UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

# APLIKÁCIA ITEM RESPONSE THEORY PRI NÁVRHU RATINGOVÝCH SYSTÉMOV

Diplomová práca

2012

Bc. Ivan Labáth

**UNIVERZITA KOMENSKÉHO V BRATISLAVE**
**FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY**

# APLIKÁCIA ITEM RESPONSE THEORY PRI NÁVRHU RATINGOVÝCH SYSTÉMOV

Diplomová práca

| | |
|---|---|
| Študijný program: | Informatika |
| Študijný odbor: | 2508 Informatika |
| Školiteľ: | RNDr. Michal Forišek, PhD. |

**Bratislava 2012**

**Bc. Ivan Labáth**


Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

# ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Bc. Ivan Labath
**Študijný program:** informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
**Študijný odbor:** 9.2.1. informatika
**Typ záverečnej práce:** diplomová
**Jazyk záverečnej práce:** slovenský

**Názov:** Aplikácia Item Response Theory pri návrhu ratingových systémov

**Cieľ:** Cieľom práce je analyzovať aktuálnu situáciu pri návrhu rôznych ratingových systémov, s dôrazom na využitie Bayesovského usudzovania a Item Response Theory. Ďalším cieľom práce je preskúmať možnosti využitia Item Response Theory pri návrhu ratingových systémov v nových prostrediach. Na základe analýzy problematiky môže byť vhodné ako súčasť práce vytvoriť praktickú implementáciu takéhoto ratingového systému, nie je to však nutné.

**Vedúci:** RNDr. Michal Forišek, PhD.
**Katedra:** FMFI.KI - Katedra informatiky

**Spôsob sprístupnenia elektronickej verzie práce:**
bez obmedzenia

**Dátum zadania:** 27.10.2010

**Dátum schválenia:** 28.10.2010 prof. RNDr. Branislav Rovan, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím citovaných zdrojov.

. . . . . .

Chcel by som sa poďakovať svojmu školiteľovi, Michalovi Foriškovi, za nadhľad a trpezlivosť, ktorými ma podporoval prí písaní tejto práce.
Pánovi Jánovi Somorčíkovi by sa chcel poďakovať za konzultácie ohľadom štatistických metód.
Pánovi Richardovi Ostertágovi by sa chcel poďakovať za poskytnutie testov a ich výsledkov z predmetu princípy počítačov.

# Abstrakt

V práci sa zaoberáme možnosťou vytvorenia ratingového systému na základe Item Response Theory. Konkrétne, ukázali sme konštrukciu boostrapových a bayesovských asymetrických intervalových odhadov presnosti schopností pridelených súťažiacim, priamo z modelu bez predpokladov rozdelenia chyby. Pre bayesovský odhad sme ukázali monotónnosť hodnotenia úlohovo porovnateľných súťažiacich, za platnosti určitých podmienok. Ďalej sme urobili prehľad dostupnosti programov potrebných na hodnotenie pomocou IRT, kde sme kvôli nedostatku implementácie JML algoritmov na odhad naprogramovali vlastný nastroj podporujúci JML ako aj L-BFGS-B algoritmus a spomínané odhady chyby. Na záver sme pomocou Item Response Theory na príklade jedného predmetu preukázali, že študenti poznajú testy z minulých skúšok.

Kľúčové slová: Item Response Theory, rating, súťaž, odhad chyby, intervalový odhad, asymetrický odhad, monotónnosť

# Abstract

We study the possibility of an Item Response Theory based rating system. In particular, we show the construction of two types of asymmetric interval error bounds of ability assessments directly from the model, without assuming prior error distribution. For one of them, constructed via Bayesian inference, we show a restricted form of monotonicity on task-comparable contestants. On the topic of IRT fitting software, we assess the availability of free implementations and provide our own implementation using JML and L-BFGS-B algorithms. Finally, correlating median ratings with task reuse we show student knowledge of prior tests.

Key words: Item Response Theory, rating, contest, reliability, interval bound, asymmetric bound, monotonicity

# Predhovor

Súťaže v riešení úloh sú populárne v rôznych oblastiach. Klasický spôsob hodnotenia je podľa počtu úloh, prípadne zohľadňujúc čas a prideľujúc rôzny počet bodov podľa náročnosti úlohy. Tento systém je jednoduchý a efektívny, na zmysluplnosť porovnania účastníkov však vyžaduje, aby všetci riešili rovnaké úlohy. Vznikli a vznikajú mnohé riešenia, ktoré tieto ohraničenia odstraňujú, každé so svojími špecifickými vlastnosťami.

V štyridsiatych a päťdesiatych rokoch minulého storočia, psychológovia sa zamýšľali nad problémom merania vnútorných charakteristík osôb a dospeli k teórii, kde test už nemeral skóre, ale hodnotu meranej charakteristiky. Táto jemná, ale podstatná zmena, bola jedena z prvých krokov k formovaniu toho, čo dnes nazývame Item Response Theory. Zmena abstrakcie však mala svoju cenu v podobe výpočtovej náročnosti, ktorá bola na tú dobu príliš vysoká.

S rozvojom počítačov sa pomaly rozvíjalo aj využitie teórie, najmä v oblasti psychometrie, keď už bolo možné týmto prístupom zhodnotiť testy rozumnej veľkosti. Dodnes je však iba malým prúdom v oceáne, kde panuje zaužívaná teória testov. Ako hovorí Streiner: „Veda sa mení jeden pohreb za druhým."

Psychometria a súťaže majú veľký prekryv v spoločnom predmete merania, podstatne sa však líšia vo svojom cieli. Kde psychológov zaujíma najmä všeobecný stav a na jednotlivcoch v meraní nezáleží, v súťažiach je práve na nich dôraz. Využitie teórie z psychometrie v súťažiach nie je priamočiara záležitosť, nosí však so sebou silné teoretické poznatky a skúsenosti.

Pre uplatnenie v súťažiach, teória musí preukázať spoľahlivosť a prínos voči zaužívanej tradícii. Naša práca je ďalším krokom v snahe sprístupniť výhody tejto teórie pre využitie v súťažiach a všeobecne skúškach schopností.

# Obsah

# Zoznam obrázkov

# Úvod

Item Response Theory ako teóra testovania pochádzajúca z psychometrie, prináša mnohé výhody, najmä ohybnosťou v tvorbe testov a väčšou výpovednosťou.

Naša práca je ďalším krokom v snahe sprístupniť výhody tejto teórie pre využitie v oblasti ratingu súťaží a všeobecne skúšok, ktoré majú mierne iné požiadavky a záujmy než v psychológii.

Konkrétne rozdiely vidíme v potrebe presne a jednoznačne určiť poradie jednotlivcov, na čo sa zaoberáme odhadmi chyby, teda presnosti namerania jednotlivcov a poukazujeme na vplyv apriórnych predpokladov v zaužívanom algoritme hodnotenia.

**Kapitola 1** má za cieľ čitateľovi poskytnúť znalosti, ktoré nie sú našim výskumom, ale poskytnú potrebný kontext pre jadro našej práce. Začína určením modelu skúšok na ktorý sa naša práca vzťahuje a uvedením pojmu ratingu. Ďalej porovnáva klasickú teóriu testov s novou Item Response Theory, ktorej sa venuje do konca v podrobnejšom rozbore.

**Kapitola 2** je zameraná na teoretický rozbor. Po krátkom uvedení modelu ratingového algoritmu, prezentuje náš prínos v oblasti odhadovania presnosti hodnotenia súťažiacich. Podáva súhrn existujúcich algoritmov, a rozbor vlastností, kde poukazuje na ich nedostatky, na ktoré následne navrhuje dve riešenia založené na Item Response Theory.

**Kapitola 3** je implementačnou a výskumnou časťou práce. Začína prehľadom existujících implementácií. Ďalej prezentuje niektoré špecifiká našej vlastnej implementácie a porovnáva dostupné algoritmy z hľadiska vhodnosti na súťaže. Na záver prezentuje výsledky jednoduchého výskumu špecifického pre Item Response Theory.

# Kapitola 1

# Rating a teória testov

V tejto kapitole najprv vyhraníme oblasť v ktorej sa budeme pohybovať. Ďalej budeme prezentovať prehľad problematiky a známe výsledky v danej oblasti.

V rámci prehľadu prezentujeme aj úvodné definície pojmov, ktoré budú v ďalších kapitolách používané. Mnohé pojmy sa týkajú spoločenskej oblasti, teda sú veľmi zložité a tematikami dlhých filozofických traktátov. Veríme, že čitateľovi nebude prekážať, keď pojmy spresníme iba v miere potrebnej na dostatočne presné vyjadrenie myšlienok a odpustíme si filozofické traktáty o ďalších súvislostiach.

## 1.1 Model skúšok

Na jednej strane by sme chceli pracovať so súťažami čo všeobecnejšie, a na druhej sme nútení používať model súťaže, ktorý skliesňuje a skresľuje samotné súťaže, aby sme mohli niečo ucelene a všeobecne tvrdiť. Definujme si teda základný model súťaže, kde budeme predpokladať čo najmenej, ale budeme s ním môcť dostatočne pohodlne pracovať.

**Poznámka.** *Súťaže, testy a všeobecne skúšky pre naše účely možno svojou formou považovať za zhodné, preto v ďalšom texte nebudeme rozlišovať medzi týmito pojmami.*

V tejto práci sa hlavne zameriame na skúšky, kde sú skúšané osoby pracujúce samostatne, teda nie v skupine, na viacerých jednotlivých úlohách, ktoré sú samostatne hodnotené a z týchto čiastkových výsledkov je určené hodnotenie súťažiaceho.

Na začiatok si definujeme niekoľko základných pojmov, s ktorými budeme ďalej pracovať.

**Definícia 1.1.** Množinu subjektov, ktorý sa zúčastnili skúšky, ktorú rozoberáme budeme nazývať množina skúšaných subjektov. Cieľom ratingový algoritmu bude hodnotiť týchto subjektov.
Pre ďalšie potreby budeme túto možinu označovať $S$.

**Definícia 1.2.** Množina úloh, ktoré sa vyskytovali na skúške a skúšaný subjekti ich mali riešiť, budeme nazývať množina úloh.
Pre ďalšie potreby budeme túto množinu označovať $U$.

**Definícia 1.3.** Hodnotenie súťažiaceho na jednej úlohe budeme nazývať čiastkový výsledok. Ratingový algoritmus bude čiastkové výsledky sumarizovať v hodnotenia.
Pokiaľ nebude povedané ináč, budeme predpokladať, že sú čiastkové výsledky z množiny $\{0, 1\}$.
Doménu čiastkových výsledkov budeme označovať $Q$.

V našej práci budeme skúšky skúmať učitým spôsobom, s využitím modelov Item Response Theory. Na tie účely budeme potrebovať vhodný tvar skúšky. Ďalej budeme predpokladať, že skúška spĺňa nasledovné kritériá.

**samostatná práca súťažiacich**

Súťažiaci pracujú jednotlivo, pokiaľ možno s čo najmenším vplyvom ostatných na ich prácu.

Chceme merať schopnosti jednotlivých súťažiacich, kde práca v tímoch a vzájomné ovplyvňovanie by skresľovalo výsledky a nútilo k zložitým, prípadne pochybným, spôsobom odstraňovania daných vplyvov.

**rozdelenie skúšky na úlohy**

Skúška sa skladá z menších jednotlivých celkov, úloh, pričom tieto čiastkové výsledky na úlohách sú jediným podstatným činiteľom ich hodnotenia.

Náš prínos je v spôsobe spracovania čiastkových výsledkov na jednotlivých úlohách, z ktorých sa skúška skladá. S ohľadom, že v základnom modeli pracujeme s binárne hodnotenými úlohami, je vhodné, aby úloh bolo aspoň zopár, aby rozdiely medzi súťažiacimi boli merateľné.

**konštantnosť úloh**

Úlohy sú stále a nemenné. Pre každú úlohu platí, že je jedna a tá istá pre každého, kto ju podstúpi riešiť.

Budeme charakterizovať vlastnosti jednotlivých úloh a preto je dôležité, aby ich vlastnosti boli nemenné. Stálosť úloh nám umožní merať vnútornú schopnosť súťažiacich na stálej škále. Na rozdiel od mnohých postupov však nepredpokladáme, že sú súťažiaci skúšaní na rovnakej sade úloh. Potrebujeme však výrazný prekryv úloh medzi súťažiacimi.

**nezávislosť úloh**

Čiastkové výsledky sú navzájom nezávislé cez všetky úlohy a súťažiacich, mimo previazanosti cez meranú veličinu.

Predpoklad, že úspechy súťažiacich na úlohách sú navzájom nezávislé je mierne nadsadený, no je nutným predpokladom štatistického spracovania pomocou Item Response Teory. Malé vzájomné závislosti by nemali významne ovplyvňovať výsledky, musia však naozaj byť malé.

Síce v žiadnej netriviálnej súťaži nebudú splnené všetky predpoklady, mnohé skúšky sa dostatočne približujú tejto idealizácii, aby sa odchýlky závažne neprejavovali na výsledku.

Príkladom skúšok, ktoré spĺňajú tieto podmienky môžu byť súťaže z programovania, matematiky, biológie alebo školské testy. Konkrétne Forišek vo

svojej práci [For09a] skúmal informatickú súťaž TopCoder a krajské kolo slovenskej olympiády v informatike a v našej práci analyzujeme testy z princípov počítačov (`1-INF-130`).

## 1.2 Rating

V súťažiach, testoch a všeobecne skúškach okrem iného býva cieľom určiť výťaza, prípadne nejaké porovnanie účastníkov. Jeden z problémov na ceste k tomuto cieľu je ohodnotenie účastníkov. Tento prechod z faktickej skutočnosti, ako účastníci konali a čo urobili, na hodnotenie, porovnanie účastníkov bude pre naše potreby rating. Síce riešenia často bývajú jednoduché, vôbec to neznamená, že samotný problém hodnotenia je jednoduchý.

Z pohľadu uskutočnenia, hlavným cieľom hodnotenia býva určiť vzájomné zoradenie alebo hodnosť resp. skóre účastníkov. Inými cieľami, ako je napríklad cvičenie, alebo učenie detí súťaživosti (viď [For09a]) sa v tejto práci nebudeme zaoberať.

Ako oficiálny výsledok môže stačiť určenie výťaza - toho najlepšieho, prípadne zopár najlepších, čo je väčšinou aj najjednoduchšie. Často je potrebné aj čiastočné alebo úplné usporiadanie účastníkov. Usporiadanie je jednoduchšie preto, že je iba relatívne, nepotrebuje škálu a nemeria veľkosť rozdielov medzi účastníkmi.

Obsiažnejšiu informáciu o účastníkoch dáva prideľovanie hodností, čo zodpovedá meraniu, nie iba porovnávaniu schopností. Cieľom je, aby hodnotenie jednoducho vyjadrovalo úspechy účastníkov. Väčšinou sa ako hodnosť používa jednoduché reálne číslo ako napr. pre šach, prípadne aj s odhadom presnosti merania ako napr. TopCoder a TrueSkill.

Hodnosť má prednosť najmä v tom, že vyjadruje schopnosti na istej škále, ktorá je viac alebo menej stála. Umožňuje to širšie porovnávanie schopností v rôznych časoch a medzi účastníkmi, ktorí neboli priamo hodnotený. Skok do diaľky dĺžky $7m$ má $7m$ teraz a mal aj pred desiatimi rokmi v Afganistane, ale výhra v prvej lige ľadového hokeju na Slovensku sa vôbec nepodobá tej v

Afganistane, ani terajšej ani tej desať ročnej, či už slovenskej alebo afganskej.

Škála tiež do určitej miery vyjadruje veľkosti rozdielov medzi účastníkmi. Ak výsledky skoku do diaľky sú: 710, 700, 410 a 400 centimetrov, hovorí nám to viacej ako poradie: prvý, druhý, tretí a štvrtý. Na druhej strane, nie je decimeter ako decimeter. Pojmy schopnosti a úspechu sú relatívne a veľmi zložité. Ako zhŕňa Streiner v [Str10], urobiť perfektne lineárnu škálu na meranie schopností je známy a dosiaľ nevyriešený problém.

Prideľovanie hodností, ak sa dobre zvolí, má prednosť väčšej výpovednosti o zúčastnených. Problémom je, že väčšia zložitosť ratingového systému, vzbudzuje aj väčšie pochybnosti o ňom.

**ratingový systém** je celkový spôsob hodnotenia skúšaných osôb.

Je to spôsob prechodu z faktickej skutočnosti, ako účastníci konali a čo urobili, na hodnotenie, porovnanie účastníkov. V prípade skúšok spĺňajúcich naše kritéria, jeho neodbytnou súčasťou je aj ratingový algoritmus.

**ratingový algoritmus** je spôsob výpočtu hodnotenia súťažiacich z čiastkových výsledkov.

Ratingový algoritmus je časť ratingového systému, ktorý bude stredobodom pozornosti našej práce. Algoritmus zhŕňa výsledky z jednotlivých úloh a pretvára ich v jednotné hodnotenie súťažiaceho.

