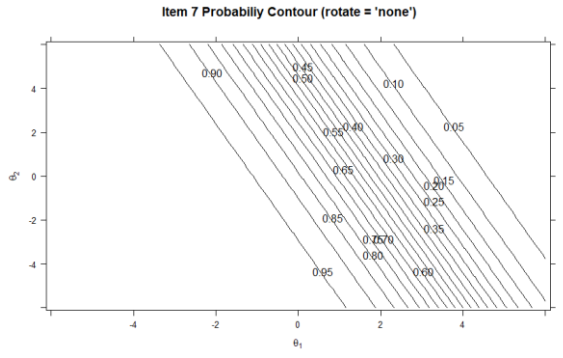
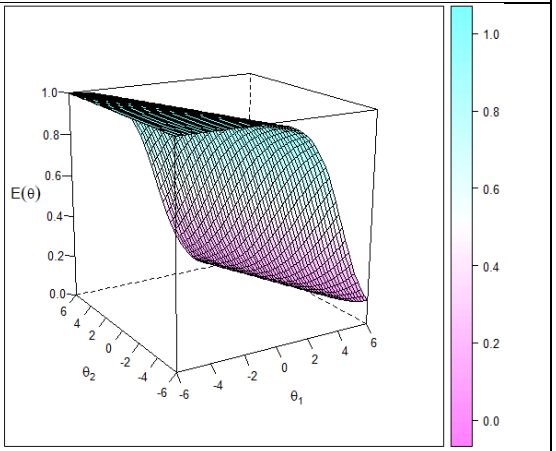
Геометричне представлення результатів, а саме контурні лінії питань та поверхні очікуваних оцінок іспитників, наведено у таблиці 4.2 для питань 1-8 та 11. Можна порівняти отримані результати графічно з аналітичними.

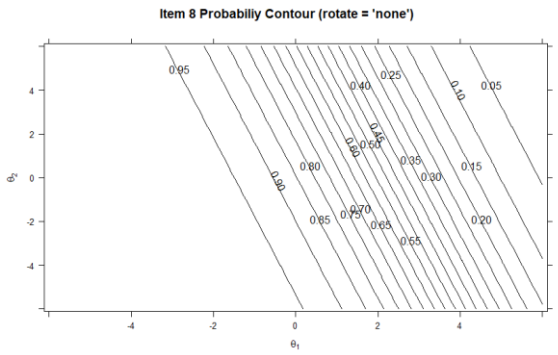
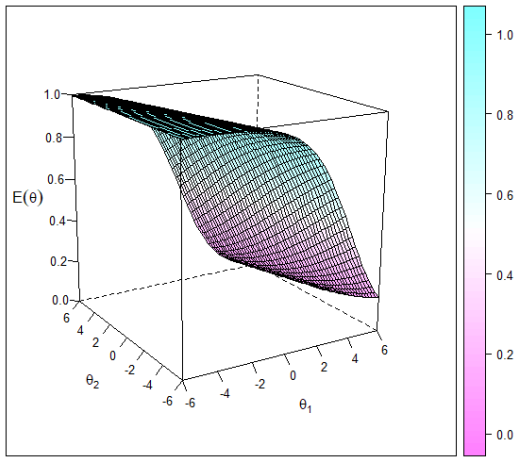
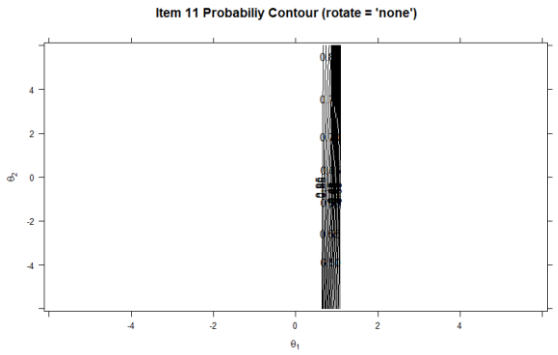
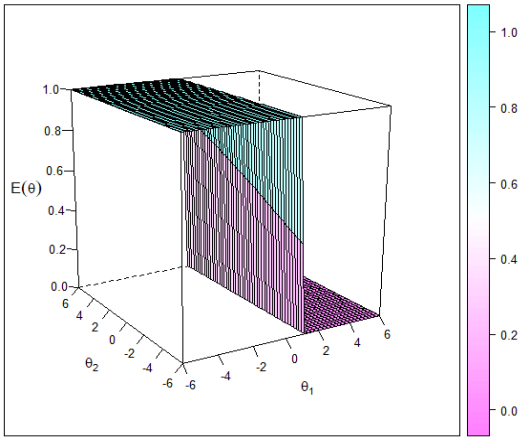
Аналізуючи результати моделювання отримані графічно, дивлячись на щільність контурних ліній можна побачити, що завдання 2, 3, 11 мають високу диференціюючу здатність, питання 6-8 – маленьку. Цей факт співпадає з висновками щодо цих питань, отриманих аналітично.

