Lesson 7: Item response theory models (part 2)

Patrícia Martinková

Department of Statistical Modelling
Institute of Computer Science, Czech Academy of Sciences

Institute for Research and Development of Education
Faculty of Education, Charles University, Prague

NMST570, November 20, 2018

Outline

Review: IRT models

Framework for

- estimating *latent traits* (ability levels) θ

  by means of *manifest* (observable) variables (item responses)

  and appropriate *psychometric* (statistical) model

Notes:

- Ability θ is often treated as random variable (but see further)
- Items: dichotomous, polytomous, multiple-choice, ...
- IRT model: describes probability of (correct) answer as function of
  - ability level and
  - item parameters

  This function is called:

  - *Item response function (IRF)*
  - *Item characteristic curve (ICC)*

Review: Introduction to IRT models

Use of IRT models

- To calibrate items (i.e. to estimate difficulty, discrimination, guessing,...)
- To assess respondents' latent trait (ability, satisfaction, anxiety,...)
- To describe test properties (standard error, test information,...)
- Test linking and equating, computerized adaptive testing, etc.

IRT model assumptions

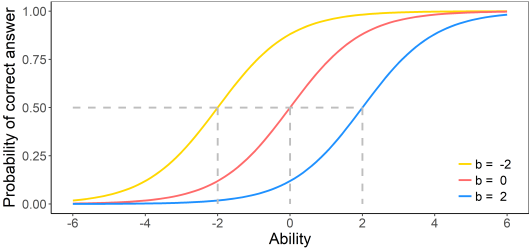
1. Model definition (functional form, usually monotonic ICC)
   - e.g. 2PL IRT model: $\mathrm{P}\left(Y_{ij}=1|\theta_i,a_j,b_j\right)=\pi_{ij}=\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}$
2. Unidimensionality of latent variable θ
3. Local independence (conditional independence)
   - e.g. $\mathrm{P}\left(Y_{i1}=1,Y_{i2}=1|\theta_i,a_j,b_j\right)=\pi_{i1}\cdot\pi_{i2}$
   - e.g. $\mathrm{P}\left(Y_{i1}=1,Y_{i2}=0|\theta_i,a_j,b_j\right)=\pi_{i1}\cdot(1-\pi_{i2})$
4. Invariance of parameters
5. Independence of respondents

Rasch Model

$$\pi_{ij} = P(Y_{ij} = 1 | \theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)} \quad (1)$$

θ_i ability of person i, for $i = 1, \ldots, I$
b_j difficulty of item j (location of inflection point) for $j = 1, \ldots, J$

Item Characteristic Curve (ICC)



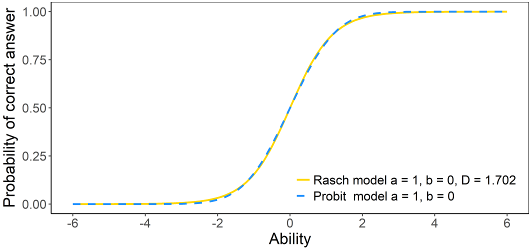
Note: Originally, Rasch model denoted as $\pi_{ij} = \frac{\tau_i}{\tau_i + \xi_j}$.
To get to (1), consider $\theta_i = \log(\tau_i)$, and $b_j = \log(\xi_j)$ for $\tau_i > 0, \xi_j > 0$

Logistic vs. Probit model (Note on Scaling parameter D)

Rasch model is sometimes defined as:

$$\pi_{ij} = P(Y_{ij} = 1|\theta_i, b_j) = \frac{\exp(D[\theta_i - b_j])}{1 + \exp(D[\theta_i - b_j])}$$

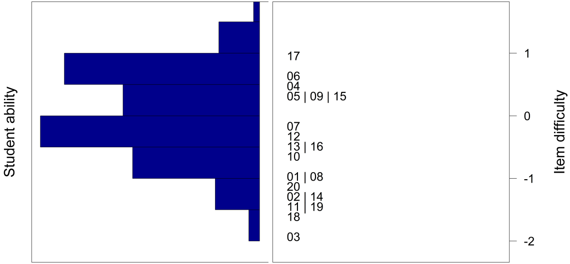
$D = 1.702$ is scaling parameter introduced in order to match logistic and probit metrics very closely (Lord and Novick, 1968)



Note: Probit (normal-ogive) model: $\pi_{ij} = \Phi(\theta_i - b_j)$, where $\Phi(x)$ is a cumulative distribution function for the standard normal distribution.