Pre ďalšiu práci si formálne definujeme základný ratingový algoritmus. Iba základný, ktorý bude ďalej rozširovaný preto, že konkrétne algoritmy môžu, za prvé, používať aj rôzne iné informácie okrem základných výsledkov na úlohách. Za druhé, môžu podávať okrem samotného hodnotenia súťažiacich aj iné užitočné informácie, ktoré ak sa vzťahujú iba na daný algoritmus a sú s ním hlboko prepojené, nemalo by zmysel oddeľovať.

**Definícia 1.4.** Doména hodnotení je množina, ktorej prvky ratingový algoritmus priraďuje subjektom na vyjadrenie hodnotenia.
Túto množinu budeme ďalej označovať $H$.

**Definícia 1.5.** Základný ratingový algoritmus je funkcia z postupnosti trojíc z množín súťažiacich, úloh a výsledkov do čiastočnej funkcie z množiny súťažiacich do množiny hodnotení.

$$(S \times U \times Q)^* \to (S \to H) \tag{1.1}$$

V základnom algoritme nerozlišujeme, či boli úlohy rozdelené na kolá, prípadne testy alebo nie. Výstup sme definovali ako čiastočnú funkciu, aby sme zohľadnili fakt, že nie všetci súťažiaci musia byť dostatočne nameraný. Napríklad pre prázdny vstup môže byť výstupom nikde nedefinovaná funkcia.

Teraz dočasne odložíme špecifiká ratingu a nahliadneme na spomínané pojmy v rámci všeobecnejšej teórie testov.

*Napíšem aj čo všetko môžeme od ratingu chcieť?*

## 1.3 Teória testov

Psychometria je oblasť vedy, ktorá sa zaoberá teóriami a technikami merania duševných vlastností ľudí. Teória testov je formalizáciou merania pomocou testov, pričom sa test chápe ako súbor úloh alebo otázok, na ktoré testovaní subjekti odpovedajú. Formalizácia pojmu testu poskytuje základ na teoretické a štatistické spracovanie samotných testov, ako aj ich výsledkov a vztahov medzi nimi a meranými veličinami.

V psychometrii sa prevažne používajú dva typy teórií testovania a to Classical Test Theory (klasická teória testov, ďalej len CTT) a Item Response Theory (teória zodpovedania otázok, ďalej len IRT). Obe z nich sú vlastne súborom množstva konkrétnych teórií, ale sú medzi nimi zásadné rozdiely. Konečný cieľ majú rovnaký, líšia sa však úrovňou základných pojmov a rozsahom teórií na nich založených.

CTT skúma výsledky na celom teste, neberúc do ohľadu špecifiká jednotlivých úloh. Pri skúmaní vzťahu medzi meranou vlastnosťou a výsledkami je meranou veličinou teoreticky očakávané ohodnotenie subjektu na teste, ktoré sa odhaduje z toho prakticky dosiahnutého, pomocou chybovosti testu

na subjektoch.

IRT skúma výsledky na samotných úlohách, nezávisle od zloženia konkrétnych testov. Predpokladá existenciu vnútornej vlastnosti (latent trait), ktorú úlohy merajú. Vlastnosti úloh modeluje pravdepodobnosťou odpovedí na úlohy vzhľadom na meranú vlastnosť funkciou zvanou charakteristická krivka úlohy (Item Characteristic Curve). V tomto modeli potom IRT odhaduje charakteristiky úloh a vnútorné vlastnosti, ktoré sedia na skúmané údaje.

*Väčšina ratingových systémov je založená na CTT, niektoré na bayesovských odhadoch ako šachovský Elo, informatický TopCoder alebo TrueSkill na XBox-och. Bližší popis týchto systémov a ďalšie odkazy sa nachádzajú vo Foriškovej práci [For09a].*

Rozhodli sme sa skúmať možnosti ratingu založeného na Item Response Theory preto, že voči Classical Test Theory prináša nasledovné výhody:

**priamejšie koncepty**

Merané schopnosti nie sú zbytočne definované cez testy, ale priamo na úlohách, ktoré boli riešené. Výskum sa nezameriava na vlastnosti testu, ale priamejšie na osoby a úlohy, prinášajúc menej skreslenia.

**jednotná stupnica**

Výsledok merania je definovaný ako hodnota meranej vlastnosti, nie ako skóre na teste. Táto hodnota je stabilná a prenositeľná na všetky testy merajúce rovnakú vlastnosť, až na lineárny posun a škálovanie a chybu merania. Skóre je definované ako počet bodov, ktorý je výrazne závislejší na zložení testu.

**kombinovanie úloh**

Meranie je definované na jednotnej stupnici a priamo sa týka subjektov a úloh, čo umožňuje subjektov testovať na skoro ľubovoľnom výbere úloh a predsa dostať porovnateľné výsledky. Nie sú potrebné štandardizované testy, keďže potrebná štandardizácia sa dosiahne na úrovni úloh pri ich spoločnom použití ako prirodzená súčasť IRT.

**výtané rozličnosti úloh**

Charakteristiky úloh sú modelované a zohľadnené rozdiely medzi nimi. CTT neskúma jednotlivé úlohy a preto požaduje, aby boli istým spôsobom podobné alebo ich rozloženie vhodne štruktúrované.

**rozsiahlejší model**

Priamym skúmaním odpovedí na jednotlivých úlohach, miesto ich sumáru, sa využíva viacej informácii, čo umožňuje získať porovnateľnú presnosť z menej úloh. Rozsiahlejší model priamo modeluje ďalšie charakteristiky úloh a subjektov ako súčasť jednotného celku.

Item Response Theory má aj svoje nevýhody a teda nie je vždy lepším výberom. Najzávažnejšie nedostatky sú:

**veľkosť vzorky**

Modely IRT majú veľa parametrov, na ktorých sú veľmi závislé. Na ich dobrý odhad je potrebná dostatočne veľká vzorka, ktorá nie je vždy dostupná. CTT nemá toto ohraničenie. V našej práci preskúmame tieto hranice, kde IRT začína mať zmysluplné výsledky.

**nezávisloť úloh**

IRT vo svojich základných modeloch predpokladá, že výsledky na úlohách nekorelují mimo spätosti cez meranú veličinu. Znamená to, že by sa v teste nemali vyskytovať podobné úlohy. Takéto úlohy sa často vyskytujú testoch robených pre CTT, kedže je v nej počet úloh hlavným prostriedkom zníženia chyby merania. Predsa, sú spôsoby ako pojať tieto spätosti v IRT a často možno aj tieto úlohy v IRT proste vynechať bez výrazného zníženia kvality testu.

**zložitosť**

Spôsob hodnotenia v IRT modeloch je výrazne zložitejší, vyžadujúc pri tom špeciálny softvér. Vypočítať hodnotenia ručne by bolo veľmi problematické, až nemožné. Súvis medzi hodnotením a výsledkami je

zložitý a ťažký na vysvetlenie. Pre laikov preto môže byť tento systém divný, nepochopiteľný, až podozrivý.

### 1.3.1 Classical Test Theory

Klasický prístup CTT sme všetci zažili. Najprv sa zostrojí test, ktorý pozostáva z množstva úloh a každá úloha má stanovené rozpätie bodového ohodnotenia. Skúšaná osoba ho vyplní, úlohy sú obodované v rámci medzí a počet bodov je sčítaný. Síce každý inú, ala predsa len podobnú štruktúru v rámci CTT majú testy, písomky, súťaže ako aj dotazníky. Priblížime si spoločné črty CTT, pre porovnanie s IRT.

Teraz už formálne. Budeme používať minulé definície subjektov $S$ (1.1), úloh $U$ (1.2), čiastkových výsledkov $Q$ (1.3) a hodnotenia $H$ (1.4). Špeciálne pre CTT budeme potrebovať definíciu testu ako základného pojmu, nad ktorým je teória definovaná.

**Definícia 1.6.** Test je usporiadaný zoznam úloh, používaný na skúšanie subjektov. V rámci CTT je potrebné jednoznačne určiť test pre potreby jeho ďalšej charakterizácie.
Množinu testov budeme označovať $T$.

**Definícia 1.7.** Funkcia $u$ je funkcia z množiny testov, do množiny konečných postupností úloh. Táto funkcia určuje z akej postupnosti úloh je test zložený.

$$u : T \to U^*_{KON} \tag{1.2}$$

Keďže subjekti môžu byť viackrát testovaný rovnakou úlohou, výsledky vyjadríme na základe testu a čísla úlohy.

**Definícia 1.8.** Funkcia $v$ je funkcia z množiny testovaných subjektov, množiny testov a množiny prirodzených čísel do množiny reálnych čísel. Táto funkcia určuje počet bodov získaných subjektom v danom teste na n-tej úlohe.

$$v : S \times T \times \mathbb{N} \to Q \tag{1.3}$$

**Definícia 1.9.** Funkcia $h$ je funkcia z množiny testovaných subjektov a množiny testov do množiny reálnych čísel. Táto funkcia určuje bodový zisk subjekta v danom teste. Bodový zisk na teste sa rovná súčtu bodových ziskov v jednotlivých úlohách.

$$h : S \times T \to \mathbb{R} \tag{1.4}$$

$$h(s,t) = \sum_i v(s,t,i) \tag{1.5}$$

CTT je súbor teórií testovania, ktoré sú vybudované na spoločnom modeli meraných veličín a vlastností testov. Základ teórie je predpoklad, že nameraný úspech pri konkrétnom testovaní sa skladá z pravej zložky, ktorá je teoreticky očakávaný stredný úspech subjekta v nekonečno testovaní a z náhodnej chyby merania. Namerané hodnoty rozkladá na tieto dve zložky podľa rovnice (1.6).[Kli05]

$$h(s,t) = r(s,t) + e(s,t) \tag{1.6}$$

**Definícia 1.10.** Funkcia $r$ je funkcia z množín testovaných subjektov a testov do množiny reálnych čísel, udávajúca pravú hodnotu testovanej veličiny.

Pravá hodnota je hodnota, ktorá by bola nameraná za neprítomnosti chýb merania.

$$r : S \times T \to \mathbb{R} \tag{1.7}$$

$$\tag{1.8}$$

**Definícia 1.11.** Funkcia $e$ je funkcia z množín testovaných subjektov a testov do množiny reálnych čísel, označujúca chybu merania pri testovaní subjektu testom.

$$e : S \times T \to \mathbb{R} \tag{1.9}$$

**Definícia 1.12.** Funkcie H, R a E sú funkcie z množín testovaných subjektov a testov do množiny náhodných premenných modelujúcich hodnoty funkcií $h$, $r$ a $e$.

CTT spracúva namerané hodnoty $h$ ako realizácie náhodných premenných. Pri ich štatistickom spracovaní CTT má určité predpoklady o náhodných premenných modelujúcich pravé a dosiahnuté úspechy ako aj chyby merania.

**náhodné chyby**

Chyba je náhodná s normálnym rozdelením okolo nuly.

$(\forall s,t)(\exists\sigma)\ \mathrm{E}(s,t) \sim \mathcal{N}(0,\sigma^2)$

Dôležité je, aby chyba mala rozumné rozdelenie okolo 0, aby výsledok testu bol aproximáciou pravej hodnoty. Všeobecne je akceptované, že chyby v (dobre urobených) testoch majú charakter normálneho rozdelenia, čo ďalej umožňuje odhadovať chybu merania.

**navzájom nezávislé chyby**

Štatistika na ktorej sa zakladá CTT vyžaduje, aby chyby boli náhodné a nezávislé. V tom prípade ich je možné rozlišovať od meraných veličín.

$(\forall s,t,s',t')((s,t) \neq (s',t')) \implies \mathrm{E}(s,t)$ a $\mathrm{E}(s',t')$ sú nezávislé.

Zanedbateľnosť systematickej chyby je nie vždy dosiahnuteľná. Pre známe systematické chyby sú teórie, ktoré sa ich snažia eliminovať rôznymi spôsobmi.

**chyby nezávislé od meranej hodnoty**

Na rozumný odhad chyby merania, chyba merania nesmie závisieť od pravej hodnoty meranej veličiny.

$(\forall s,t,s',t')\ \mathrm{E}(s,t)$ a $\mathrm{R}(s',t')$ sú nezávislé

Nepríjemný dôsledok tohto predpokladu je, aj to, že disperzia chýb by mala byť rovnaká pre všetky merania, čo nemusí byť pravda, ale za niektorých okolností to môže byť akceptované.

Nie všetky testy a výskumy sú rovnaké. Podľa požiadavok a realizácie testovania sú potrebné rozdielne modely CTT.

Samotná CTT nemodeluje vzájomné súvislosti rozdielov medzi úlohami a rozdielov v dosiahnutých úspech pri testovaní, čo je hlavný prínos Rasch a IRT modelov. Detailnejšie skúmanie ponecháva na externé metódy, ktoré nie sú integrálnou časťou CTT teórií, ale sú podstatnou súčasťou CTT prístupu k testovaniu.

Nameraný úspech $h$ a teda aj pravý úspech $r$ je sumou nameraných hodnôt $v$ (rovnica 1.5), prípadne lineárnou funkciou ak sú pridané koeficienty. Väčšina CTT teórií na odhad pravej hodnoty ($r$) nerobí nič zložitejšie ako zpriemerovanie nameraných hodnôt ($v$), ak ich je viacej pre príslušné osoby resp. testy, podľa toho čo je cieľom výskumu.

CTT hlavne skúma reliabilitu, variabilitu, veľkost chyby a odhaduje, s akou pravdepodobnosťou namerané výsledky niečo hovoria, alebo nie. Zodpovedá to jednoduchšej predstave testovania, kde za zostaví test, namerajú sa výsledky, pričom výsledky sú dostupné priamo (napr. suma bodov) a štatisticky sa zhodnocuje, či je vôbec niečo namerané.

### 1.3.2 Item Response Theory

Spôsob hodnotenia v IRT je zložitejší než v CTT. Namiesto jednoduchej sumy, hodnotenie je definované cez pravdepodobnostný model, modelujúci pravdepodobnosti odpovedí subjektov vzhľadom na ich schopnosti a vlastnosti úloh. Subjektom sú následne pridelené ako hodnotenia modelované vlastnosti, ktoré maximalizujú pravdepodobnosť výskytu nameraných odpovedí.

Vo všeobecnosti, IRT sa zaoberá meraním nepriamo merateľných vlastností, ktoré nie sú iba schopnosti, ale aj citové nálady, presvedčenia alebo činitele ovplyvňujúce zložité chemické a biologické reakcie. Síce prezentovaná teória je planá všeobecnejšie, ohraničíme sa na oblasť hodnotenia, používajúc príslušnú terminológiu.

Uvedieme a formálne zadefinujeme teóriu IRT a konkrétne modely, ktoré budeme používať. Z časti model skúšok budeme používať definície množín subjektov $S$ (1.1), úloh $U$ (1.2) a domény čiastkových výsledkov $Q$ (1.3), a

z ratingového algoritmu doménu hodnotení $H$ (1.4). Keďže definície v CTT boli zamerané na testy, čo nie je prípad v IRT, zavedieme si nové definície výsledkov, ktoré sa priamo týkajú úloh.

Budeme pracovať so základnými IRT modelmi, kde je úloha buď vyriešená, alebo nie. Čiastkové výsledky preto definujeme ako boolovské.

$$Q = \{0, 1\} \tag{1.10}$$

**Definícia 1.13.** Funkcia $v$ je funkcia z množiny testovaných subjektov $S$ a množiny úloh $U$ do domény čiastkových výsledkov $Q$, podávajúca čiastkový výsledok subjekta na danej úlohe.

$$v : S \times U \to Q \tag{1.11}$$

Modely IRT popisujú subjektov pridelením vnútorných charakteristík, ktoré sú predmetom skúmania. V našom modeli to bude hodnosť v tvare reálneho čísla, predstavujúceho schopnosť riešiť úlohy, akými testujeme.

$$H = \mathbb{R} \tag{1.12}$$

**Definícia 1.14.** Funkcia $\theta$ je funkcia, ktorá každému subjektu priradí jeho hodnosť.

$$\theta : S \to H \tag{1.13}$$

Nasleduje základný pojem IRT. Funkcia nazývaná charakteristickou krivkou úlohy (Item Characteristic Curve), ktorá modeluje pravdepodobnosť vyriešenia úlohy vzhľadom na hodnosť.

**Definícia 1.15.** Funkcia $p$ je funkcia, ktorá každej úlohe pridelí pravdepodobnostnú funkciu označujúcu s akou pravdepodobnosťou bude úloha vyriešená pre danú hodnotu vnútornej charakteristiky.

$$p : U \times H \to (0, 1) \tag{1.14}$$

Funkcie $p$ a $\theta$ tvoria základ IRT modelov. IRT má za cieľ odhadnúť tieto dve funkcie tak, aby čo najlepšie sedeli na namerané hodnoty. Až potom je možné konať štatistickú analýzu.

Funkcie $p$ musí spĺňať určité požiadavky. Ako pravdepodobnostná funkcia musí byť monotónna, mať hodnoty v rozmedzí 0 až 1 a musí mať špeciálny tvar, určený konkrétnym IRT modelom.