Аналізуючи контурні лінії питань, можна побачити що для перших 8 питань маємо від’ємний кут нахилу прямих. Це свідчить про те, що питання 1-8 оцінюють обидві здібності іспитника, а низький рівень володіння навичкою 1 не можна компенсувати високим рівнем навички 2. Контурні

Таблиця 4.2. *Результати моделювання контурних ліній та поверхонь очікуваних оцінок іспитників за допомогою компенсаторної MIRT-моделі*

№ пит	Контурні лінії питання	Поверхня очікуваних оцінок іспитників
1		
2		
3		

4	 Item 4 Probabiliy Contour (rotate = 'none') θ1 θ2	 E(θ) θ1 θ2
5	 Item 5 Probabiliy Contour (rotate = 'none') θ1 θ2	 E(θ) θ1 θ2
6	 Item 6 Probabiliy Contour (rotate = 'none') θ1 θ2	 E(θ) θ1 θ2
7	 Item 7 Probabiliy Contour (rotate = 'none') θ1 θ2	 E(θ) θ1 θ2

8		
11		

лінії питання 11 направлені вертикально – питання характеризує рівень володіння навичкою 1. Питання 21, 22 мають низьке значення параметру диференціюючої здатності за двома здібностями, 11, 25, 27 – за першою, але 25 та 27 мають велике значення диференціюючої здатності за другою здібністю. Питання 2, 3, 21, 22, 27 є дуже складними, 11 має помірну складність.

4.2. Результати моделювання частково компенсаторної MIRT-моделі

Під час моделювання вхідних даних, представлених на рис. 4.1, за допомогою двовимірної частково компенсаторної MIRT-моделі, було отримано наступні результати.

Аналітично отримані значення параметрів диференціюючої здатності a_1, a_2 та складності питань d наведено у таблиці 4.3 для усіх питань.

Таблиця 4.3. *Результати моделювання параметрів питань за допомогою частково компенсаторної MIRT-моделі*

N питання	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{d_1}$	$\boldsymbol{d_2}$
1	2.334	2.334	3.428	3.428
2	2.262	2.262	2.709	2.709
3	1.808	1.808	2.154	2.154
4	3.016	3.016	5.251	5.251
5	3.124	3.124	3.47	3.47
6	1.267	1.267	1.864	1.864
7	1.222	1.222	1.586	1.586
8	1.001	1.001	1.745	1.745
9	1.521	1.521	1.435	1.435
10	3.624	3.624	0.986	0.986
11	5.825	5.825	1.788	1.788
12	4.468	4.468	1.389	1.389
13	5.302	5.302	1.754	1.754
14	1.439	1.439	1.985	1.985
15	1.267	1.267	2.294	2.294
16	1	1	2.452	2.452
17	1.514	1.514	1.26	1.26
18	1.694	1.694	0.723	0.723
21	4.485	4.485	1.699	1.699

22	4.543	4.543	1.4	1.4
23	1.127	1.127	0.959	0.959
24	1.954	1.954	1.814	1.814
25	3.846	3.846	3.444	3.444
26	2.138	2.138	1.994	1.994
27	2.321	2.321	1.821	1.821
28	2.172	2.172	1.995	1.995

У таблиці відсутні рядки для питань 19 та 20, тому що їх треба було вилучити із вхідних даних для моделі через наявність проблеми зі збіжністю – так само як і для першої моделі. Проте ця проблема має трохи іншу природу.