Item-Person Map (Wright Map)

IRT models allow us to put *items* and *persons* on the same scale

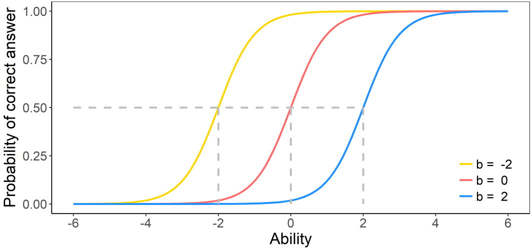


Note: See an example of „32-item test of body height“ (van der Linden, 2017), compare to Figure 2.4

1PL IRT Model

$$\pi_{ij} = P(Y_{ij} = 1 | \theta_i, a, b_j) = \frac{\exp[a(\theta_i - b_j)]}{1 + \exp[a(\theta_i - b_j)]}$$

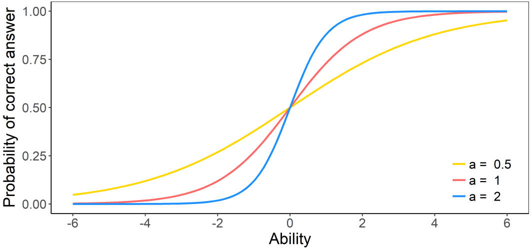
θ_i ability of person i for $i = 1, \ldots, I$
b_j difficulty of item j (location of inflection point) for $j = 1, \ldots, J$
a discrimination common for all items (slope at inflection point)



2PL IRT Model

$$\pi_{ij} = P(Y_{ij} = 1|\theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$$

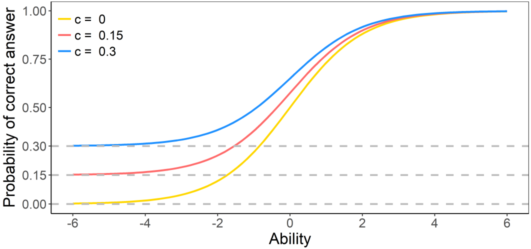
θ_i ability of person i for $i = 1, \ldots, I$
b_j difficulty of item j (location of inflection point)
a_j discrimination of item j (slope at inflection point) for $j = 1, \ldots, J$



3PL IRT Model

$$\pi_{ij} = P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$$

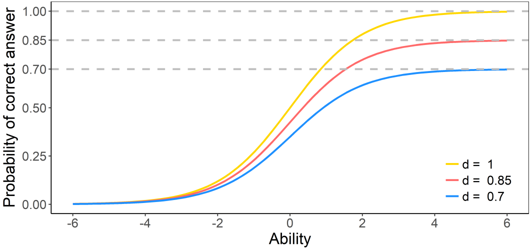
θ_i ability of person i for $i = 1, \dots, I$
b_j difficulty of item j (location of inflection point)
a_j discrimination of item j (slope at inflection point)
c_j pseudo-guessing parameter of item j (lower/left asymptote), $j = 1, \dots, J$



4PL IRT Model

$$\pi_{ij} = P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j, d_j) = c_j + (d_j - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$$

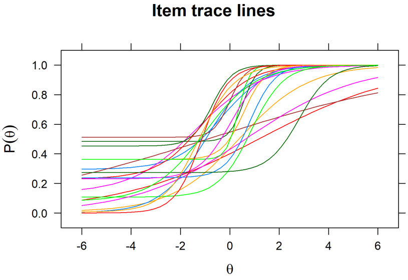
θ_i ability of person i for $i = 1, \ldots, I$
b_j difficulty of item j (location of inflection point)
a_j discrimination of item j (slope at inflection point)
c_j pseudo-guessing parameter of item j (lower/left asymptote)
d_j inattention parameter of item j (upper/right asymptote), for $j = 1, \ldots, J$

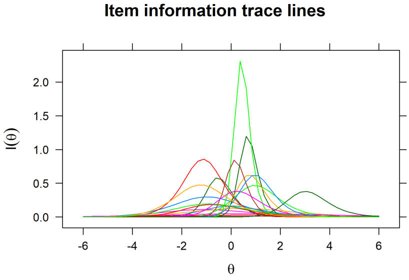


Information Function

$$P(\theta, a_j, b_j, c_j, d_j) = c_j + (d_j - c_j)\frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]},$$