Distribučné funkcie normálneho rozdelenia boli používané do konca 50. rokov, kedy boli nahradené logistickou funkciou (1.15) hlavne zo štatistických a praktických dôvodov.[vdLH96]

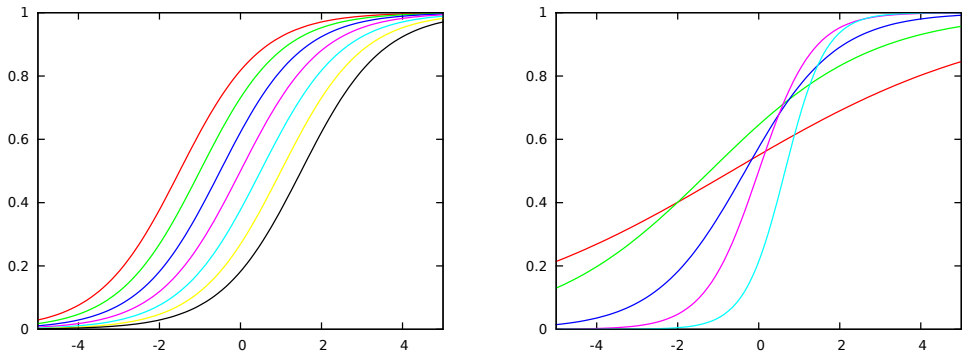
$$p(u, f) = \frac{1}{1 + e^{-b_u(f+a_u)}} \tag{1.15}$$

Nasádzanie IRT modelu na dáta nie je zrovna triviálna záležitosť.

Sú dva rôzne pohľady na tvorbu IRT modelov. Jeden názor je, že modely by sa mali prispôsobovať dátam, čo vedie k zložitým modelom, kde už začína byť otázne, či je daná interpretácia zrovna tá pravá. Druhý názor, vyzdvihovaný najmä v Rasch modeloch je ten, že by mali byť dostatočne jednoduché, s príjemnými štatistickými vlastnosťami ako obojstranná monotónnosť a ak model nesedí pre daný test, tak test bol zle zostavený.

Obojstranná monotónnosť je vlastnosť, že všetci subjekti generujú rovnaké usporiadanie ťažkosti úloh a zároveň všetky úlohy generujú rovnaké usporiadanie schopností subjektov. Znamená to, že by výber úloh nemal vplývať na usporiadanie subjektov a výber subjektov by nemal ovplyvňovať usporiadanie úloh (Obr. 1.1). V príklade je ukážka 1PL modelu, ktorý má konštantnú zakrívenosť, t.j. diskrimináciu a úlohy sa líšia iba svojou zložitosťou. Tento model je v podstate zhodný s pôvodným Rasch modelom.

Okrem usporiadania úloh a subjektov, IRT modely umožňujú aj zmysluplne navzájom porovnávať zložitosť úloh a vlastnosti subjektov na spoločnej škále. Zaujímavý parameter úloh, konkrétne odhadnutej $p$ funkcie je hodnota vnútornej vlastnosti, pri ktorej má subjekt 0.5 pravdepodobnosť ju vyriešiť. Vtedy sa subjekt a úloha považujú za rovnako hodných. CTT podobné porovnávanie neumožňuje.




(a) model s dvojitou monotónnosťou (b) model bez dvojitej monotónnosti

Obr. 1.1: Ukážka dvojitej monotónnosti

Keďže v súťažiach nie je testovaná iba jedna hypotéza, ale chceme získať informácie o každom súťažiacom a zároveň tvrdiť, o každom odhade vlastnosti subjektu, že je s určitou pravdepodobnosťou správny, potrebujeme dostatočnú prevahu nameraných údajov nad stupňami voľnosti, ktoré zavedie použitý IRT model.

Budú pre nás teda zaujímavé najmä jednoduché IRT modely, kde bude iba jedna vnútorná vlastnosť subjektov, pravepodobnostná funkcia $p$ úloh bude jednoduchá, definovaná pomocou čo najmenej parametrov, ale nie menej ako je potrebné a úlohy budú podľa možnosti hodnotené dichotomicky, teda buď ich subjekt vyrieši, alebo nie. Na takéto účely sa zdá byť najvhodnejšia trieda 1PL, 2PL a 3PL modelov.

### 1PL, 2PL a 3PL modely

Tieto modely sú základné logistické modely s jedným až troma parametrami. Presný tvar pravepodobnostných funkcií úloh je logistická funkcia (1.16) s tromi parametrami, prípadne dvomi, keď je $c_u$ konštantne 0 alebo jedným keď $b_u$ je 1.

$$p(u, f) = c_u + \frac{1 - c_u}{1 + e^{-b_u(f+a_u)}} \tag{1.16}$$

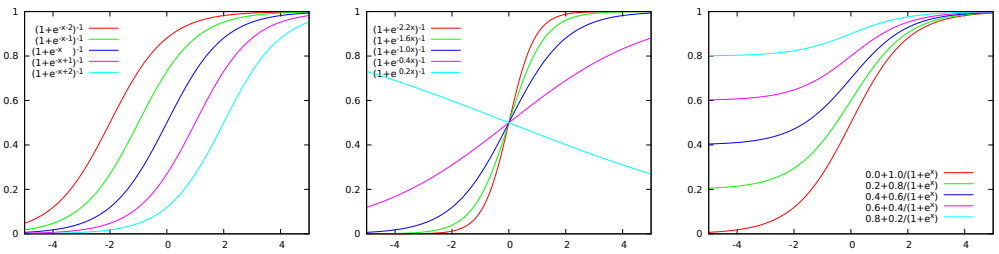
Model s jedným parametrom - 1PL pôvodne vytvoril Rasch v 50-tych rokoch hlavne pre niektoré špeciálne vlastnosti, ktoré iné modely nemajú.

Modely 2PL a 3PL vytvoril Birnbaum koncom 50-tych rokov, ale známymi sa stali až koncom 70-tych. Jeho práca bola založená na Lordových modeloch používajúcich normálne rozdelenie.[vdLH96]

Najzákladnejší parameter, ktorý sa vyskytuje vo všetkých troch modeloch je $a$. On udáva ťažkosť danej úlohy, konkrétne je hodnota meranej vlastnosti subjektov, pri ktorej je 0.5 pravdepodobnosť úspešného vyriešenia. V 1PL modeli sa teda môžu vyskytovať rôzne ťažké úlohy, ale aby model dobre sedel, musí schopnosť vyriešenia mať podobný tvar závislosti od testovanej vlastnosti. Príkladom je obrázok (1.2a).

Druhý parameter je $b$, ktorý označuje smer alebo zakrívenosť krivky (Obr. 1.2b). Tento parameter sa zvykne nazývať diskriminácia úlohy s možnou interpretáciou ako miera rozdielu úspešnosti riešenia vzhľadom na odchýlku vlastnosti subjektu od ťažkosti úlohy - $a$. Môže byť vhodné používať model aj s týmto parametrom napríklad, ak sú úlohy rôzneho charakteru alebo merajú vnútornú vlastnosť rôznymi spôsobmi a zároveň je dostatočné množstvo nameraných údajov.

Tretí parameter je užitočný v prípade, že je náchylnosť k náhodnému riešeniu úlohy, napríklad v prípade otázok s viacerými moznosťami. Prípadne hocikde inde, kde je určitá pravdepodobnosť, že úloha bude splnená bez ohľadu na meranú vlastnosť (Obr. 1.2c). Tento paramater sa podľa potreby môže určiť pred analýzou dát, prípadne sa môže odhadovať spolu s ostatnými, ak sa neberie ako predpoklad určitá náchylnosť riešenia bez znalostí (teda meranej vlastnosti).



(a) $(-2, -1, 0, 1, 2)$ (b) $(-0.2, 0.4, 1, 1.6, 2.2)$ (c) $(0, 0.2, 0.4, 0.6, 0.8)$

Obr. 1.2: Parametre PL modelov

### 2PL model

Nás bude prevažne zaujímať 2PL model, teda model s parametrami $a$ a $b$ funkcie odozvy 1.17.

$$p(u, f) = \frac{1}{1 + e^{-b_u(f+a_u)}} \tag{1.17}$$

**Definícia 1.16.** Funkcia $a$ je funkcia z množiny úloh do reálnych čísel, ktorá odhaduje pri akej hodnote testovaného vnútorného faktoru bude subjekt mať polovičnú pravdepodobnosť správne riešiť úlohu. Určuje teda stred krivky $p$.

$$a : U \to \mathbb{R} \tag{1.18}$$

V zložitých matematických výrazoch budeme písať $a_u$ miesto $a(u)$.

$$a_u \equiv a(u) \tag{1.19}$$

**Definícia 1.17.** Funkcia $b$ je funkcia z množiny úloh do reálnych čísel, ktorá určuje sklon funkcie odozvy $p$.

$$b : U \to \mathbb{R} \tag{1.20}$$

V zložitých matematických výrazoch budeme písať $b_u$ miesto $b(u)$.

$$b_u \equiv b(u) \tag{1.21}$$

Na analýzu dát 2PL modelom, potrebujeme model nasadiť na tie dáta. Konkrétne potrebujeme nájsť funkcie $a$ a $b$ (parametre úloh) a funkciu $\theta$ (vlastnosti subjektov), ktoré vhodne popisujú namerané dáta.

Vhodné môžu byť rôzne odhady týchto funkcií. Hodnoty funkcie $\theta$ je možné škálovať, posúvať a zmeniť im znamienko, s príslušnými zmenami na parametroch $a$ a $b$ bez zmien v ich výpovednosti. Pri odhadovaní hodnôt si teda zvolíme ľubovoľnú konkrétnu škálu. Navyše by sme chceli, aby vlastnosť pozitívne korelovala s výsledkami a budeme teda požadovať, aby väčšina

parametrov $b$ bola kladná. Ak pre nejakú úlohu $u$ parameter $b$ nie je kladný, teda funkcia $p(u)$ nie je rastúca, znamená, že subjekti s väčším odhadnutím faktorom majú tendenciu ju horšie riešiť.

Budeme hľadať funkcie $a$, $b$ a $\theta$, ktoré sú, až na horeuvedené ekvivalencie, odhadom s najväčšou pravdepodobnosťou. Na to si najprv definujeme pojem Fisherovej informácie.

Sir Ronald Aylmer Fisher roku 1925 definoval štatistický pojem informácie, ktorý sa dnes volá Fisherova informácia. Intiutívne ju možno chápať ako prevrátenú hodnotu presnosti, s akou je možné parameter odhadnúť.

Formálne, nech $X$ je náhodná premenná, $\theta$ je meraný parameter a $P(X, \theta) = P(X|\theta)$ je pravedpodobnostná funkcia premennej $X$ vzhľadom na $\theta$. Pre konkrétne namerané $X$ sa môžeme na $P$ pozerať ako na funkciu vierohodnoti $a$. Definujeme skóre ako deriváciu logaritmu tejto funkcie (1.22).

$$V(X, \theta) = \frac{\partial}{\partial \theta} \ln P(X, \theta) \tag{1.22}$$

Fisherovu informácia $\mathcal{I}$ je definovaná ako rozptyl skóre. Keďže očakávaná hodnota skóre je 0, Fisherova informácia bude štvorec očakávaného skóre. Pre všeobecný 3PL model platí (1.24).

$$E(V(X, \theta)) = \sum_{x \in X} \frac{\partial}{\partial \theta} \left[ P(x, \theta) \ln P(x, \theta) \right] = 0 \tag{1.23}$$

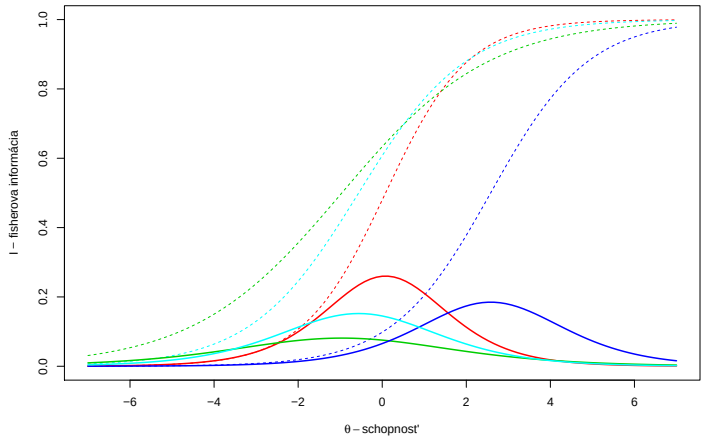
$$\begin{aligned} \mathcal{I}(a) &= D(V(X, a)) = E\left[ \left( \frac{\partial}{\partial \theta} \ln P(X, \theta) \right)^2 \right] \\ &= P(0, \theta) \left( \frac{\partial}{\partial \theta} \ln P(0, \theta) \right)^2 + P(1, \theta) \left( \frac{\partial}{\partial \theta} \ln P(1, \theta) \right)^2 \\ &= \frac{1}{P(0, \theta)} \left( \frac{\partial}{\partial \theta} P(0, \theta) \right)^2 + \frac{1}{P(1, \theta)} \left( \frac{\partial}{\partial \theta} P(1, \theta) \right)^2 \end{aligned} \tag{1.24}$$

Funkcia Fisherovej informácie nám hovorí, koľko informácie bude získané o odhadovanom parametri $a$ pri jednom meraní, ak má parameter danú pravú hodnotu. Keďže rozptyl nezávislých meraní je sčítateľný, aj Fisherova infor-

mácia je sčítateľná a Fisherova informácia množiny nezávislých úloh je suma informácií jednotlivých úloh.

Konkrétne pre 2PL model platí

$$\mathcal{I}(f) = a_u^2 \cdot p(u, f) \cdot (1 - p(u, f)) \tag{1.25}$$




Obr. 1.3: Ukážka Fisherovej informácie pre 2PL model

## 1.4 Nasádzanie IRT modelu

Model IRT pre konkrétne údaje je definovaný parametrami úloh a subjektov, ktoré dosahujú maximálnu vierohodnosť. Síce boli niektoré IRT modely známe už v 50-tych rokoch, používať sa začali omnoho neskoršie, keďže nájsť parametre dosahujúce maximálnu vierohodnosť je náročné.

Známe sú hlavne dva algoritmy cielené na nasádzanie modelov, ktoré používame. Joint Maximum Likelihood (JML) je jednoduchší algoritmus, ktorý spočíva v stredavom optimalizovaní parametrov subjektov a parametrov úloh. Výhodou je, že je jednoduchší a nemá apriori predpoklady rozloženia parametrov.

Marginal Maximum Likelihood (MML) je častejšie používaný algoritmus.

Jedna zo zásadných čŕt je, že odhaduje rozloženie schopností subjektov jeho apriórnym predpokladom. Má výhodu zaručenej konvergencie, kvôli čomu môže byť vhodnejší v prípade menších súborov dát. Pokiaľ je možné, o súťažiacich nechceme predpokladať rozloženie a preto sme implementovali JML.

### 1.4.1 Joint Maximum Likelihood algoritmus

Budeme predpokladať, že $R : (S, U, V)^*$ obsahuje namerané údaje. Parametre $a$, $b$, $c$ a $\theta$ modelu budeme zapisovať ako vektory reálnych čísel $\bar{a}$, $\bar{b}$, $\bar{c}$ a $\bar{\theta}$, indexované prislúchajúcimi úlohami a súťažiacimi.

Na maximalizáciu vierohodnosti potrebujeme nájsť vektory, kde funkcia (1.26) nadobúda maximum. Algoritmus Joint Maximum Likelihood tento problém optimalizácie funkcie s $|S| + 3|U|$ premenných rozbíja na striedavú optimalizáciu $|S|$ funkcií s jednou premennou a $|U|$ funkcií s tromi premennými.

$$L(\bar{\theta}, \bar{a}, \bar{b}, \bar{c}) = \prod_{(s,u,v)\in R} P(v, \bar{\theta}_s, \bar{a}_u, \bar{b}_u, \bar{c}_u) \tag{1.26}$$

Ak do funkcii $L$ dosadíme konštanty ako všetky argumenty až na tie prislúchajúce konkrétnej úlohe resp. súťažiacemu a zanedbáme činitele súčinu, ktoré sa týmto stanú konštantami, dostaneme sady funkcií $L_u$ resp. $L_s$. Každá úloha a každý súťažiaci bude mať svoju funkciu vierohodnosti jeho parametrov.

$$(\forall u \in u)\ L_u(a, b, c) = \prod_{\substack{(s,u',v)\in R \\ u'=u}} P(v, \bar{\theta}_s, a, b, c) \tag{1.27}$$

$$(\forall s \in S)\ L_s(\theta) = \prod_{\substack{(s',u,v)\in R \\ s'=s}} P(v, \theta, \bar{a}_u, \bar{b}_u, \bar{c}_u) \tag{1.28}$$

Ak sú dosadenými konštantami parametre, kde funkcia $L$ nadobúda maximum, zo spôsobu odvodenia funkcií $L_u$ a $L_s$ je zrejmé, že aj ony budú

nadobúdať maximum v rovnakom bode. Podobne aj opačne, ak v nejakom bode všetky funkcie $L_u$ a $L_s$ nadobúdajú maximum, je to maximum aj funkcie $L$.

Tieto funkcie majú výhodu, že sa im ľahšie hľadá maximum, problém sú však neznáme konštanty, ktoré do nich potrebujeme dosadiť. Všimnime si, že funkcie úloh závisia iba na konštantách súťažiacich a opačne. Problém vyriešime začínaním v nejakom začiatočnom odhade parametrov a striedavým optimalizovaním parametrov úloh a súťažiacich, vždycky používajúc parametre nájdené v predchádzajúcom kole ako konštanty, pokiaľ sa parametre neustália.