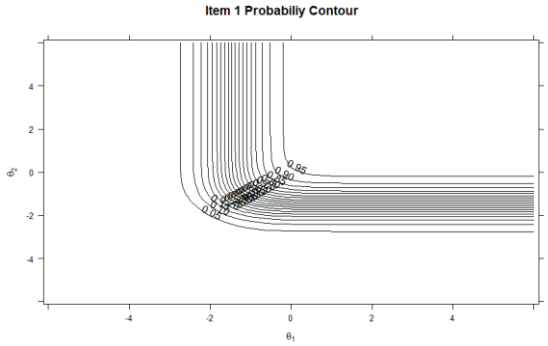
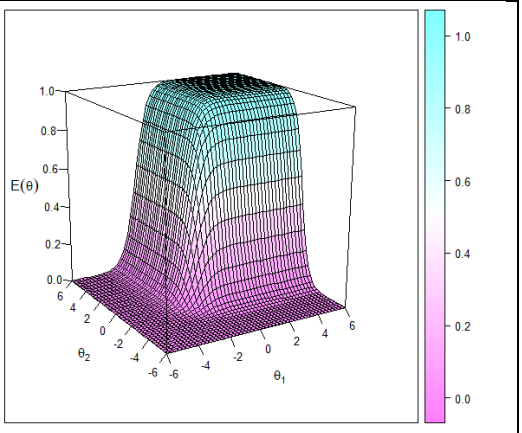
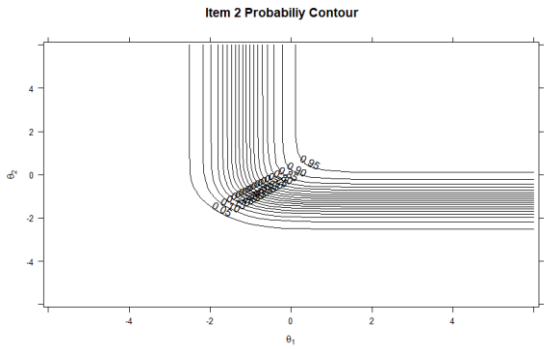
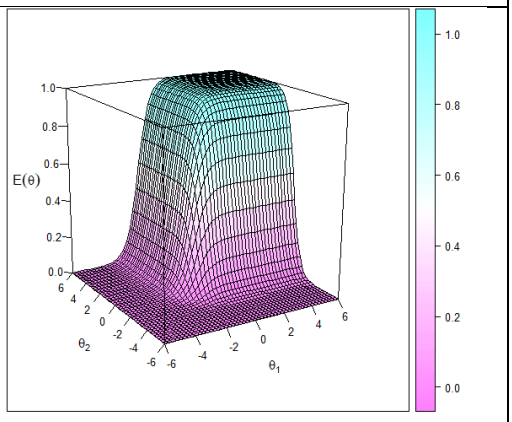
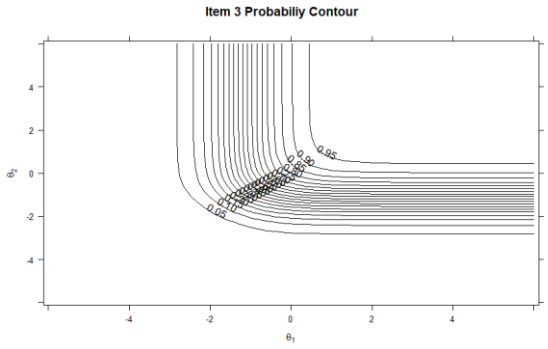
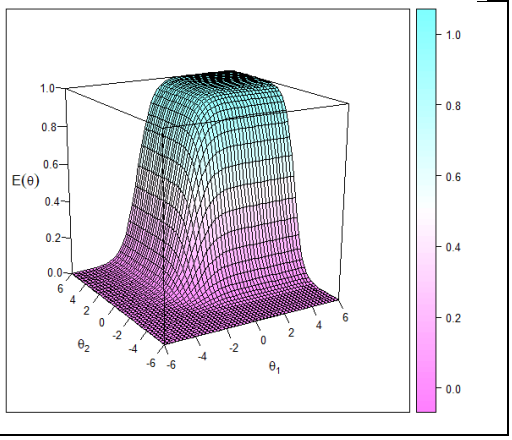
Для алгоритму MH-RM, коли кількість ітерацій стає дуже високою (наприклад, більше 1500) або коли Max Change = 0.25, то значення друкуються на консолі занадто часто (вказуючи на те, що параметри були обмежені, оскільки вони рухались із кроком, більшим за 0,25. В такому випадку модель може бути погано визначеною або мати дуже плоску інформаційну поверхню, а справжні оцінки параметрів максимальної правдоподібності може бути важко знайти. Стандартні помилки обчислюються після збіжності моделі шляхом передачі SE = TRUE, щоб виконати додатковий етап MH-RM, але розглядаючи оцінки максимальної правдоподібності як фіксовані точки.