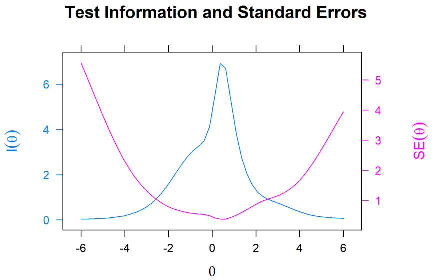
$$\mathrm{I}_j(\theta, a_j, b_j, c_j, d_j) = \frac{bP}{\delta\theta} = a_j(d_j - c_j)\frac{\exp[a_j(\theta - b_j)]}{\{1 + \exp[a_j(\theta - b_j)]\}^2}$$

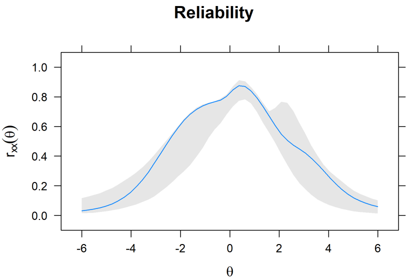




Test Information and Reliability

$$\mathrm{I}(\theta) = \sum_j \mathrm{I}_j(\theta, a_j, b_j, c_j, d_j)$$





Note: Standard error $SE(\hat{\theta}|\theta) = 1/\sqrt{I(\hat{\theta}|\theta)}$

Reliability $SE(\hat{\theta}|\theta) = \sigma\sqrt{(1 - r_{xx}(\hat{\theta}|\theta))}$

Maximum Likelihood Estimation

Once the data have been collected, we can ask: „Which (item/person) parameters would most likely produce these results?“

Estimating ability parameters:

- Assume five items with known item parameters
- Assume response pattern 11000
- Student with what ability is most likely to produce these responses?

Estimating item parameters:

- Assume 20 students with known abilities $\theta_1, ..., \theta_{20}$
- Assume responses to the first item 11000011110101001110
- Item with what difficulty b is most likely to lead to these student responses?

Estimating ability parameter θ

Problem

- Assume five items (obeying Rasch model) with known item parameters $b_1 = -1.90, b_2 = -0.60, b_3 = -0.25, b_4 = 0.30, b_5 = 0.45$.
- Assume *response pattern* 11000.
- How likely is average student ($\theta = 0$) to produce these responses?
- How likely is weaker student ($\theta = -1$) to produce these responses?
- Which student is more likely to produce these responses?

Solution

1. calculate probability for each response in the pattern
2. calculate probability of the response pattern
   - use assumption of conditional independence:
     product of probabilities of individual responses in the pattern

Estimating ability parameter θ

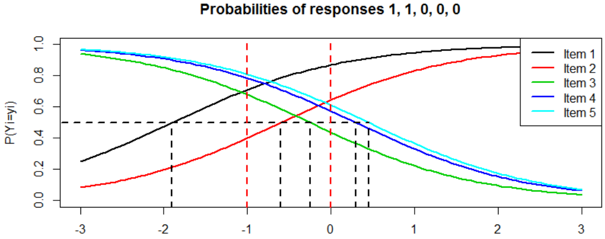
Solution

- $\mathrm{P}\left(Y_1 = 1|\theta = 0\right) = \frac{e^{(0-(-1.9))}}{1+e^{(0-(-1.9))}}, \dots$

- $\mathrm{P}\left(\mathbf{Y} = 11000|\theta = 0\right) =$
$\frac{e^{1.9}}{1+e^{1.9}} \cdot \frac{e^{0.6}}{1+e^{0.6}} \cdot \left(1 - \frac{e^{0.25}}{1+e^{0.25}}\right) \cdot \left(1 - \frac{e^{-0.3}}{1+e^{-0.3}}\right) \cdot \left(1 - \frac{e^{-0.45}}{1+e^{-0.45}}\right) =$

  $= 0.87 \cdot 0.65 \cdot 0.44 \cdot 0.57 \cdot 0.61 = 0.086$

- $\mathrm{P}\left(\mathbf{Y} = 11000|\theta = -1\right) = 0.71 \cdot 0.40 \cdot 0.68 \cdot 0.79 \cdot 0.81 = 0.123$



Estimating ability parameter θ

Problem

- Student with what ability θ is most likely to produce responses 11000?