Ak algoritmus skončí, nájde nejaké maximum funkcie $L$. Nie je úplne zrejmé, na akých vstupoch algoritmus skončí [vdLH96]. V praxi sa však ukazuje ako funčný, keď údaje dostatočne jasne určujú model.

Algoritmus JML nedefinuje spôsob voľby začiatočných parametrov a maximalizácie unárnych a trinárnych funkcií. V našej implementácii sme vyskúšali niekoľko možností.

### 1.4.2 Marginal Maximum Likelihood algoritmus

Pôvodný algoritmus Marginal Maximum Likelihood (MML) zverejnili Bock a Lieberman v roku 1970. Popularitu získal najmä kvôli neistým vlastnostiam JML algoritmu a stal sa štandardným algoritmom na odhad 2PL a 3PL modelov. Mal exponenciálnu zložitosť vzhľadom na počet úloh, ale vylepšili ho Bock a Aitkin roku 1981 pridaním Expectation Maximization (EM) algoritmu. [vdLH96]

S ohľadom, že je dostupná implementácia tohto algoritmu v Hansonovej knižnici ETIRM, nebudeme sa venovať detailom tohto algoritmu. Pre podrobnejší popis si záujemci môžu pozrieť pôvodný Bockov a Aitkinov článok [BA81], prípadne článok autora knižnice [WH97]. Podstatné však pre nás je, že tento algoritmus používa apriori predpoklady o rozložení meraných vlastností subjektov.

Na nasádzanie zložitejších modelov je známe aj vylepšenie MML s náz-

vom Adaptive Quadrature Expectation Maximization (ADQEM) popísaný Schillingom a Bockom a nový algoritmus Li Cai-a zvaný Metropolis-Hastings Robbins-Monro (MH-RM). Algoritmy boli autormi popísané v Psychometrike roku 2005 resp. 2010. Implementáciu oboch algoritmov možno nájsť napr. v programe IRTPro. Ako motivácia zavedenia týchto algoritmov je uvádzaná faktorová analýza, kde sú súčasne modelované viaceré vlastnosti subjektov. Keďže nepoužívame tieto modely, algoritmy sme zhodnotili ako príliš a zbytočne zložité na naše účely.

# Kapitola 2

# Teória

V tejto kapitole najprv v časti 2.1 definujeme jeden možný model IRT ratingového algoritmu inšpirovaný Foriškom [For09a]. Uvádzame ho však najmä na ilustráciu.

Od podkapitoly 2.2 Odhad presnosti, začneme prezentovať vlastný výskum. Rozoberieme zaužívané spôsobov počítania presnosti odhadov schopností jednotlivých súťažiacih. Rozbor je plod výskumu autora a prínosom svojím novým spôsobom pohľadu na odhady presnosti a požiadavky na ne. Sformulujeme tu vlastnosti, ktoré by sme od dobrých odhadov presnosti očakávali, ale v rámci bežných CTT metód neboli dosiahnuteľné.

Ďalšie dve podkapitoly venujeme našim dvom návrhom na odhadovanie presnosti merania, kde preskúmame ich vlastnosti a ukážeme, že sú výrazným priblížením k splneniu požiadaviek sformulovaných v podkapitole 2.2. Konkrétne bayesovský odhad bude spĺňať všetky požiadavky, až na všeobecnú monotónnosť hodnotenia, ktorú zaručíme iba za istých predpokladov, a výpovednosti, ktorú je najvhodnejšie skúmať empiricky.

## 2.1 Ratingový algoritmus

Definujeme model ratingového algoritmu. Bude to jednoduchý model, ktorý má jedno kolo. Tento model je možné rozšíriť na zložitejší s viacero kôl,

aký používal Forišek [For09a]. Náš výskum je však platný v oboch a preto použijeme jednoduchší.

Nech sú množiny $S$, $U$ a $Q$ množiny súťažiacich, úloh a výsledkov, ako bolo definované v kapitole 1, potom IRT ratingový algoritmus je funkcia na množine

$$(S, U, Q)^* \to ((S \to H), (U \to (\mathbb{R}^n))) \tag{2.1}$$

kde $n$ je 1, 2 resp. 3, pre 1PL, 2PL resp. 3PL IRT model.

## 2.2 Odhad presnosti

Začneme úvodom do cieľov použitia a možných spôsobov počítania odhadov presnosti v oblasti súťaží prehľadom rozvoja ich využitia. Poukážeme tým na smerovanie vývoja tejto oblasti a navrhneme požiadavky na dobrý odhad presnosti. Následne ukážeme dva možné odhady rôznych vlastností, ktoré sa približujú týmto požiadavkám.

V prípade klasických testov vo forme CTT, sa autor nestretol s vyjadrovaním presnosti merania jednotlivých súťažiacich. Štandardnou požiadavkou v športe je minimálny bodový rozdiel súťažiacich pre potreby porovnávania, ako napríklad vo volejbale a tenise, ale výrazne ďalej to nezachádza. Ako možný dôvod vidíme štadardný predpoklad CTT o normálnom a rovnakom rozdelení chýb merania súťažiacich.

Arpad Elo v algoritme na šachový rating používa jedno číslo na ohodnotenie šachistu, javia sa však náznaky zohľadňovania presnosti merania v tvare K-faktoru. Tento K-faktor sa prejavuje ako citlivosť hodnosti na nové súťaže, pričom je určovaný na základe hodnosti oboch hráčov a počtu ich doterajších hodnotení.

Významný krok v modelovaní presnosti merania urobil Mark Glickman, keď na základe Elo algoritmu vytvoril najprv Glicko a následne Glicko-2 [Gli12] algoritmus. Tieto algoritmy pomocou dodatočnej jednej resp. dvoch veličín charakterizujú stabilitu a presnosť hodnosti hráča.

Elo a Glicko algoritmy sú používané nielen v šachu, ale aj širšie, hodnotia

však jedine súboje dvoch súťažiacich. Zaujímavým rozšírením Glickmanovho algoritmu je algoritmus TrueSkill [HMG07], ktorý hodnotí tímy hráčov. Vyvinuli ho v Microsofte, uvádzajúc ako motiváciu výber vhodných súperov a tvorenie tímov súťažiacich v online hrách.

Spomínané algoritmy sú stavané na hodnotenie priamych súbojov súťažiacich a teda nie sú vhodné pre naše súťaže. V oblasti súťaží jednotlivcov proti systému je informatikom známy predstaviteľ TopCoder, ktorého hodnotenie súťažiacich tiež obsahuje dve zložky. Mierne podobný bayesovským vzorcom, TopCoder počíta hodnosť (rank) a mieru jej neurčitosti (volatility). Síce užitočný, TopCoder štatisticky nie je dobre podložený. Forišek ukazuje závažné nedostatky ako napríklad nemonotónnosť v [For09a].

Ďalší známy spôsob hodnotenia presnosti odhadu je cez očakávanú mieru získanej informácie. Forišek týmto prístupom, v článku o ratingu pomocou Item Response Theory [For09b], definuje smerodajnú odchýlku merania (standard error of measurement) rovnicou (2.2), kde $\hat{\theta}$ je odhad schopnosti súťažiaceho a $\mathcal{I}$ je funkcia udávajúca mieru Fisherovej informácie. Tento prístup používal Forišek v práci [For09a] pri porovnaní IRT ratingu s TopCoder algoritmom.

$$SEM(\hat{\theta}) = \sqrt{\frac{1}{\mathcal{I}(\hat{\theta})}} \tag{2.2}$$

Teraz uvedieme jeden príklad, kde poukážeme na teoretické nedostatky odhadu Fisherovou informáciou.

**Príklad 2.1.** Majme troch súťažiacich $S$, dve úlohy $U$ a výsledky $R$.

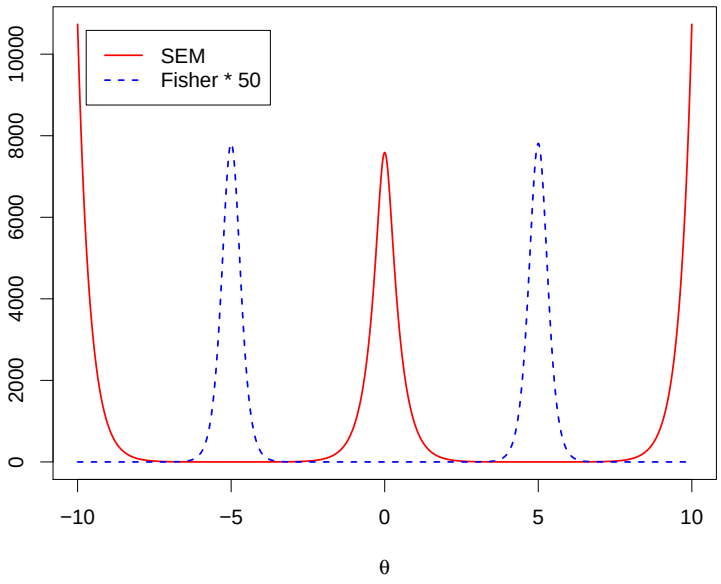
$$S = \{s_1, s_2, s_3\}$$
$$U = \{u_1, u_2\}$$
$$R = \left[\begin{array}{c|cc} & u_1 & u_2 \\ \hline s_1 & 0 & 0 \\ s_2 & 0 & 1 \\ s_3 & 1 & 1 \end{array}\right]$$

Foriškov ratingový algoritmus pre tieto údaje dáva nasledovné výsledky:

| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
|---|---|---|---|
| $u_1$ | 5 | $-5$ | 0 |
| $u_2$ | 5 | 5 | 0 |

| | $\hat{\theta}$ | $S\hat{E}M$ |
|---|---|---|
| $s_1$ | $-10$ | 10733.49 |
| $s_2$ | 0 | 7589.72 |
| $s_3$ | 10 | 10733.49 |

Súťažiaci sú rovnomerne rozmiestnený na povolenom intervale, úlohy sú v strede medzi nimi a majú maximálnu diskriminačnú schopnosť. Fisherova informácia týchto úloh je znázornená na obr. 2.1. Možno ich chápať ako ostré meradlo, ktoré pri meraní skoro určite povie, či má súťažiaci väčšiu alebo menšiu hodnosť ako $-5$ resp. 5. Podľa modelu súťažiaci majú schopnosti v príslušných intervaloch $(-\infty, -5)$, $(-5, 5)$ a $(5, \infty)$. Síce súťažiaci $s_2$ je určený na interval dĺžky 10, odchýlka tohto merania počítaná Fisherovou informáciou je vyše 7000, asi 100 krát viacej, než by mala byť.




Obr. 2.1: Znázornenie Fisherovej informácie a odhadu chyby pre príklad 2.1

K príkladu 2.1 sa ešte vrátime, keď budeme rozoberať jednotlivé vlastnosti odhadov presnosti. Najprv ujasníme, čo by sme od odhadov presnosti chceli

a akými spôsobmi to môžeme dosiahnuť. Odhady presnosti majú hlavne dve využitia, ktoré by sme v našom riešení chceli podchytiť:

**spoľahlivosť merania**

Poznanie presnosti merania nám umožňuje zhodnotiť či a do akej miery sme schopnosti súťažiacich namerali. Použitie zložitého štatistického nástroja bez možnosti zhodnotenia jeho spoľahlivosti na nameraných údajoch je nezodpovedné. Dostupnosť odhadov chyby môže pomôcť vyžitiu IRT v súťažiach.

**dolné odhady**

Odhadovanie presnosti hodnotenia jednotlivých hráčov prináša možnosť robiť pesimistické, prípadne optimistické odhady ich schopností. Dolný odhad hráčov pre potreby zoradenia používa TrueSkill. Rating hráča je sa počíta formulou (2.3), kde $\mu$ je odhad schopnosti a $\sigma$ je smerodajná odchýlka.

$$l(\mu, \sigma) = \mu - 3\sigma \tag{2.3}$$

Tento odhad môže byť férovejší tým, že od hráča požaduje, aby sa dokázal ako naozaj dobrý, na čo mu nestačí zopár náhodných výhier. Pri naivnom použití na IRT sa však môže oplatiť aj menší výkon.

Používané spôsoby odhadu presnosti sa líšia medzi rôznymi systémami. Zhrnieme ich do troch zásadne odlišných prístupov, na ktoré sa budeme odkazovať pri analýze dostupných a návrhu vlastného riešenia. Na vysvetlenie budeme predpokladať, že $\theta$ je meraná schopnosť. Túto schopnosť budeme modelovať ako neznámu náhodnú premennú $\Theta$, aposteriórne z pohľadu po hodnotení, ktorého výstupom bola $\hat{\theta}$.

**smerodajná odchýlka**

Väčšina spomínaných algoritmov vyjadruje chybu jedným číslom, označíme ho $\hat{\sigma}$, ktoré je mierou neistoty výstupného hodnotenia. V bayesovských systémoch sa od tejto hodnoty očakáva, že je odhadom smerodajnej odchýlky $\Theta$ od $\hat{\theta}$, teda že platí (2.4). Ak by chyba mala normálne

rozdelenie, tak by slušným odhadom pre $\Theta$ bolo rozdelenie $\hat{\Theta}$ (2.5).

$$\sigma^2 = E((X - \hat{\theta})^2) \approx D(X) \tag{2.4}$$

$$\hat{\Theta} \sim \mathcal{N}(\hat{\theta}, \hat{\sigma}) \tag{2.5}$$

Systémy, ktoré sa nesnažia odhadovať smerodajnú odchýlku, ale majú podobné vlastnosti, ako napr. TopCoder, zaradíme do tejto kategórie, keďže vlastnosti, ktoré nás zaujímajú má podobné.

**interval spoľahlivosti**

Presnosť odhadu je možné vyjadriť ako interval spoľahlivosti, teda interval v ktorom leží skutočná hodnota meranej vlastnosti $\theta$ s určitou pravdepodobnosťou $p$, za platnosti predpokladov ratingového systému (2.7). Tento prístup má prednosť najmä v tom, že interval nemusí byť symetrický okolo $\hat{\theta}$ a môže byť aj poloohraničený. Častým javom je, že je súťažiaci lepšie zmeraný z jednej, než druhej strany. Stáva sa to najmä v prípade súťažiacich, ktorí sú veľmi dobrí, prípadne zlí. Týchto súťažiacich možno z jednej strany dobre ohraničiť hranicou spoľahlivosti skúšky, ale druhá je pritom neistá. Smerodajná odchýlka toto nevie dobre vyjadriť.

$$p = 0.95 \tag{2.6}$$

$$P(\theta \in (a, b)|\Theta) > p \tag{2.7}$$

Síce ratingové systémy Glicko a TrueSkill vyjadrujú výsledky ako interval spoľahlivosti, resp. jeho dolnú hranicu, počítajú ho zo smerodajnej odchýlky za predpokladu normálneho rozdelenia. Ide teda iba o spôsob prezentácie.

**distribučná funkcia**

Pokiaľ sme schopný odhadnúť o rozložení $\Theta$ viacej informácie ako podávajú minulé dve riešenia, napríklad ak vieme odhadnúť jej distribučnú funkciu, môže byť vhodným riešením posunúť celú túto informáciu na

spracovanie užívateľom, napríklad v tvare grafu hustoty pravdepodobnosti (2.8).

$$f(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} P(\Theta = \theta) \tag{2.8}$$

Nepoznáme prípady vyjadrenia ratingu ako distribučnej funkcie resp. jej znázornenia pomocou grafu, nepredpokladáme však, že neexistujú. Spomínané ratingové systémy predpokladajú normálne rozdelelenie chyby. Ich grafické zobrazenie by mohla byť zaujímavá záležitosť, ale všetka informácia je už obsiahnutá v smerodajnej odchýlke.

Bdelý čitateľ si mohol všimnúť, že všetky doteraz spomínané ratingové systémy sú zaradené do prvej skupiny. Máme dva dôvody na uvedenie skupín s intervalom spoľahlivosti a distribučnou funkciou. Jeden je, že sú to zaužívané postupy v štatistike, ako aj vyhodnocovaní niektorých testov a druhý, že chceme vytvoriť takéto odhady vo vlastnom systéme.

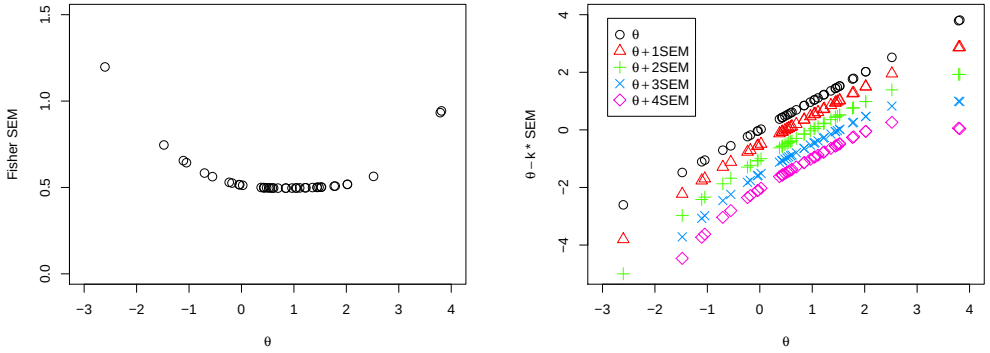
Vráťme sa k odhadu presnosti pomocou Fisherovej informácie. Uvedieme tri vlastnosti odhadu chyby mierou očakávanej informácie v zmysle rovnosti (2.2), ktoré považujeme za zdroje nepresností a chceli by sme odstrániť.