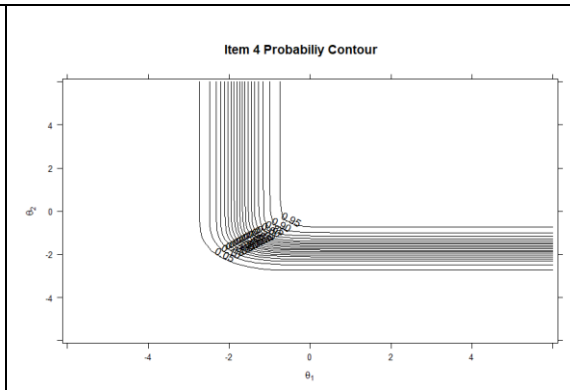
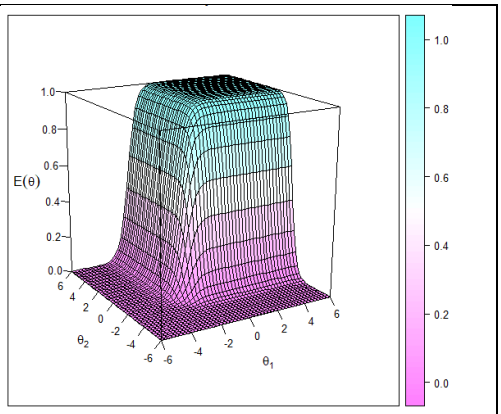
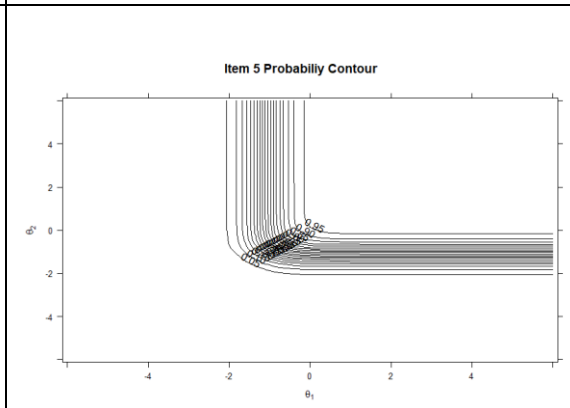
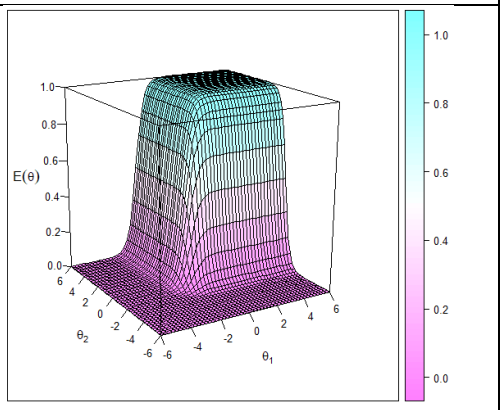
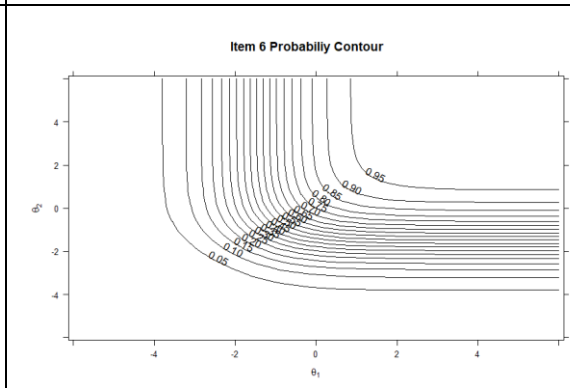
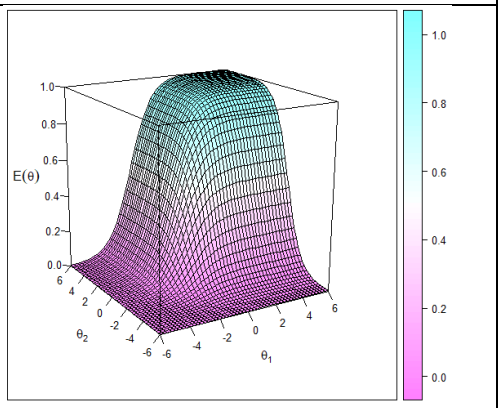
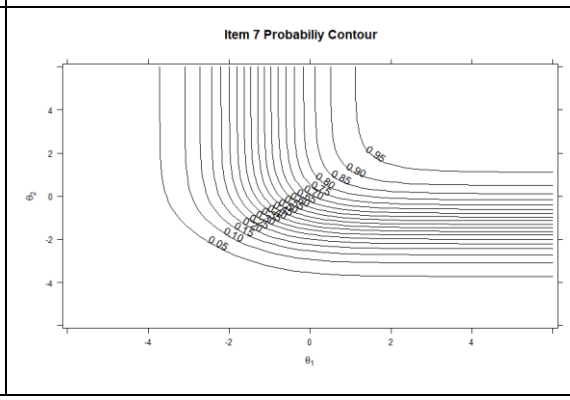
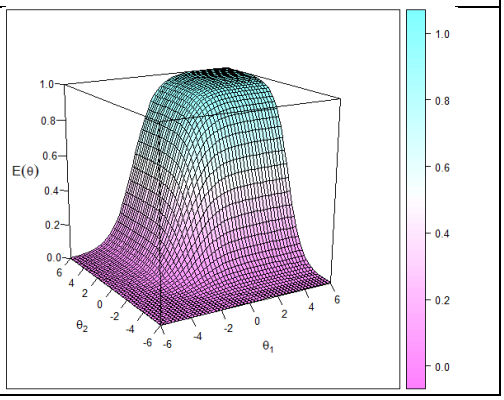
Геометричне відображення, а саме контурні лінії питань та поверхні очікуваних оцінок іспитників, наведено у таблиці 4.4 для питань 1-11. Значення логарифмічної правдоподібності моделі дорівнює -536.583.