Solution

1. calculate probability for each response in the pattern (as function of θ)
2. calculate probability of the response pattern (as function of θ)
   - this function is known as **likelihood function** L
   - use assumption of conditional independence:
   - $\mathrm{P}\,(11000|\theta) = L(11000|\theta) = p_1 \cdot p_2 \cdot \ (1 - p_3) \cdot \ (1 - p_4) \cdot \ (1 - p_5)$
3. find the maximum value of the likelihood function

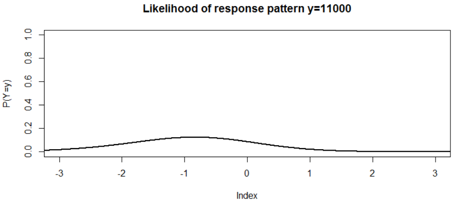
Estimating ability parameter θ

Problem

- Student with what ability θ is most likely to produce responses 11000?

Solution

- $\mathrm{P}\left(\mathbf{Y}=11000|\theta\right)=$
$$\frac{e^{\theta+1.9}}{1+e^{\theta+1.9}}\cdot\frac{e^{\theta+0.6}}{1+e^{\theta+0.6}}\cdot\left(1-\frac{e^{\theta+0.25}}{1+e^{\theta+0.25}}\right)\cdot\left(1-\frac{e^{\theta-0.3}}{1+e^{\theta-0.3}}\right)\cdot\left(1-\frac{e^{\theta-0.45}}{1+e^{\theta-0.45}}\right)$$

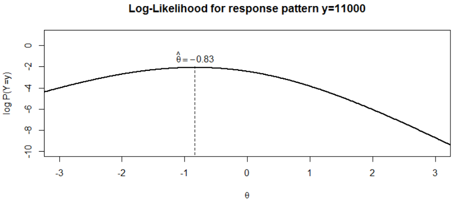


- For which θ is the likelihood the highest?

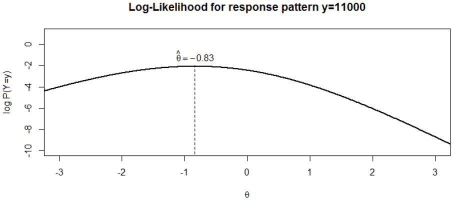
Estimating ability parameter θ

Log-likelihood

- Reaches maximum for the same θ as likelihood
- Easier to handle
- $\log \mathrm{P}\left(\mathbf{Y} = 11000|\theta\right) = \log \frac{e^{\theta+1.9}}{1+e^{\theta+1.9}} + \log \frac{e^{\theta+0.6}}{1+e^{\theta+0.6}} + \log \left(1 - \frac{e^{\theta+0.25}}{1+e^{\theta+0.25}}\right) + \log \left(1 - \frac{e^{\theta-0.3}}{1+e^{\theta-0.3}}\right) + \log \left(1 - \frac{e^{\theta-0.45}}{1+e^{\theta-0.45}}\right)$



Maximum Likelihood Estimation - Technical details



For which θ is the likelihood the highest?

- Empirical MLE
  - method of brackets
  - does not provide with standard error of estimate
- Newton-Rhapson
  - looks for $\log L' = 0$ (zero derivative of $\log L$)
  - uses second derivative $\log L''$ to find it quickly: $\theta_{new} = \theta_{old} - \frac{\log L'}{\log L''}$
  - derivatives can be further used for estimation of item information and standard error

Estimation of item parameters

Problem (Estimating item difficulty b)

- Assuming that person abilities are known, item with what difficulty b is most likely to produce student responses 110010011000?

Solution

1. calculate probability of student response pattern (as function of b)
    - this function is again known as **likelihood function** L
    - use assumption of conditional independence
2. find the maximum value of the likelihood function

Note: For 2PL models likelihood-function is 2-dimensional!

Three types of ML Estimates in IRT models

Usually, both **person** and **item** parameters need to be estimated.