**bodový podľa $\hat{\theta}$**

Odhad presnosti v zmysle (2.2) sa počíta jedine na základe hodnoty informačnej funkcie v bode $\hat{\theta}$. Takýto odhad predpokladá, že informačná funkcia vo významnom okolí $\theta$ je približne konštantná (2.9). Na rozdiel od príkladu 2.1, tento predpoklad býva v bežných testoch splnený na rozsahu spoľahlivosti testu a významným je prínosom, ak boli súťažiaci hodnotený na rôznych množinách úloh.

$$SEM(\theta) \approx SEM(\hat{\theta}) \tag{2.9}$$

Problém však môže nastať, keď je schopnosť súťažiaceho pri konci intervalu spoľahlivosti množiny úloh, ktoré riešil. V tomto prípade je kvalita




(a) Odhad smerodajnej odchýlky merania vzhľadom na pridelenú schopnosť

(b) Dolný odhad schopnosti vzhľadom na pôvodný odhad schopnosti

Obr. 2.2: Odhad chyby merania Fisherovou informáciou na teste z princípov počítačov 2012-01-12 modelované 3PL modelom, kde neodpovedanie je považované za nesprávnu odpoveď.

odhadu otázna a zmeny v odhade schopnosti $\hat{\theta}$ výrazne ovplyvňujú odhad presnosti.

**nemonotónny dolný odhad**

Funkcia Fisherovej informácie v IRT modeli nie je konštantná, pri konci merného rozsahu testu sa výrazne znižuje. Nemožno teda priamo používať konzervatívne odhady ako má TrueSkill (2.3), lebo aj pri malých konštantách $k$ môže byť hodnotenie nemonotónne. Teda môže sa súťažiacemu $s_1$ oplatiť neurobiť úlohu, aby jeho odhad $\hat{\theta}_1$ mal lepší odhad presnosti $SEM(\hat{\theta}_1)$ a teda mal väčší konzervatívny odhad (2.10).

$$\begin{gathered}(\forall u)\ v(s_1,u) \leq v(s_2,u) \\ (\exists u)\ v(s_1,u) < v(s_2,u) \\ \hat{\theta}_1 < \hat{\theta}_2 \\ \hat{\theta}_1 - k \cdot SEM(\hat{\theta}_1) > \hat{\theta}_2 - k \cdot SEM(\hat{\theta}_2)\end{gathered} \tag{2.10}$$

Táto nerovnosť by sa vyskytla v príklade 2.1 pre ľubovoľné $k > 0.004$. Pre menej vykonštruovaný príklad, uvádzame výsledky testu z princípov počítačov zobrazené na obrázku 2.2b. Dolný odhad už pre hodnotu

$k = 3$ vymení poradie najlepších dvoch študentov. V tomto prípade síce monotónnosť ešte nie je porušená, lebo študenti sú úlohovo neporovnateľný, ale algoritmus Fisherovho odhadu to nevie.

Nemonotónnosť pôvodí z toho, že bodový odhad odhaduje chybu symetricky (2.11), aj keď taká nie je. Totiž odhad chyby počítaný z celej informačnej funkcie by bol monotónny.

$$SEM(\theta - \delta) \not\approx SEM(\theta + \delta) \tag{2.11}$$

**apriórny podľa $\theta$**

Fisherova informácia je definovaný ako očakávané množstvo získanej informácie na teste, pre súťažiaceho hodnosťou $\theta$. Pri tomto výpočte sa nezohľadňujú výsledky na úlohách, až na $\hat{\theta}$, ktorým sa odhaduje $\theta$.

Nech sú dvaja súťažiaci $s_1$ a $s_2$ skúšaný na rovnakých úlohách, pričom prvý, $s_1$, vyrieši ľahké a nevyrieši ťažké úlohy, ale druhý, $s_2$ má problém s ľahkými, a ťažké vyrieši. Ak dostanú rovnaký odhad schopnosti $\hat{\theta}_1 = \hat{\theta}_2$, tak budú mať aj rovnaký odhad chyby $SEM(\hat{\theta}_1) = SEM(\hat{\theta}_2)$. Tento stav nie je žiadúci, lebo odhad schopnosti $\hat{\theta}_2$ druhého súťažiaceho $s_2$ môže byť iba náhoda, keď výsledok jednej úlohy pre odhad znamená mnoho.

Z apriórneho pohľadu je situácia so súťažiacim $s_2$ podľa modelu nepravdepodobná, čakáme, že súťažiaceho so schopnosťou $\hat{\theta}_2$ dobre zmeriame, ale aposteriórne už vieme povedať, že nastal ten prípad, keď sme ho nezmerali.

**jedno číslo ako výstup**

Spomínané spôsoby odhadu chyby mali za výstup jedno číslo. Prednosťou je, že sa s ním ľahšie pracuje, ale týmto zjednodušením sa stráca informácia.

Problémom je nemožnosť vyjadriť asymetrickosť, a teda ani poloohraničené intervaly spoľahlivosti, ktoré sa vyskytujú v prípade subjektov

mimo merného rozsahu používaných úloh.

Je otázne, či je možné vytvoriť monotónny dolný odhad schopnosti na základe iného ako výlučne jednostranného odhadu chyby.

Spomínané klasické ratingové systémy modelujú hodnotenie odhadom jeho rozdelenia, ktorý je vyjadrený dvomi (Glicko, TrueSkill) resp. tromi (Glicko-2) parametrami. Pri hodnotení sa z predchádzajúceho odhadu a výsledkov získa nový odhad. O týchto algoritmoch predpokladáme, že nemajú problém s monotónnosťou dolného odhadu, kvôli spôsobu modelovania chyby. Predpokladajú však symetrické, normálne rozdelenie odhadu schopnosti, ktoré v našom modeli nepovažujeme za vhodné. Charakterizovať schopnosť súťažiaceho, ktorý vyriešil všetky úlohy normálnym rozdelením je zaujímavá záležitosť.

Pri ratingu súťažiacich na úlohách, výhodou IRT je možnosť zohľadniť špecifiká jednotlivých úloh. Totiž v CTT sú všetky úlohy považované za rovnaké. Práve túto skutočnosť využíva odhad chyby na základe informačnej funkcie, ako je Foriškov. Poznanie vlastností úloh a výsledkov súťažiaceho na nich považujeme za dostatočné na odhad chyby hodnotenia súťažiaceho a nebudeme ďalej odhadovať chyby úloh. Možnosti zohľadnenia nepresností odhadu úloh na odhad schopností skúmali Tsutakawa a Johnsonová [TJ90], keď počítali bayesovské odhady vplyvu neistoty parametrov na hodnotenie cez bodový odhad informačnej matice. Ich odhad však bol symetrický a počítal smerodajnú odchýlku.

Naším cieľom je vytvoriť odhad schopnosti pre IRT ratingový systém, ktorý bude umožňovať zhodnotiť spoľahlivosť merania súťažiacich a robiť dolné odhady schopností. Od týchto odhadov budeme požadovať, aby zvládali asymetrické intervaly spoľahlivosti a pokiaľ je to možné, vhodné bude aj vyjadrenie odhadu aposteriórnej distribúcie hodnotenia pre súťažiacich. Odhad by teda nemal byť bodový, aby sme mohli zaručiť monotónnosť a robiť asymetrické odhady. Navrhovaný odhad bude aposteriórny, zohľadňujúci konkrétne výsledky súťažiaceho a mieru ich podivnosti.

Ukážeme dva prístupy odhadu chyby. Jeden je založený na bootstrap spôsobe odhadovania neznámeho rozdelenia a druhý bayesovským prístupom zhodnocuje aposteriórne pravdepodobnostné rozdelenie schopnosti súťažiaceho. Oba aposteriórne odhadujú rozdelenie schopnosti, nie sú však ekvivalentné.

## 2.3 Bootstrap odhad presnosti

Bootstrap je zaužívaný spôsob odhadu spoľahlivosti merania z malého množstva údajov, nepredpokladajúc pri tom konktrétne rozdelenie. Používa sa najmä na odhad rozptylu a tvorenie intervalov spoľahlivosti. My využijeme tento postup na vytvorenie aposteriórneho odhadu rozptylu a intervalu spoľahlivosti pre hodnosti $\theta_i$ súťažiacich.

Najprv neformálne predstavíme bootstrap pomocou analógie a potom formálne zadefinujeme štýl bootstrapu, ktorý budeme používať.

Ak by sme mali vrece orechov a zaujímala by nás priemerná hmotnosť orechu, mohli z nej náhodne vybrať 30 kusov, spočítať ich priemernú hmotnosť a odhadnúť tým priemernú hmotnosť orechov v celom vreci. Tento údaj sa nazýva prvý moment a počíta sa pomerne jednoducho. Zo spomínaných 30 orechov vieme niečo usúdiť aj o presnosti tohto odhadu. Keď sú tieto orechy podobnej hmotnosti, asi bude odhad spoľahlivejší, než keď sú úplne divoké. Problém však je toto pozorovanie sformalizovať a číselne vyjadriť. Ak by sme napríklad vedeli, že hmotnosti orechov majú normálnu distribúciu, mohli by sme spočítať rozptyl týchto orechov, z čoho by sme spočítali formálny odhad rozloženia orechov vo vreci a následne aj očakávanej spoľahlivosti merania. Keďže nevieme tvar rozloženia hmotností orechov vo vreci, alebo ho nechceme predpokladať, privoláme si na pomoc bootstrap metódu.

Všimnime si, že získané (empirické) rozloženie našich 30-tich orechov sa do určitej miery podobá na nám neznáme rozloženie orechov vo vreci. Presnosť merania odhadneme spočítaním rozptylu, nebudeme však počítať rozptyl hmotností našich 30-tich orechov, tým by sme získali odhad na rozptyl

hmotností orechov vo vreci. Nás zaujíma teoretický rozptyl merania, teda priemernej hmotnosti náhodne vybraných 30-tich orechov z vreca. Kto nemá v hlave, má v pätách, vyskúšame urobiť veľa meraní. Keďže môžeme odhnúť rozloženie orechov vo vreci rozložením našich dostupných 30-tich orechov, nebudeme ťahať z vreca, ale z týchto 30-tich. Samozrejme, vo vreci je orechov veľa a preto musíme ťahať s opakovaním. Týmto spôsobom sme schopný vytvoriť mnoho sád orechov, ktoré sú síce mierne zdegenerované, ale podobajú sa rozloženiu možných výberov 30-tich orechov z vreca. Zbavili sme sa potreby myslieť a môžeme priamo spočítať rozptyl priemernej hmotnosti na týchto vytvorených sadách, počítaním na všetkých možných $30^{30}$ sád, prípadne na nejakej dostatočne veľkej náhodnej vzorke. Tento rozptyl na simulovaných sadách je potom dobrým odhadom na rozptyl všetkých možných meraní na vreci orechov.

Bootstrap metóda nám žiadne nové poznatky priniesť nemôže, umožňuje nám však odhadnúť ako sa správa nejaká funkcia, ako napríklad priemer, na údajoch podobného rozloženia, ako sú dostupné údaje, bez potreby jeho formalizácie.

Známych je mnoho rôznych bootstrapových metód. Uvedieme formálnejšiu definíciu neparametrického bootstrapu, ktorý je hore popísaný a z ktorého vytvoríme odhad presnosti. Keďže na vyjadrenie predpokladov je potrebná zložitejšia štatististická teória, predpoklady nebudeme dopodrobna špecifikovať, iba ich spomenieme. Pre podrobnejší popis odkazujeme čitateľa na knihu Casellu a Bergera [CB02].

**Definícia 2.1.** Nech $X$ je náhodná premenná a vektor $\bar{x} = x_1, \ldots, x_n$ je $n$ nezávislých realizácií tejto premennej. Nech $f : X^n \to \mathbb{R}$ je funkcia na definičnom obore hodnôt týchto premenných. Nech $\bar{x}_1^*, \ldots, \bar{x}_m^*$ je $m$ vektorov dĺžky $n$, ktoré vznikli rovnomerným náhodným výberom s opakovaním prvkov vektora $\bar{x}$.

Pokiaľ funkcia $f$ spĺňa určité vlastnosti normality a distribúcia $X$ je v istom zmysle slušná, potom platí (2.12) a (2.13), pričom $Var$ je rozptyl nameraných hodnôt a $Q_z$ je kvantilová funkcia určujúca najmenšiu hodnotu

väčšiu než nadobudne $z$-ta čiastka prípadov.

$$(\forall z)\ Q_z^*(f(\bar{x}^*)) \approx Q_z(f(\bar{x})) \tag{2.12}$$

$$Q_z^*(f(\bar{x}^*)) = min\Big\{y \ \Big|\ \big|\{i \,|\, f(\bar{x}_i^*) \leq y\}\big| \ < \ z \cdot m\Big\}$$

$$Var^*(f(\bar{x}^*)) \approx Var(f(\bar{x})) \tag{2.13}$$

$$Var^*(f(\bar{x}^*)) = \frac{1}{m-1}\sum_{j=1}^{m}\Big(f(\bar{x}_j^*) - \overline{f(\bar{x}^*)}\Big)^2$$

$$\overline{f(\bar{x}^*)} = \frac{1}{m}\sum_{j=1}^{m}\big(f(\bar{x}_j^*)\big)$$

Špeciálne platí, že odhad maximálnej vierohodnosti spĺňa podmienky kladené na funkciu $f$, ako aj medián a rôzne iné štatistiky.

Navrhujeme teda používať bootstrap metódu na odhadovanie presnosti vypočítaných schopností súťažiacich, pričom presnosť odhadu každého súťažiaceho budeme počítať zvlášť. Pre jednoduchosť budeme ďalej rozoberať jedného súťažiaceho $s$.

V zmysle definície 2.1, obor hodnôt náhodnej premenej $X$ je množina usporiadaných dvojíc úloh a úspechov súťažiaceho na nich. Známou realizáciou $\bar{x}$ tejto náhodnej premennej sú všetky známe výsledky $\bar{v}$ súťažiaceho . Odhadovanou funkciou $f$ je odhad maximálnej vierohodnosti, ktorým bežne počítame schopnosti súťažiacich. Treba poznamenať, že opakované výskyty výsledku na rovnakej úlohe budeme považovať za osobitné merania pre potreby počítania pravdepodobnosti. V opačnom prípade by zanikli a výpočet by sa podobal znáhodneniu jackknife metódy.

$$\bar{v} \ : \ (U, B)^*$$

$$\bar{x} = \bar{v}$$

$$f(\bar{v}) = MLE(\theta|\bar{v})$$

Týmto spĺňame predpoklady bootstrap metódy (2.1) a môžeme ju používať na odhadovanie rozptylu a kvantilov odhadu schopnosti $\hat{\theta}$.

Teraz už môžeme definovať odhad chyby, smerodajnú odchýlku $\sigma$, ako odmocninu rozptylu (2.14). Tento odhad však budeme používať najmä na porovnanie s inými algoritmami.

$$\begin{aligned} \bar{\theta}^* &= f(\bar{x}^*) \\ \sigma^2 &= Var^*(\bar{\theta}^*) \end{aligned} \tag{2.14}$$

**Definícia 2.2.** Bootstrap odhad presnosti súťažiaceho $s$ miery pravdepodobnosti $p$ je interval $(d_s, h_s)$, ktorý obsahuje strednú $p$-tu čiastku bootstrapovaných odhadov.

$$d_s = Q_{(1-p)/2}(\bar{\theta}_s^*) \tag{2.15}$$

$$h_s = Q_{(1+p)/2}(\bar{\theta}_s^*) \tag{2.16}$$

Štandardná voľba miery $p$ je napríklad 0.95.

$$p = 0.95$$

Ďalšou možnosťou je prenásobenie bootstrapovaných výskytov $\theta_i^*$ vhodne úzkym normálnym rozdelením a ich sčítanie, čím sa získa uhladený graf vhodný na zobrazenie. Zameriame sa však na intervalový odhad.

Zmienime sa o voľných parametroch v nášom algoritme na intervalové odhady. Parameter $m$ zodpovedá počtu kôl Monte Carlo simulácií. V závislosti od požadovanej presnosti, pre bootstrap je odporúčané vyše 5000 kôl. V prípade využitia na iné ako výskumné účely, pre stabilné výsledky odporúčame desiatky alebo až stovky tisíc kôl. Naša implementácia na testoch z princípov počítačov urobila 100 000 bootstrapov približne za minútu, dala by sa však podstatne zrýchliť.

Parameter $p$ určuje požadovanú očakávanú pravdepodobnosť, že sa meraná schopnosť súťažiaceho nachádza vo výslednom intervale. Veľkosť $p$ má vplyv na potrebný počet kôl na zaručenie rovnakej stability odhadu. Z tohto dôvodu, ako aj z predpokladaných zväčšujúcich sa odchýlok pri konci bo-

otstrapového rozloženia schopností, od ozajstného rozloženia schopností, neodporúčame tento parameter nastavovať príliš veľký. Vhodný hodnota však závisí od konkrétnej situácie.