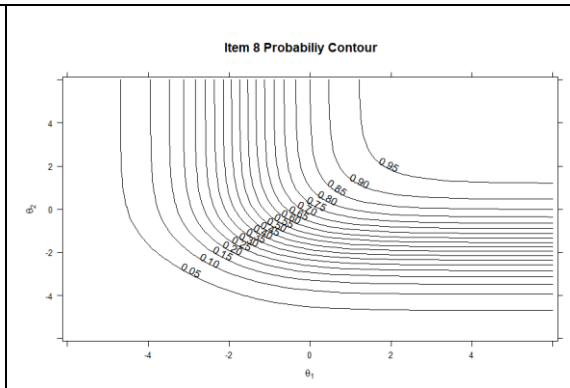
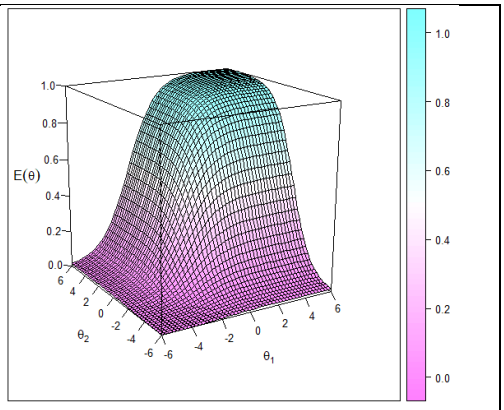
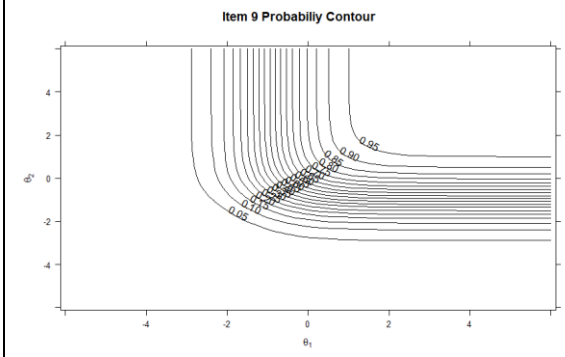
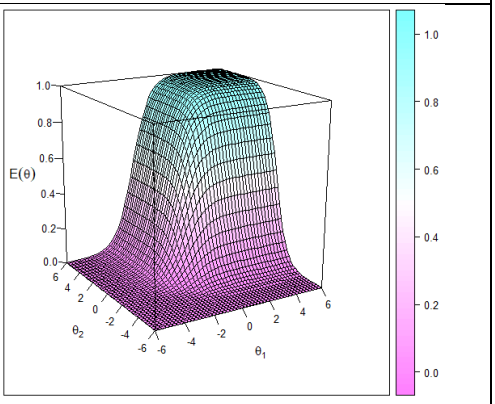
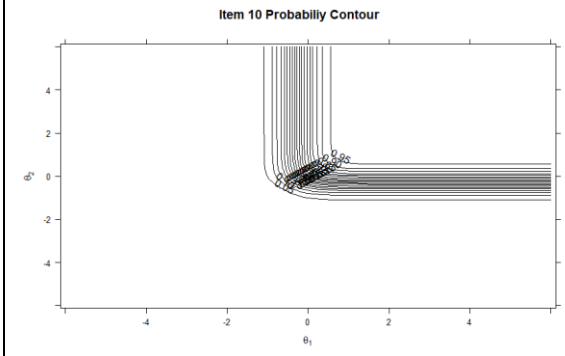
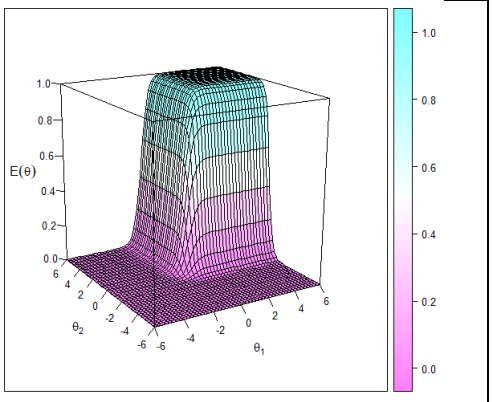
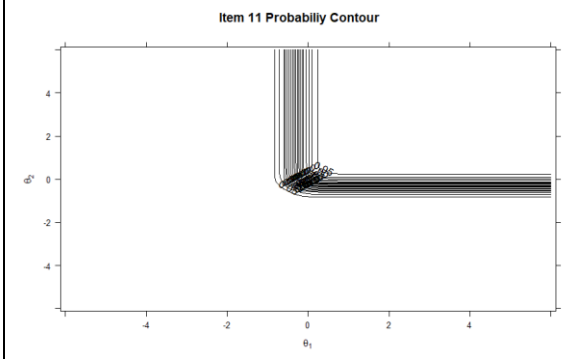
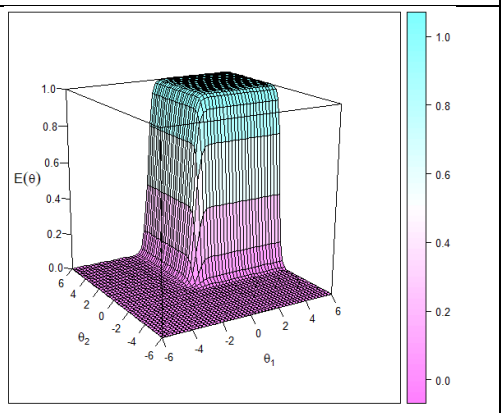
Проаналізуємо результати моделювання, отримані графічно, для питань 1-11. Можна побачити, що ансамблі контурних лінії питань мають Г-образну форму. Тому робиться висновок, що вони добре оцінюють обидві ознаки. Дивлячись на щільність контурних ліній, можна побачити, що