- Joint Maximum Likelihood
  - Used in Winsteps
  - Ping-pong between person and item MLE
  - With increasing number of examinees, number of parameters to be estimated increases
  - May lead to inconsistent, biased estimates
- Marginal Maximum Likelihood
  - Used in IRTPRO, `ltm`, `mirt`
  - Assumes *prior* ability distribution (usually $N(0,1)$)
  - Ability is „integrated out“ to get ML estimates of item parameters
  - *Expected a posteriori* estimates of abilities
- Conditional Maximum Likelihood
  - Used in `eRm`
  - Only applicable in 1PL (Rasch) models, where:
    - Total score is sufficient statistics for ability
    - Percent correct is sufficient statistics for difficulty

Joint Maximum Likelihood

Mathematical and technical details:

$$L = \mathrm{P}\left(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{y_{ij}}\left(1-\pi_{ij}\right)^{(1-y_{ij})}$$

- logarithm simplifies the above expression to sum:

$$lnL = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\cdot ln(\pi_{ij}) + (1-y_{ij})\cdot ln(1-\pi_{ij})$$

- maximization incorporates computation of partial derivatives
- indeterminacy of parameter estimates in the origin and unit
  - person centering
  - item centering

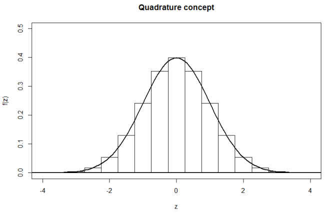
Marginal Maximum Likelihood

Mathematical and technical details:

- marginal likelihood ($\boldsymbol{\theta}$ is integrated out)

$$L = \mathrm{P}\left(\boldsymbol{Y}\right) = \int_{-\infty}^{\infty} \mathrm{P}\left(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\right) \cdot g(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{b}) d\boldsymbol{\theta}$$

- $g(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{b})$ is so called *prior* distribution (usually assumed N(0,1))
- integration solved using *Gauss-Hermite quadrature* (numerical integration)



Can be understood as weighted sum: at each theta interval, the likelihood of response pattern rectangle is weighted by that rectangle's probability of being observed

- L does not depend on $\boldsymbol{\theta}$ and can be maximized with respect to $\boldsymbol{a}, \boldsymbol{b}$

Assessing Ability Levels

Once the item parameters are known (estimated)

- Maximum likelihood estimator (MLE)
- Weighted likelihood
- Bayes model estimator (BME), maximum a posteriori (MAP)
- Expected a posteriori (EAP)

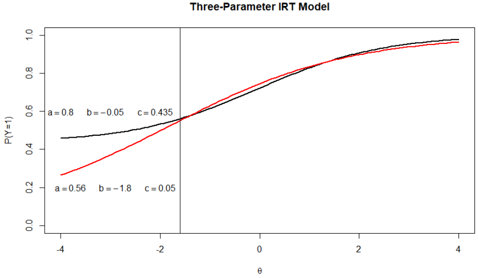
MLE - iterative process

1. Select set of **starting values**
    - randomly or intelligently
    - the closer the starting values are to the actual values the better
2. Maximize the likelihood - get new estimates
3. Check the stopping rule - stop if:
    - maximal number of runs is reached
    - likelihood does not change *too much*

Model selection

- Log-likelihood - the bigger the better
- AIC (Akaike information criterion), BIC (Bayesian information criterion) - the smaller the better
- LRT (likelihood ratio test): if significant $(p < 0.05)$ - submodel is rejected, use model with more parameters

- 3PL model: possibly problems with local maxima, problems to distinguish between models



Item and Person Fit Assessment

- Ames & Penfield (2015)
- Comparing ICC of the fitted model to observed proportion of correct responses
- Detection of improbable response patterns
- Comparing number of respondents with given response pattern to what is expected by the model (X^2 test)

Further Topics

Further models

- Polytomous IRT models (ordinal/nominal)
- Multidimensional IRT models
- Hierarchical IRT models, etc.
- Accounting for Differential item functioning, etc.

Applications

- Test equating
- Computerized adaptive testing, etc.

Vocabulary

- Rasch model, 1PL, 2PL, 3PL, 4PL IRT models
- Item Characteristic Curve (ICC)
- Item Response Function (IRF)
- Item Information Function (IIF)
- Test Information Function (TIF)
- Likelihood function
- Parameter estimation: JML, CML, MML
- Model fit, Item fit, Person fit

Thank you for your attention!

www.cs.cas.cz/martinkova