Naša bootstrapová metóda odhaduje očakávané rozloženie schopnosti súťažiaceho, s predpokladom, že úlohy na meranie schopností majú vlastnosti a rozdelenie podobné vypočítanému odhadu úloh a súťažiaci na úlohy odpovedá v sklade s nameranými údajmi. Teda nepredpokladá, že sú úlohy správne namerané, predpokladá však, že existujú iba také úlohy, ako boli namerané, v podobnom rozložení ako ich riešil súťažiaci, pričom úspech súťažiaceho je neoddeliteľne spojený s úlohou ktorú riešil.

Bootstrapová metóda teda nie je bodová a umožňuje robiť asymetrické intervalové odhady, teda aj dolné odhady a využíva aposteriórne poznanie výsledkov súťažiaceho na jednotlivých úlohách. Monotónnosť, vo všeobecnosti žiaľ nevieme zaručiť, iba za istých predpokladov. Z teoretickej stránky, o spoľahlivosti môžeme povedať jedine, že sa spoliehame na teóriu bootstrapu, pričom predpoklady a merané veličiny sa zdajú byť rozumné.

*Monotónny môže byť, keď riešili rovnaké úlohy a budeme rovnakým spôsobom vyberať nové podmnožiny.*

## 2.4 Bayesovský aposteriórny odhad presnosti

Vytvoríme odhad presnosti bayesovským usudzovaním priamo z IRT modelu. Výhodou tohto prístupu je, že budeme priamo využívať údaje existujúceho modelu, ktorý sa snaží modelovať všetky podstatné údaje a nepotrebujeme na to dodatočné predpoklady o rozdelení chyby.

Základné údaje poskytované IRT modelom sú parametre úloh a súťažiacich. Tieto parametre sú bodom maximálnej vierohodnosti v mnohorozmernom priestore pravdepodobností IRT modelu. Skúmaním tvaru tohto priestoru sme však schopný získať viacej informácií, presnejšie, ktoré parametre sú ako vyhranené.

Keď sa pozeráme na jedného súťažiaceho $s$, IRT model nám dáva apri-

órnu pravdepodobnosť, že vyrieši jednotlivé úlohy, ako funkciu jeho schopnosti. Aposteriórne, po súťaži, síce aj ďalej nevieme jeho schopnosť $\theta_s$, vieme jeho výsledky na jednotlivých úlohách, z ktorých získame odhad maximálnej vierohodnosti $\hat{\theta}_s$. Táto funkcia vierohodnosti nemusí dávať celkovú pravdepodobnosť jedna, ak ju však normalizujeme, dostaneme funkciu aposteriórnej pravdepodobnosti schopnosti súťažiaceho, za predpokladu platnosti modelu a dobrého odhadu úloh.

Odvodíme tento fakt matematicky. Budeme predpokladať že schopnosť $\theta$ súťažiaceho $s$ leží v intervale $<-10, 10>$, ako sme si model definovali. Ďalej budeme predpokladať znalosť funkcie $P_u$, ktorú sme odhadli pri nasádzaní modelu, udávajúcu pravdepodobnosť vyriešiť úlohu vzhľadom na schopnosť.

Funkcia $P_{\bar{u},\bar{v}}$ určuje pravdepodobnosť výskytu výsledkov $\bar{v}$ na úlohách $\bar{u}$.

$$P_{\bar{u},\bar{v}}(\theta) = \prod_i P_{u_i}(\theta)^{v_i} \left(1 - P_{u_i}(\theta)\right)^{1-v_i} \tag{2.17}$$

Náhodná premenná $\Theta$ modeluje schopnosť súťažiaceho a náhodná premenná $V$ modeluje vektor výsledkov. Aby sme sa vyhli miešaniu spojitých a diskrétnych premenných, schopnosť modelujeme diskrétne s $M$ rovnomerne rozdelených hodnôt na intervale $<-10, 10>$.

$$\begin{aligned} (\forall k, 0 \leq k < M) \qquad & \tau_k = \frac{20k}{M} - 10 \\ (\forall k, 0 \leq k < M) \qquad & P(\Theta = \tau_k) = \frac{1}{M} \end{aligned}$$

Dosadením do bayesovho vzorca dostávame

$$\begin{aligned} P(\Theta = \theta \mid V = \bar{v}) &= \frac{P(V = \bar{v}|\Theta = \theta)P(\Theta = \theta)}{\sum_\tau P(V = \bar{v}|\Theta = \tau)P(\Theta = \tau)} \\ &= \frac{P_{\bar{u},\bar{v}}(\theta)\frac{1}{M}}{\sum_\tau P_{\bar{u},\bar{v}}(\tau)\frac{1}{M}} \end{aligned}$$

Horeuvedená rovnosť je diskrétna, teda iba približne zodpovedá modelu. Jej transformáciou a limitným výpočtom, ktorý veríme, že nie je potrebné zo-

brazovať sa už dostaneme k presnej, spojitej formulácii.

$$P(\Theta < \theta) = \frac{\sum_{\tau<\theta} P_{\bar{u},\bar{v}}(\tau)\frac{1}{M}}{\sum_{\tau} P_{\bar{u},\bar{v}}(\tau)\frac{1}{M}}$$
$$P(\Theta < \theta) = \frac{\int_{-10}^{\theta} P_{\bar{u},\bar{v}}(x)\,\mathrm{d}x}{\int_{-10}^{10} P_{\bar{u},\bar{v}}(x)\,\mathrm{d}x} \tag{2.18}$$

Z distribučnej funkcie (2.18) vyjadríme funkciu hustoty.

$$f_s(\theta) = \frac{P_{\bar{u},\bar{v}}(\theta)}{\int_{-10}^{10} P_{\bar{u},\bar{v}}(x)\,\mathrm{d}x} \tag{2.19}$$

Funkcia hustoty $f_s$ vyjadruje aposteriórnu pravdepodobnosť schopnosti súťažiaceho v našom modeli, za predpokladu správneho odhadu úloh. Jej tvar presne zodpovedá našej slovnej úvahe.

Potrebovali sme ohraničiť súťažiaceho na konečný interval, aby sme zaručili konečnú hodnotu integrálu $\int P_{\bar{u},\bar{v}}(x)\,\mathrm{d}x$ a teda aj konečnú hustotu $f_s$. Tento predpoklad nemusí byť nutný, pokiaľ je schopnosť súťažiaceho dobre ohraničená úlohami, ako býva prípad väčšiny súťažiacich v 1PL a 2PL modeloch, no neplatí všeobecne.

Kvantily funkcie hustoty môžeme počítať nasledovne

$$Q_z(f_s) = min\left\{ y \;\middle|\; \frac{\int_{-10}^{y} f_s(x)\,\mathrm{d}x}{\int_{-10}^{10} f_s(x)\,\mathrm{d}x} \;<\; z \right\}$$

**Definícia 2.3.** Bayesovský odhad presnosti súťažiaceho $s$ miery pravdepodobnosti $p$ je interval $(d_s, h_s)$, ktorý obsahuje strednú $p$-tu čiastku hustoty aposteriórnej pravdepodobnosti $f_s$ (2.19).

$$d_s = Q_{(1-p)/2}(f_s) \tag{2.20}$$
$$h_s = Q_{(1+p)/2}(f_s) \tag{2.21}$$
$$\tag{2.22}$$

Štandardná voľba $p$ je napríklad 0.95, ako aj v bootstrapových odhadoch,

závisí však od počtu úloh.

O tomto odhade ukážeme, že je monotónny. Na to si najprv dokážeme lemu.

**Lema 1.** (O posúvaní masy) *Nech $f$ a $g$ sú kladné funkcie na intervale $(a, c)$ a majú definované konečné integrály na každom jeho podintervale.*

1. *Ak $g$ je rastúca, potom pre všetky $b$ v intervale $(a, c)$ platí*

$$\frac{\int_a^b f(x)g(x)\,\mathrm{d}x}{\int_a^c f(x)g(x)\,\mathrm{d}x} < \frac{\int_a^b f(x)\,\mathrm{d}x}{\int_a^c f(x)\,\mathrm{d}x} \tag{2.23}$$

2. *Ak $g$ je klesajúca, potom pre všetky $b$ v intervale $(a, c)$ platí*

$$\frac{\int_a^b f(x)g(x)\,\mathrm{d}x}{\int_a^c f(x)g(x)\,\mathrm{d}x} > \frac{\int_a^b f(x)\,\mathrm{d}x}{\int_a^c f(x)\,\mathrm{d}x} \tag{2.24}$$

*Dôkaz.* Najprv dokážeme bod 1. Keďže je funkcia $g$ rastúca, a funkcia $f$ je kladná, môžeme na ľavej časti intervalu v integrále funkciu $g$ zhora ohaničiť jej hodnotou v bode $b$ a vysunúť z integrálu.

$$\begin{aligned}
&\big(\forall x \in (a,b)\big) && f(x) > 0 \\
&\big(\forall x \in (a,b)\big) && g(x) < g(b) \\
&\big(\forall x \in (a,b)\big) && f(x)g(x) < f(x)g(b) \\
& && \int_a^b f(x)g(x)\,\mathrm{d}x < \int_a^b f(x)g(b)\,\mathrm{d}x \\
& && \int_a^b f(x)g(x)\,\mathrm{d}x < g(b)\int_a^b f(x)\,\mathrm{d}x
\end{aligned}$$

Podobne, na pravej časti môžeme integrál ohraničiť zdola.

$$\begin{aligned}
&\big(\forall x \in (b,c)\,\big) && f(x) > 0 \\
&\big(\forall x \in (b,c)\,\big) && g(b) < g(x) \\
&\big(\forall x \in (b,c)\,\big) && f(x)g(b) < f(x)g(x) \\
& && \int_b^c f(x)g(b)\,\mathrm{d}x < \int_b^c f(x)g(x)\,\mathrm{d}x \\
& && g(b)\int_b^c f(x)\,\mathrm{d}x < \int_b^c f(x)g(x)\,\mathrm{d}x
\end{aligned}$$

Využitím predpokladu, že sú funkcie kladné, môžeme prenásobiť tieto dve nerovnosti a zbaviť sa $g(b)$.

$$\begin{aligned}
g(b)\int_b^c f(x)\,\mathrm{d}x \int_a^b f(x)g(x)\,\mathrm{d}x &< g(b)\int_a^b f(x)\,\mathrm{d}x \int_b^c f(x)g(x)\,\mathrm{d}x \\
\frac{\int_b^c f(x)\,\mathrm{d}x}{\int_a^b f(x)\,\mathrm{d}x} &< \frac{\int_b^c f(x)g(x)\,\mathrm{d}x}{\int_a^b f(x)g(x)\,\mathrm{d}x}
\end{aligned}$$

Nerovnosti upravíme do požadovaného tvaru.

$$\begin{aligned}
\frac{\int_b^c f(x)\,\mathrm{d}x}{\int_a^b f(x)\,\mathrm{d}x} + 1 &< \frac{\int_b^c f(x)g(x)\,\mathrm{d}x}{\int_a^b f(x)g(x)\,\mathrm{d}x} + 1 \\
\frac{\int_a^c f(x)\,\mathrm{d}x}{\int_a^b f(x)\,\mathrm{d}x} &< \frac{\int_a^c f(x)g(x)\,\mathrm{d}x}{\int_a^b f(x)g(x)\,\mathrm{d}x} \\
\frac{\int_a^b f(x)g(x)\,\mathrm{d}x}{\int_a^c f(x)g(x)\,\mathrm{d}x} &< \frac{\int_a^b f(x)\,\mathrm{d}x}{\int_a^c f(x)\,\mathrm{d}x}
\end{aligned}$$

Dôkaz bodu 2 vyplýva z dôkazu bodu 1 pre funkcie $f^-(x) = f(-x)$ a $h^-(x) = h(-x)$ na intervale $(-c,-a)$. □

**Definícia 2.4.** Úloha $u$ je kladne hodnotená, ak je v modeli pravdepodobnosť jej vyriešenia $P_u$ rastúcou funkciou schopnosti súťažiaceho $\theta$.

**Definícia 2.5.** Úloha $u$ je záporne hodnotená, ak je v modeli pravdepodobnosť jej vyriešenia $P_u$ klesajúcou funkciou schopnosti súťažiaceho $\theta$.

**Lema 2.** *Nech súťažiaci $s_1$, $s_2$ a $s_3$ riešili úlohy $\bar{u}_1$, $\bar{u}_2$ a $\bar{u}_3$ s výsledkami $\bar{v}_1$,*

*$\bar{v}_2$ a $\bar{v}_3$. Nech $u$ je kladne hodnotená úloha a platí*

$$\bar{u}_1 = \bar{u}_2||u$$
$$\bar{v}_1 = \bar{v}_2||0$$
$$\bar{u}_3 = \bar{u}_2||u$$
$$\bar{v}_3 = \bar{v}_2||1$$

*Potom platí*

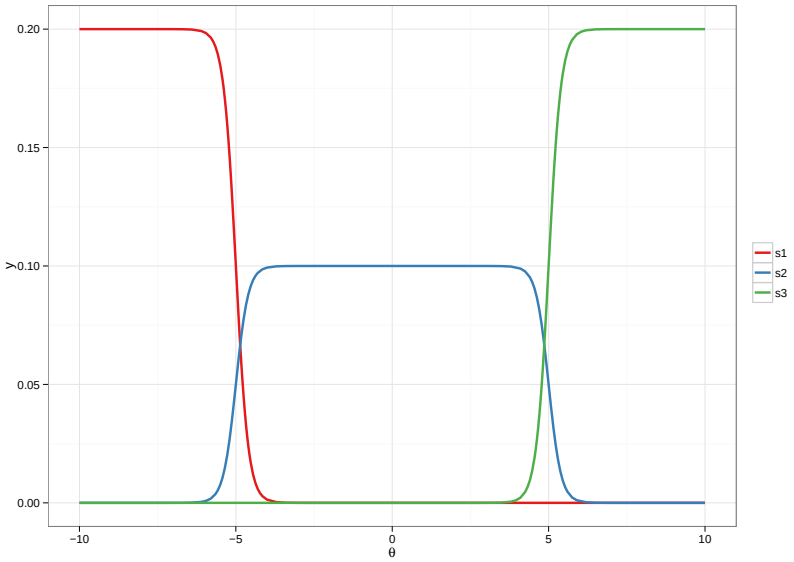
$$(\forall z \in (0,1)) \quad Q_z(f_{s_1}) < Q_z(f_{s_2})$$
$$(\forall z \in (0,1)) \quad Q_z(f_{s_3}) < Q_z(f_{s_2})$$

*Proof.* Z lemy 1, definície kvantilu a kladného hodnotenia úlohy $u$ vyplýva požadované tvrdenie. □

Na základe predchádzajúcej lemy je bayesov odhad chyby dvoch úlohovo porovnateľných súťažiacich monotónny na kladne hodnotených úlohách.

*V porovnaní dvoch možných výsledkov súťaže, možno si môže súťažiaci polepšiť neriešením alebo zlým vyriešením vzhľadom na iných, úlohovo neporovnateľných súťažiacich.*

*Zaujímavé je, že tento dolný odhad môže byť väčší ako odhad maximálnej vierohodnosti. Usporiada aj tých, čo nič nevyriešili, podľa toho, ako zlý sa stihli ukázať.*




Obr. 2.3: Bayesovský odhad schopností pre príklad 2.1

# Kapitola 3

# Implementácia a výsledky

Táto kapitola začína prehľadom existujúcich programov na nasádzanie IRT modelov, kde poukazujeme na nedostatok voľne šíriteľných implementácií.

Začínajúc podkapitolou 3.2 prezentujeme vlastný prínos, kde najprv ukážeme naše optimalizácie rýchlosti a presnosti výpočtu základných IRT kriviek, ktoré sú časovo najnáročnejšou operáciou programu.

Následne rozoberieme algoritmy na nasádzanie modelov, konkrétne JML, MML a vlastnú ideu priameho použitia všeobecného algoritmu optimalizácie L-BFGS-B. Porovnáme ich čas behu a ukážeme vplyv apriórnych predpokladov distribúcie v MML algoritme na poradie.

Na záver urobíme ukážku použitia IRT na tvorenie úsudkov, ktoré by v klasickej teórii bolo ťažké poukázať. Konkrétne, pomerne jednoduchým ratingom a úsudkom o zložení testov ukážeme, že študenti poznajú minulé testy z princípov počítačov na základe výsledkov testov.

## 3.1 Existujúce implementácie

Streiner vo svojej práci [Str10] píše ako v štatistike dominuje Classical Test Theory, pokiaľ sa iba zriedkavo používa Item Response Theory. Pritom jeden z dôvodov je aj to, že všetky bežne používané štatistické nástroje umožňujú používať CTT bez premýšľania, a IRT vyžaduje použitie špecializovaných a

často drahých počítačových programov.