питання 4,5,10,11 мають більше значення диференціюючої здатності, питання 6-8 - менше.

Таблиця 4.4. *Результати моделювання контурних ліній та поверхонь очікуваних оцінок іспитників за допомогою частково компенсаторної MIRT-моделі*

№ пит	Контурні лінії питання	Поверхня очікуваних оцінок іспитників
1		
2		
3		

4	 Item 4 Probabiliy Contour θ1 θ2	 E(θ) θ1 θ2
5	 Item 5 Probabiliy Contour θ1 θ2	 E(θ) θ1 θ2
6	 Item 6 Probabiliy Contour θ1 θ2	 E(θ) θ1 θ2
7	 Item 7 Probabiliy Contour θ1 θ2	 E(θ) θ1 θ2

8		
9		
10		
11		

Загалом можна зробити висновки, що компенсаторна та частково компенсаторна моделі дають співставні результати для аналізу оцінок

іспитників. Проте з першої моделі було легше та наочніше зробити декілька висновків.

4.3. Аналіз результатів моделювання

Для аналізу результатів моделювання використовувалися критерії AIC, SABIC, HQ, BIC та метод оцінки максимальної правдоподібності, розглянуті у розділі 2. При моделюванні вхідних даних, зображених на рисунку 4.1, за допомогою компенсаторної та частково компенсаторної двовимірних MIRT-моделей було отримано значення критеріїв, перелічених у таблиці 4.5.

Таблиця 4.5. *Критерії порівняння моделей*

	AIC	**SABIC**	**HQ**	**BIC**	**logLik**
C2PL	1074.348	993.427	1137.427	1235.613	-460.174
PC2PL	1281.166	1171.870	1366.364	1498.977	-536.583

Як бачимо, результати першої моделі є кращими за всіма критеріями (оскільки мають менші значення), та мають більше значення максимальної правдоподібності. Тому для використаних даних слід обрати компенсаторну двовимірну MIRT-модель.

ВИСНОВКИ

У даній магістерській дисертації було виконано оцінку латентних параметрів некомпенсаторних MIRT-моделей за допомогою статистичних методів. Проведене дослідження дозволило зробити висновки щодо результатів модульної контрольної роботи з предмету «Теорія ймовірностей».

Основні наукові та практичні результати роботи полягають у наступному:

1. Було виконано огляд одновимірних та багатовимірних IRT-моделей, наведено приклади їх застосування до роботи з даними.

2. Було розглянуто методи оцінювання якості моделювання, зокрема оцінку моделі методом граничної максимальної правдоподібності, наведено методи визначення необхідних розмірностей моделей (паралельний аналіз, критерій χ^2, кластерний метод, критерій Кайзера-Гутмана).

3. Розроблено математичний опис компенсаторних та некомпенсаторних MIRT-моделей, що використовувалися при моделюванні.

4. Для моделювання було використано компенсаторну 2PL-модель та некомпенсаторну 2PL-модель, вхідні дані для моделювання були попередньо оброблені. Було здійснено моделювання в середовищі RStudio, було отримано аналітичні результати та їх графічна візуалізація.

5. Було виконано аналіз результатів моделювання аналітичним та геометричним способом.

6. Порівнюючи результати моделювання двох моделей, було обрано як кращу компенсаторну двовимірну MIRT-модель. Для порівняння було використано критерій AIC, SABIC, HQ, BIC та оцінку максимальної правдоподібності моделей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Ackerman T. Using multidimensional item response theory to understand what items and tests are measuring / T.A Ackerman. // Applied Measurement in Education. – 1994. – №7. – С. 255–278.

2. Ackerman T. An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. / T. Ackerman, M. Gierl, C. Walker. // Educational Measurement: Issues and Practice. – 2003. – №22. – С. 37–53.

3. Hambleton R.K Fundamentals of item response theory. / R.K Hambleton, H. Swaminathan, H. Roger. – Newbury Park: SAGE Publications, 1991.

4. Beguin A. MCMC estimation and some model-fit analysis of multidimensional IRT models. / A. Beguin, C. Glas. // Psychometrika. – 2001. – №66. – С. 541–562.

5. Traub R. Classical test theory in historical perspective. / R. Traub. // Educational Measurement: Issues and Practice. – 1997. – №16. – С. 8–14.

6. Ackerman T. A didactic explaination of item bias, item impact, and item validity form a multidimensional perspective. / T. A. Ackerman. // Journal of Educational Measurement. – 1992. – №29. – С. 67–91.

7. Segall D. General ability measurement: An application of multidimensional item response theory. / D. O. Segall. // Psychometrika. – 2001. – №66. – С. 79–97.

8. Allen D. Introducing multidimensional item response modeling in health behavior and health education research. / D. Allen, M. Wilson. // Health Education Research. – 2006. – №21. – С. 73– 84.

9. Bolt D. Conditional covaariance-based representation of multidimensional test struccture. / D. M. Bolt. // Applied Psychologial Measurement. – 2001. – №25. – С. 244–257.

10. Vollema M. The multidiemsionalty of self-report schizotypy. / M. Vollema, H. Hoijtink. // Schizophrenia Bulletin. – 2000. – №26. – С. 565–575.

11. Zhang J. Calibration of response data using MIRT models with simple and mixed structure. / J. Zhang. // Applied Psychological Measurement. – 2012. – №36. – С. 375–398.

12. Reckase M. Multidimensional item response theory. / M. D. Reckase. – New York: Springer, 2009.

13. Sympson J. A model for testing with multidimensional items. / J. Sympson, D. Weiss // proceedings of the 1977 computerized adaptive testing conference / J. Sympson, D. Weiss. – Minneapolis: University ofMinnesota, 1978. – С. 82–98.

14. Hartig J. Multidimensional IRT models for the assessment of competencies. / J. Hartig, J. Hohler. // Studies in Educational Evaluation. – 2009. – №35. – С. 57–63.

15. Babcock B. Estimating a noncompensatory IRT model using metropolis within Gibbs sampling. / B. Babcock. // Applied Psychological Measurement. – 2011. – №35. – С. 317–329.

16. Reckase M. The difficulty of test items that measure more than one ability. / M. D. Reckase. // Applied Psychological Measurement. – 1985. – №9. – С. 401–412.

17. Svetina D. Assessing dimensionality of noncompensatory item response thoery with complex structures. / D. Svetina. // Educational and Psycholgical Measurement. – 2012. – №73. – С. 312–338.

18. McKinely R. A comparison of several goodness-of-fit statistics. / R. McKinely, C. Mills. // Applied Psychological Measurement. – 1985. – №9. – С. 49–57.