Našli sme niekoľko štatistických nástrojov na analýzu pomocou IRT, z čoho bolo zopár veľkých komerčných nástrojov, za ktoré sa domnievame, že by všetky mali podporovať nami potrebnú funkčnosť a mnoho malých bezplatných nástrojov zameraných na jednotlivé činnosti, ktoré podporujú iba jeden, prípadne niekoľko podobných IRT modelov. Väčšina z týchto malých programov sú pre nás nezaujímavé, lebo sú zamerané na iné činnosti ako napríklad faktorovú analýzu a pracujú s inými modelami.

V našom prieskume sme našli nasledovné programy, ktoré poskytujú možnosť odhadu parametrov IRT modelov s ktorými pracujeme, konkrétne 2PL a 3PL.

1. IRTPro je v terajších časoch najznámejším IRT programom. Ako vlajková loď spoločnosti Scientific Software International je nástupcom ich predchádzajúcich známych IRT programov Bilog-MG a Parscale. Cena akademickej licencie pre jedného používateľa je 500$.
   http://www.ssicentral.com/irt/index.html

2. Mplus je tiež známy IRT program. Vyrába ho spoločnosť Muthén & Muthén a akademická licencia pre jedného používateľa je 600$.
   http://www.statmodel.com/programs.shtml

3. Xcalibre je výrobkom spoločnosti Assessment Systems Corporation. Jedna akademická licencia tiež stojí okolo 500$.
   http://www.assess.com/xcart/product.php?productid=415

4. ETIRM je opensource knižnica voľne dostupná pod BSD licenciou. Napísal a održiaval ju Brad Hanson v rokoch 2000 - 2002. Ďalšiu údržbu na nej robil Werner Wothke v roku 2008, odkedy je dostupná na jeho stránke. Knižnica podporuje MML odhad parametrov úloh pre 1PL, 2PL, 3PL, PCM (partial credit) a GPCM (generalized partial credit) modely a odhad schopností EAP (expected a posteriori) a MLE (maximum likelihood estimate). Brad Hanson na nej založil aj svoj program

ICL, ktorého vývoj skončil v roku 2002.
http://www.smallwaters.com/software/cpp/etirm.html
http://www.b-a-h.com/software/irt/icl

5. LTA-2 je voľne dostupný program v tvare EXE súbora od autora J. S. Uebersax-a. Podporuje odhad parametrov úloh a bayesovský EAP (expected a posteriori) odhad schopnosti osôb v 1PL a 2PL modeloch.
http://john-uebersax.com/stat/papers.htm

Pre nás boli možnosti 1 až 4 nevhodné tým, že boli drahé a teda nedostupné. Možnosť 6 (LTA-2) je ohraničený program, ktorý nepodporuje 3PL model a nemá dostupný zdrojový kód.

Ako Mead a spol. píšu v [MMB07], program ICL a teda aj knižnica ETIRM poskytuje veľa možností, pre toho, kto ju vie používať. Ovládanie nie je zrovna triviálne ale ICL má 90 stranový manuál, ktorý poskytuje návod na jeho použitie. Knižnica bola pre nás užitočná, poskytujúc možnosť MML odhadov, ktoré sme používali pri malých vzorkách. Domnievame sa, že a priori predpoklady o distribúcii odhadovaných parametrov, ktoré MML predkladá navyše mu umožňujú dávať stabilné odhady aj pri malých vzorkách.

Rozhodli sme sa implementovať vlastný nástroj, aby sme vyskúšali JML odhady, ktoré nemajú dodatočné predpoklady o normálnej distribúcii skúšaných osôb, o ktorej Forišek v práci [For09a] ukazuje že v situáciách ako sú súťaže neplatia. Ďalej sme skúmali vylepšenia v algoritmoch a počítali štatistiky, ktoré dostupné nástroje priamo nepodporujú.

## 3.2 Numerické výpočty

Základom IRT modelov je pravdepodobnostná funkcia odozvy. Pre 3PL model je to funkcia $P$ (3.1). Naším cieľom je maximalizovať celkovú pravdepodobnosť, ktorá je súčinom pravdepodobností jednotlivých výsledkov. Pri tom sa môžu vyskytovať veľmi malé čísla, ktoré nie sú vyjadriteľné v počítači ako bežné reálne premenné. Vyhneme sa tomu problému štandardným tri-

kom počítania s logaritmom pravdepodobnosti. Logaritmus budeme počítať priamo pri výpočte funkcie odozvy, funkciou $logP$ a jej komplementom $logQ$.

$$P(\theta, a, b, c) = c + \frac{1-c}{1+e^{-a(\theta-b)}} \tag{3.1}$$

$$logP(\theta, a, b, c) = \ln\left(c + \frac{1-c}{1+e^{-a(\theta-b)}}\right) \tag{3.2}$$

$$logQ(\theta, a, b, c) = \ln\left(1 - c - \frac{1-c}{1+e^{-a(\theta-b)}}\right) \tag{3.3}$$

Pri nasádzaní modelu je časovo najnáročnejšou operáciou počítanie funkcií $logP$, $logQ$ a jej derivácií. Budeme sa preto snažiť rýchlo počítať tieto funkcie, ale pritom zaručiť aj ich numerickú stabilitu, aby sme sa vyhli nepríjemným prekvapeniam.

Na zjednodušenie funkcie $logP$ a $logQ$ vyjadríme ako kompozície dvoch funkcií $lgp$ resp. $lgq$ a $pos$. Funkcie $lgp$ a $lgq$ dávajú tvar krivky a $pos$ počíta polohu na základe parametrov úlohy a schopnosti súťažiaceho. V prípade 2PL modelu môžeme $lpq$ dostať z $lgp$ zmenením znamienka funkcie $pos$ resp. premennej $a$. Model $3PL$ však nie je symetrický, budeme teda tieto dve funkcie počítať osobitne.

$$logP(\theta, a, b, c) = lgp(pos(\theta, a, b), c) \tag{3.4}$$

$$logQ(\theta, a, b, c) = lgq(pos(\theta, a, b), c) \tag{3.5}$$

$$pos(\theta, a, b) = -a(\theta - b) \tag{3.6}$$

$$lgp(x, c) = \ln\left(c + \frac{1-c}{1+e^{x}}\right) \tag{3.7}$$

$$lgq(x, c) = \ln\left(1 - c - \frac{1-c}{1+e^{x}}\right) \tag{3.8}$$

Funkcia $pos$ stráca relatívnu presnosť v okolí $\theta \to b$, ale iba do miery v akej sú definované tieto dva parametre. Numerickej stabilite to neprekáža, pokiaľ sa takto počíta dôsledne.

Funkcia *lgp* je numericky problematická v okolí $c \to 1$ a pre záporné hodnoty $x$. V prípade $c \to 1$ je relatívna presnosť malá, ale strata absolútnej presnosti ohraničená a pomerne nevýznamná. Pre záporné hodnoty $x$ má stupňujúce sa chyby, až okolo $x = -37$ začína byť konštantnou. Vylepšili sme ju pozmenením v tvar (3.9), ktorý je na počítači už pomerne presný vďaka podpore počítania funkcie $\ln(1+x)$.

$$\begin{aligned} lgp(x,c) &= \ln\left(c + \frac{1-c}{1+e^x}\right) \\ &= \ln\frac{c + ce^x + 1 - c}{1+e^x} \\ &= \ln\frac{1+ce^x}{1+e^x} \\ &= \ln(1+ce^x) - \ln(1+e^x) \end{aligned} \tag{3.9}$$

Na druhej strane, tvar (3.9) má dva logaritmy, ktoré sú pomalé operácie. Experimentálne výsledky ukazujú, že sa tento tvar počíta asi 15% pomalšie. Ďalšími úpravami dostaneme tvar (3.10), ktorý už obsahuje iba jeden logaritmus, zvláda veľké rozpätie hodnôt $x$ ($-709$ až 709 namiesto $-36$ až 709) a počíta sa ešte rýchlejšie ako pôvodný nepresný tvar (3.7).

$$\begin{aligned} lgp(x,c) &= \ln\frac{1+ce^x}{1+e^x} \\ &= -\ln\frac{1+e^x}{1+ce^x} \\ &= -\ln\left(1 + \frac{1+e^x-1-ce^x}{1+ce^x}\right) \\ &= -\ln\left(1 + (1-c)\frac{e^x}{1+ce^x}\right) \end{aligned} \tag{3.10}$$

Poučený faktom, že potrebujeme použiť funkciu $\ln(1+x)$, aby bola zachovaná presnosť keď sú výsledné hodnoty v blízkosti nuly, upravíme funkciu *lgq*. Presnejšie, chceme aby to bol tvar $-\ln(1+x)$, čím dosiahneme, že argument $x$ nebude v intervale $<-1, 0>$, kde nemôžeme presne vyjadriť hodnoty

v okolí $-1$, ale v intervale $<0,\infty)$, kde môžeme využiť celý rozsah registrov počítača. Takto sa priamo dopracujeme k tvaru (3.11).

$$\begin{aligned} lgq(x,c) &= \ln\left(1-c-\frac{1-c}{1+e^x}\right) \\ &= \ln\frac{e^x-ce^x}{1+e^x} \\ &= -\ln\frac{1+e^x}{e^x-ce^x} \\ &= -\ln\left(1+\frac{1+ce^x}{(1-c)e^x}\right) \end{aligned} \tag{3.11}$$

Potrebujeme ešte spočítať parciálne derivácie funkcií $logP$ a $logQ$. Urobíme to tak, že najprv zderivujeme funkcie $lgp$, $lgq$ a $pos$ a z nich vyskladáme výsledné funkcie. Keďže je to pomerne jednoduchý a monotónny proces, nebudeme ho tu rozpisovať. Uvádzame iba výsledky pre budúce použitie.

$$\frac{\partial}{\partial x}\, lgp(x,c) = \frac{(c-1)e^x}{(1+ce^x)(1+e^x)} \tag{3.12}$$

$$\frac{\partial}{\partial c}\, lgp(x,c) = \frac{e^x}{1+ce^x} \tag{3.13}$$

$$\frac{\partial}{\partial x}\, lgq(x,c) = \frac{e^{-x}}{1+e^{-x}} \tag{3.14}$$

$$\frac{\partial}{\partial c}\, lgq(x,c) = \frac{1}{c-1} \tag{3.15}$$

$$\frac{\partial}{\partial a}\, pos(\theta,a,b) = b-\theta \tag{3.16}$$

$$\frac{\partial}{\partial b}\, pos(\theta,a,b) = a \tag{3.17}$$

$$\frac{\partial}{\partial \theta}\, pos(\theta,a,b) = -a \tag{3.18}$$

Hodné povšimnutia je, že po derivácii funkcie $lgq$ podľa $x$ sme rovnicu z pôvodného tvaru $\frac{1}{1+e^x}$ zmenením znamienka $x$ dali do tvaru (3.14), aby sme zachovali presnosť.

## 3.3 Algoritmus

Implementovali JML algoritmus, popísaný v kapitole 1, ktorý striedavo optimalizuje parametre súťažiacich a úloh. Pri tejto maximalizácii vierohodnosti je potrebné hľadať maximum jednej až troch premenných, na čo sme použili dostupné algoritmy z knižnice GNU Scientific Library (GSL) z jednoduchého dôvodu, že implementácia je v jazyku haskell a pre knižnicu sú vytvorené haskellovské väzby. Knižnica ponúka na výber niekoľko algoritmov, z ktorých najužitočnejšie boli VectorBFGS2, ktorý je gradientovou metódou a NMSimplex2, ktorý využíva iba funkčné hodnoty.

Prekvapivo, na optimalizáciu funcie viacero premenných sa lepšie ukázala simplexová metóda, ktorá sa nemala obyčaj zatúlať, na rozdiel od gradientovej BFGS. Na druhej strane, síce BFGS je zložitým algoritmom na optimalizovanie parametrických funkcií, ukázal sa ako vhodná voľba na optimalizovanie funkcie jednej premennej. Predpokladáme, že má náskok tým, že využíva gradient funkcie a počíta z neho a funkčnej hodnoty lokálny odhad. Chceli by sme však poznamenať, že tieto výsledky môžu byť aj vecou implementácie týchto algoritmov, na ktorej kvalitu sú určité poznámky.

V prítomnosti hraničných hodnôt parametrov, ktoré sa vyskytujú pri súťažiacich a úlohách s perfektným alebo nulovým skóre, JML algoritmus sa dostane do stavu, kde robí veľmi malé, skoro bezvýznamné zmeny. Predpokladáme, že je to tým, že jedna, alebo veľmi málo, hodnôt posúva celú masu ostatných. Veľké problémy to nerobí a zmeny hodnôt sú malé, vyžaduje to však od operátora aby stanovil kritérium na zastavenie. Z dôvodu, že sa parametre súťažiacich a úloh optimalizujú striedavo, teoreticky sa môže stať, že sa parametre budú cyklicky posúvať v opačnom smere. Implementovali sme preto možnosť nastaviť požadovanú presnosť a maximálny počet kôl.

Z tohto dôvodu sme ako ďalšiu možnosť vyskúšali priamu optimalizáciu celkovej vierohodnosti modelu, pomocou Limited memory Broyden-Fletcher-Goldfarb-Shanno Bounded (L-BFGS-B) algoritmu. Použili sme implementáciu vo fortrane autorov Zhu et al. [MN11] pre ktorú sme vytvorili haskellovské väzby. Ukázalo sa, že tento spôsob optimalizácie je niekoľkokrát pomalší

než JML. Priama podpora ohraničenia povoleného rozsahu premenných však umožnila dosiahnuť predpokladanú vlastnosť zastavenia.

*Neporovnávali sme Cedalovu implementáciu L-BFGS-B s implementáciou BFGS knižnice GSL.*

## 3.4 Porovnanie algoritmov MML, JML a BFGS

Našu implementáciu sme porovnali s referenčnou implementáciou Foriška a implementáciou algoritmu MML v knižnici ETIRM autora Brad-a Hanson-a.

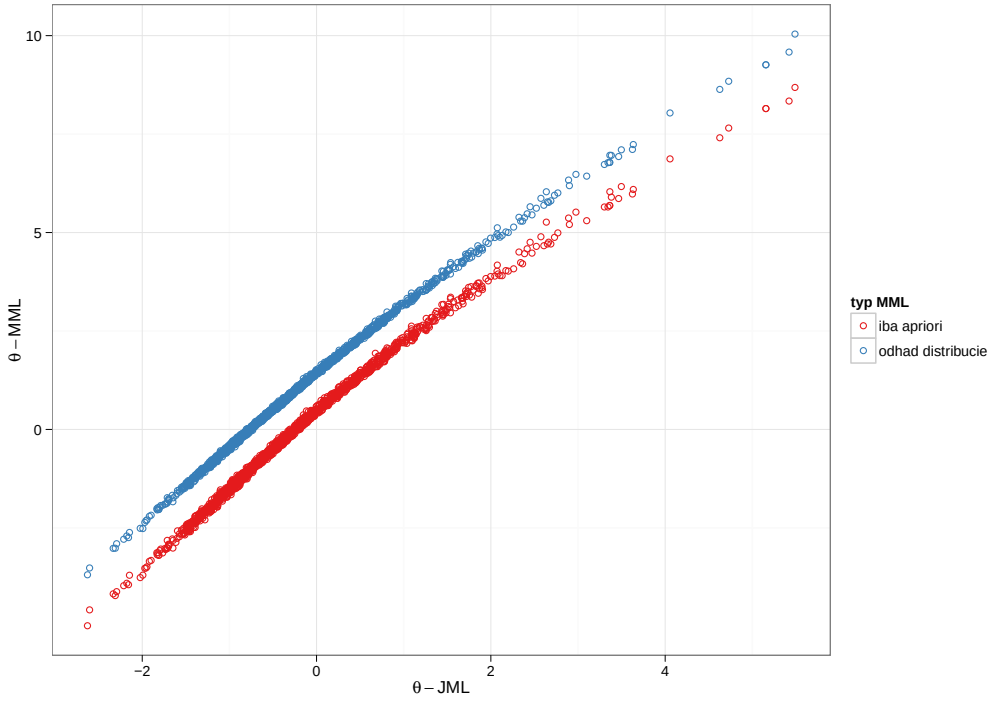
Ukázalo sa, že naša implementácia JML a priamej optimalizácie pomocou L-BFGS-B dáva zhodné výsledky s referenčnou implementáciou algoritmu JML pre 2PL model obsiahnutú v práci [For09a] Foriška, do stanovenej presnosti výpočtu, až na škálovanie a posun.

Očakávajúc rozdiely v algoritmoch MML a JML, kvôli modelom, ktoré sa odlišujú dodatočnými predpokladmi apriórnej distribúcie v algoritme MML, sme porovnali ich odhady schopností, nasadením 2PL modelu na štandardnú testovaciu sadu mondatx odpovedí 1655 osôb na 36 otázok od M. Kolena a R. Brenanna (1995), ktorá šla s knižnicou ETIRM. Na grafe 3.1 je znázornené odhadované schopnosti týmito algoritmami. Vidno, že dodatočný predpoklad apriórnej distribúcie pozmenil odhad úloh, čo sa následne prejavilo v odhade schopností, ktoré sú už odhadované zaužívanou metódou maximálnej vierohodnosti, bez apriórnych predpokladov. Algoritmom MML bolo získané iné usporiadanie subjektov, ktoré sa prejavilo už na 9-om z 1655 miest.

Knižnica ETIRM poskytuje v rámci MML algoritmu možnosť priebežného dolaďovania distribúcie počas odhadu, na základe aktuálne dosiahnutej distribúcie, v základnom nastavení je však vypnutá. Dôvodom môže byť dlhší beh algoritmu a podobný problém s neistotou konvergencie ako pri JML algoritme. V prípade zapnutia tejto možnosti odhad bol iný ako bez nej a viacej sa priblížil JML odhadu. Prvý výskyt zmeny sa posunul na 12-te miesto. Apriórna distribúcia je voliteľná a tvorila by ešte ďalšie možnosti usporiadania, nie sme však experti na nastavovanie distribúcií a preto sme použili

prednastavené hodnoty.

Z tohto vidíme, že apriórne distribúcie v MML algoritmoch majú vplyv na usporiadanie subjektov, čím sa JML algoritmus stáva lákavejším, keďže má predpoklady jedine samotného IRT modelu a tvorí jednoznačné usporiadanie. Predpokladom však je, aby model bol dobre definovaný, čo vyžaduje dostatočné množstvo údajov. Výhodou MML je, že je jednoznačne definovaný a konverguje pre ľubovoľné množstvo údajov, otázne sú však nastavenia apriórnych distribúcií.




Obr. 3.1: Porovnanie odhadov schopností získaných algoritmami JML a MML pre 2PL model. Výsledky MML so zapnutým odhadom distribúcie sú pre viditeľnosť posunuté smerom hore.

## 3.5 Testy z princípov počítačov

Na jednoduchom rozbore testov z princípov počítačov ukážeme niektoré možnosti poskytované v IRT, ktoré nie sú poskytované v CTT.

Aby sme ujasnili vzťah medzi CTT a IRT, skonštatujeme, že všetko čo poskytuje IRT, je principiálne možné aj v CTT, otázkou však je, aká zložitá analýza by bola na to potrebná a do akej miery by teória potrebná na to vlastne bola ekvivalentom IRT.

Budeme rozoberať testy z povinného informatického predmetu princípy počítačov `1-INF-130`. Zadania testov a výsledky nám sprístupnil R. Ostertág, ktorý učí predmet a robil testy. Konkrétne sa jedná o päť krúžkovacích testov z troch rokov vyučovania, z ktorých každý obsahuje dvadsať otázok. Otázky majú štyri možné odpovede, z ktorých by mala byť aspoň jedna správna a jedna nesprávna. Otázka sa považuje za správne odpovedanú, keď obsahuje práve všetky platné odpovede. Navyše, študenti majú možnosť neodpovedať, dávajúc výsledný počet bodov získaných na otázke z množiny {-2, -1, 1}.

Pre spoločné spracovanie testov sme prešli všetkými testami a pooznačovali opakujúce sa úlohy. Keďže niektoré úlohy neboli úplne rovnaké, ale pritom boli miernou obmenou starej, urobili sme dve tabuľky, kde v jednej sme prihliadali na tieto rozdiely a v druhej nie. Výsledky sú v tabuľke 3.1.

Z odpovedí všetkých študentov v tých piatich testoch sme urobili jeden súbor dát, kde sme rovnakých študentov na rôznych testoch považali za rôznych, predpokladajúc, že sa možno medzi časom niečo naučili. Na možnosť neodpovedať sme neprihliadali a všetkých sme hodnotili podľa toho, či odpovedali správne na položenú otázku alebo nie.

Na tieto údaje sme následne nasadili 3PL model pomocou algoritmu MML, knižnice ETIRM. Model 3PL sme použili preto, že sme chceli vidieť, či majú otázky typovací charakter napriek 14 teoreticky možných odpovedí. Algoritmus MML sme zvolili preto, že model 3PL obsahuje veľký počet voľných parametrov vzhľadom na množstvo údajov. Keďže cieľom bolo získať výskumné údaje, podobne ako psychológov, konkrétne usporiadanie študentov nám nebolo veľmi podstatné. V prípade, že záleží na ich hodnotení, skôr by sme odporúčali 1PL model a JML algoritmus, pokiaľ je to možné.

Celkovo študentov, ktorý odpovedali na aspoň jednu otázku správne bolo 119. Študenti, ktorí nemali ani jednu správnu odpoveď nám pri tvorbe modelu

| | 1 | 2 | 3 | 4 | 5 | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.01 | 2.01 | 3.01 | 3.01 | 5.01 | 1 | 1.01 | 2.01 | 3.01 | 3.01 | 1.01 |
| 2 | 1.02 | 2.02 | 3.02 | 3.02 | 5.02 | 2 | 1.02 | 1.02 | 3.02 | 3.02 | 5.02 |
| 3 | 1.03 | 2.03 | 3.03 | 3.03 | 1.03 | 3 | 1.03 | 1.03 | 3.03 | 3.03 | 1.03 |
| 4 | 1.04 | 2.04 | 3.04 | 3.04 | 1.04 | 4 | 1.04 | 2.04 | 3.04 | 3.04 | 1.04 |
| 5 | 1.05 | 2.05 | 3.05 | 3.05 | 1.05 | 5 | 1.05 | 1.05 | 3.05 | 3.05 | 1.05 |
| 6 | 1.06 | 2.06 | 3.06 | 4.06 | 1.06 | 6 | 1.06 | 1.06 | 3.06 | 3.06 | 1.06 |
| 7 | 1.07 | 2.07 | 3.07 | 3.07 | 1.07 | 7 | 1.07 | 2.07 | 3.07 | 3.07 | 1.07 |
| 8 | 1.08 | 1.08 | 3.08 | 3.08 | 1.08 | 8 | 1.08 | 1.08 | 3.08 | 3.08 | 1.08 |
| 9 | 1.09 | 2.09 | 3.09 | 4.09 | 3.05 | 9 | 1.09 | 1.09 | 3.09 | 3.09 | 3.05 |
| 10 | 1.10 | 2.10 | 3.10 | 3.10 | 1.10 | 10 | 1.10 | 1.10 | 1.06 | 1.06 | 1.10 |
| 11 | 1.11 | 2.11 | 3.11 | 4.11 | 1.11 | 11 | 1.11 | 2.11 | 3.11 | 4.11 | 1.11 |
| 12 | 1.12 | 1.12 | 3.12 | 3.12 | 1.12 | 12 | 1.12 | 1.12 | 3.12 | 3.12 | 1.12 |
| 13 | 1.13 | 1.13 | 3.13 | 3.13 | 1.13 | 13 | 1.13 | 1.13 | 3.13 | 3.13 | 1.13 |
| 14 | 1.14 | 1.14 | 3.14 | 3.14 | 1.14 | 14 | 1.14 | 1.14 | 3.14 | 3.14 | 1.14 |
| 15 | 1.15 | 1.15 | 2.11 | 2.11 | 1.15 | 15 | 1.15 | 1.15 | 2.11 | 2.11 | 1.15 |
| 16 | 1.16 | 2.16 | 1.12 | 4.16 | 5.16 | 16 | 1.16 | 1.16 | 1.12 | 1.12 | 1.16 |
| 17 | 1.17 | 2.17 | 1.13 | 1.13 | 1.17 | 17 | 1.17 | 2.17 | 1.13 | 1.13 | 1.17 |
| 18 | 1.18 | 1.18 | 1.14 | 4.18 | 1.18 | 18 | 1.18 | 1.18 | 1.14 | 4.18 | 1.18 |
| 19 | 1.19 | 2.19 | 2.16 | 4.19 | 1.19 | 19 | 1.19 | 2.19 | 1.16 | 1.16 | 1.19 |
| 20 | 1.20 | 1.20 | 3.20 | 3.20 | 1.20 | 20 | 1.20 | 1.20 | 3.20 | 3.20 | 1.20 |
| nerov-nakých | 20 | 13 | 15 | 12 | 3 | zásadne nových | 20 | 6 | 14 | 2 | 1 |

Tabuľka 3.1: Výskyt úloh na teste

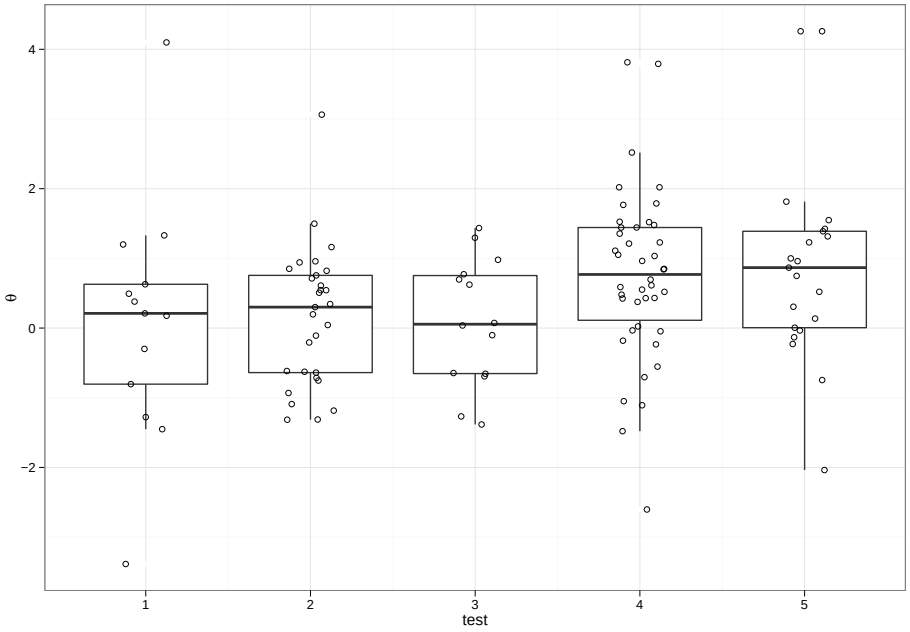
neprekážali. Týchto 24 študentov, ktorí sa z rôznych príčin vo výsledkovke nachádzali z analýzy vynecháme.

V klasickej teórii testov sú dva rôzne testy neporovnateľné. Použitie IRT dáva výsledky, ktoré by nemali byť závislé na konkrétnom teste. Keďže meranou veličinou je schopnosť riešiť otázky, ktoré sa na testoch vykytujú. Ak študenti poznajú minulé testy, dá sa čakať, že ich schopnosti budú vzrastať. Navyše, ak sa opakujúce otázky vyskytujú hlavne z posledných testov a tento proces sa dlhšie opakuje, pri nasadení IRT modelu by mal vzniknúť lineárny posun hodnotení vzhľadom na test. Vzhľadom na štruktúru otázok, v týchto testoch by sa lineárny posun nemal prejaviť.

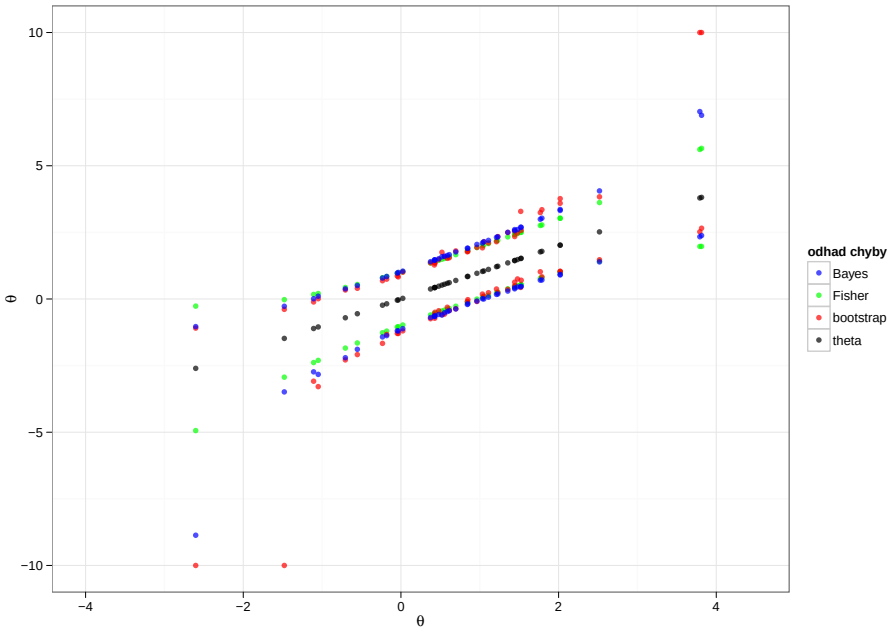
Naším predpokladom teda je, že sú schopnosti študentov v tých troch ročníkoch približne rovnaké a schopnosti skupiny riešiť test by mali korelovať so znovupoužitím starých otázok. Na overenie tejto hypotézy sme urobili graf 3.3 odhadnutých schopností študentov vzhľadom na test.

Na grafe vidno posun a to najmä na treťom a štvrtom teste, kde tretí bol skoro úplne nový test a štvrtý bol prevažne recyklácia tretieho.

Na záver uvádzame ukážku intervalových odhadov pre štvrtý test. Nedali sme ich všetky spolu, lebo bol graf nečitateľný.




Obr. 3.2: Porovnanie odhadov schopností študentov na testoch. Čiara je medián, študenti sú guľôčky.




Obr. 3.3: Intervalové odhady schopností pre 4. test

# Záver

V práci sme poukázali na nevýhody symetrických odhadov presnosti pridelených hodností, ktoré používajú všetky nám známe ratingové algoritmy.

Následne sme ukázali, že v modele Item Response Theory je možné vytvoriť asymetrický intervalový odhad, konštrukciou konštrukciu bootstrapového a Bayesovského odhadu, na čo sme nepoužívali ďalšie predpoklady než samotný IRT model. Pre Bayesovský odhad presnosti sme ukázali, že za rozumných podmienok je monotónny na úlohovo porovnateľných súťažiacich.

Urobili sme prehľad dostupných implementácií programov na nasádzanie IRT modelov. Kvôli nedostupnosti implementácie pre rating vhodného, ale inak menej používaného algoritmu JML sme implementovali vlastný nástroj, ktorý podporuje aj výpočet našich odhadov presnosti.

Na ukážku sme využitím IRT na rozbor výsledkov testov ukázali, že študenti poznajú testy z minulých skúšok.

O Bayesovskom odhade sme ukázali zaujímavé teoretické vlastnosti, ktoré ho presnosťou a stabilitou môžu urobiť vhodnejším kandidátom než ostatné systémy, zostáva však na budúci výskum jeho praktické porovnanie s inými systémami.

Bolo by vhodné do budúcna bližšie preskúmať vhodnosť JML a MML algoritmov na použitie v súťažiach, kvôli otáznej konvergencie jedného a apriórnym predpokladom druhého algoritmu, ktoré spochybňujú jednoznačnosť výsledkov.

*Okrem na katedrovej stránke registrovaných prác a stránke autora*[1]*, naša*

[1] http://people.ksp.sk/~ivan/hirt/

*implementácia je dostupná ako štandardný haskell-ovský balík. Na systéme s funkčným haskellom, potrebným na kompiláciu nášho kódu, na jeho inštalovanie by mal stačiť príkaz:*

```
cabal install hirt
```

*Prípadne iba na stiahnutie a rozbalenie:*

```
cabal unpack hirt
```

# Literatúra


[BA81] R. D. Bock and M. Aitken. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46:443–459, 1981. 1.4.2

[CB02] G. Casella and R.L. Berger. *Statistical inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. 2.3

[For09a] Michal Forišek. *Theoretical and Practical Aspects of Programming Contest Ratings*. PhD thesis, Comenius University Bratislava, 2009. 1.1, 1.2, 1.3, 2, 2.1, 2.2, 3.1, 3.4

[For09b] Michal Forišek. Using Item Response Theory To Rate (Not Only) Programmers. In Valentina Dagiene, editor, *Olympiads in Informatics*, volume 3, pages 3–16, 2009. 2.2

[Gli12] Mark E. Glickman. Example of the Glicko-2 system, 2012. http://www.glicko.net/glicko.html. 2.2

[HMG07] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, Cambridge, MA, 2007. 2.2

[Kli05] Theresa J.B. Kline. *Psychological Testing: A Practical Approach to Design and Evaluation*, chapter 5. SAGE Publications, 2005. 1.3.1

[MMB07] Alan D. Mead, Scott B. Morris, and David L. Blitz. Open-source IRT: A comparison of BILOG-MG and ICL features and item parameter recovery, 2007. 3.1

[MN11] José Luis Morales and Jorge Nocedal. Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". *ACM Transactions on Mathematical Software*, 38(1):7:1–7:4, November 2011. 3.3

[Str10] David L Streiner. Measure for Measure: New Developments in Measurement and Item Response Theory. *Canadian journal of psychiatry Revue canadienne de psychiatrie*, 55(3):180–186, 2010. 1.2, 3.1

[TJ90] Robert K. Tsutakawa and Jane C. Johnson. The Effect of Uncertainty of Item Parameter Estimation on Ability Estimates. *Psychometrika*, 55:371–390, 1990. 2.2

[vdLH96] Wim J. van der Linden and Ronald K. Hambleton. *Handbook of Modern Item Response Theory*. Springer, 1996. 1.3.2, 1.3.2, 1.4.1, 1.4.2

[WH97] David Woodruff and Bradley A. Hanson. Estimation of Item Response Models Using the EM Algorithm for Finite Mixtures. In *Annual Meeting of the Psychometric Society*, 1997. 1.4.2