19. Stone C. Monte Carlo based null distribution for and alternative hoodness-of-fit test statistic in IRT models. / C. A. Stone. // Journal of Educational Measurement. – 2000. – №37. – С. 58–75.

20. Bolt D. Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. / D. Bolt, V. Lall. // Applied Psychological Measurement. – 2003. – №27. – С. 395–414.

21. Comparison of two logistic multidimensional item response theory models. / [J. Spray, T. Davey, M. Reckase та ін.]. – Iowa City: ACT Inc, 1990.

22. Zhang B. Evaluating item fit for multidimensional item response models. / B. Zhang, C. Stone. // Educational and Psychological Measurement. – 2008. – №68. – С. 181–196.

23. Conditional covariance-based nonparametric multidimensionality assessment. / [W. Stout, B. Habing, J. Douglas та ін.]. // Applied Psychological Measurement. – 1996. – №20. – С. 331–354.

24. Finch H. Multidimensional item response theory parameter estimation with nonsimple structure items. / H. Finch. // Applied Psychological Measurement. – 2011. – №35. – С. 67–82.

25. Orlando M. Likelihood based item-fit indices for dichotomous item response theory models. / M. Orlando, D. Thissen. // Applied Psychological Measurement. – 2000. – №24. – С. 50–64.

26. DeAyala R. The theory and practice of item response theory. / R. DeAyala. – New York: Guilford Press, 2009.

27. Hong H. Nonnested model selection criteria. / H. Hong, B. Preston. – Durham: Working paper, 2006.

28. Clarke K. Testing nonnested models of international relations: Reevaluating realism. / K. A. Clarke. // American Journal of Political Science. – 2001. – №45. – С. 724–744.

29. Osteen P. An introduction to using multidimensional item response theory to assess laten factor structures. / P. Osteen. // Journal of the Society for Social Work and Research. – 2010. – №1. – С. 66–82.

30. Czado C. Non-nested model selection for spatial count regression models with application to health insurance. / C. Czado, H. Schabenberger, V. Erhardt. // Statistical Papers. – 2014. – №55. – С. 1– 22.

31. Akaike H. A new look at the statistical model identification. / H. Akaike. // Free Transactions on Cutomatic Control. – 1974. – №19. – С. 716–723.

32. Long J. Regression models for categorical and limited dependent variables. / J. S. Long. – Thousand Oaks: SAGE Publications, 1997.

33. Fox J. Applied regression analysis and generalized linear models. / J. Fox. – Thousand Oaks: SAGE Publications, 2008.

34. Clarke K. Nonnested model testing for world politics assessing binary choice models. / K. A. Clarke // American Political Science Association Conference. / K. A. Clarke. – Boston, 1998.

35. Vuong Q. Likelihood ratio tests for model selection and non-nested hypotheses. / Q. H. Vuong. // Econometrica: Journal of the Econometric Society. – 1989. – №57. – С. 307–333.

36. Genius M. Model selection and tests for non-nested contingent valuation models: An assessment of methods. / M. Genius, E. Strazzera. – Milan, Italy: Fondazione Eni Enrico Mattei, 2001.

37. Pesaran M. Non-nested hypothesis. / M. Pesaran, R. Ulloa. // The New Palgrave Dictionary of Economics. – 2008. – №2. – С. 609–642.

38. Clarke K. Nonparametric model discrimination in international relations. / K. A. Clarke. // Journal of Conflict Resolution. – 2003. – №47. – С. 72–93.

39. Clarke K. Discriminating methods: Tests for non-nested discrete choice models. / K. Clarke, C. Signorino. // Political Science. – 2010. – №38. – С. 368–388.

40. Clarke K. A simple distribution-free test for nonnested model selection. / K. A. Clarke. // Political Analysis. – 2007. – №15. – С. 347–363.

41. Ledesma R. Determining the number of factors to retain in EFA: an easy- to-use computer program for carrying out parallel analysis. / R. Ledesma, P. Valero-Mora. – 2007: Practical Assessment, Research & Evaluation.

42. Schilling S. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. / S. Schilling, R. Bock. // Psychometrika. – 2005. – №70. – С. 533–555.

43. Ul Hassan M. Discrimination with unidimensional and multidimensional item response theory models for educational data [Електронний ресурс] / M. Ul Hassan, F. Miller // Communications in Statistics - Simulation and Computation. – 2019. – Режим доступу до ресурсу: https://www.tandfonline.com/doi/full/10.1080/03610918.2019.1705344.