dortmund university

FAIR

Ministry of Culture and Science of the State of North Rhine-Westphalia

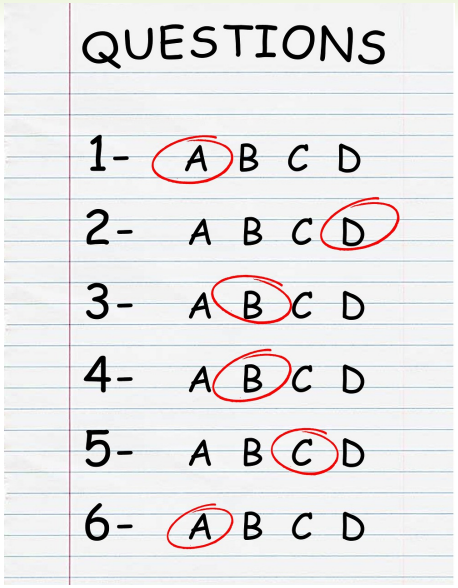
Scalable Learning of Item Response Theory Models

Susanne Frick, **Amer Krivošija** and **Alexander Munteanu**,
TU Dortmund University

AISTATS, Valencia, Spain, May 2, 2024

Motivation

Examinees $j \in [n]$ solve an exam with Items $i \in [m]$



What is given?

For each (i, j): a label if the item was correctly answered?

IRT goal: Estimate the latent item parameters and the latent examinee ability parameter, based on the given input.

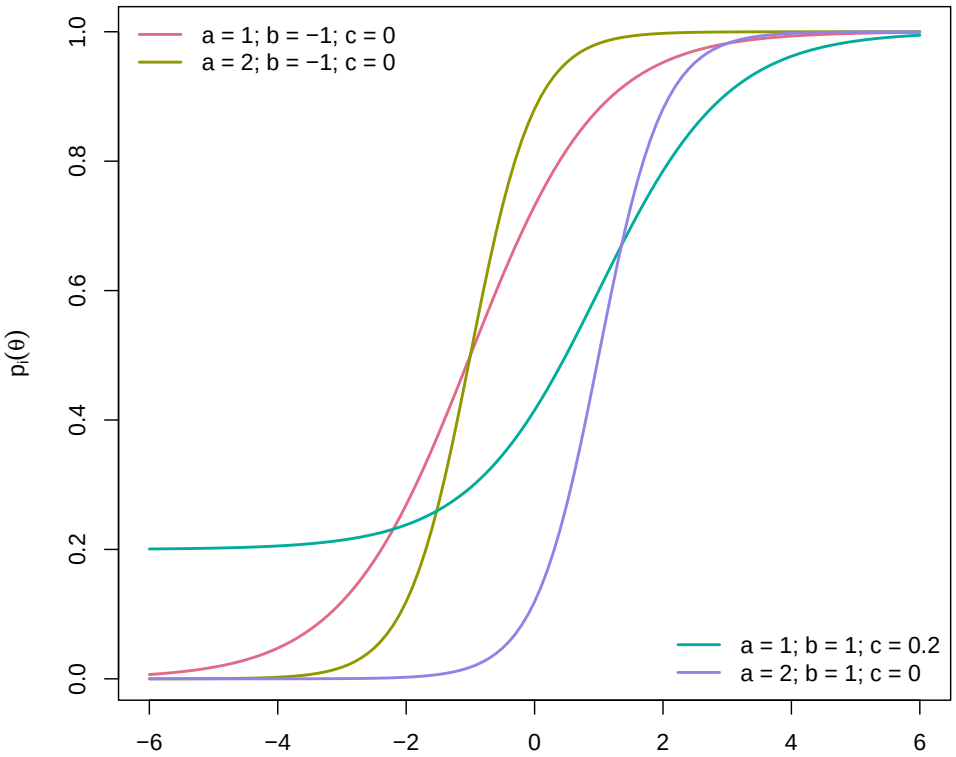
The number of examinees/items in real-world problems can be large:

PISA:	$n \approx 600\,000$	$(m \approx 30)$	AutoML:	$m \approx 5\,000$	$(n \approx 150)$
SHARE:	$n \approx 140\,000$	$(m \approx 10)$			
NEPS:	$n \geq 15\,000$	$(m \approx 100)$			

Storing labels for $(n, m) = (500\,000, 5\,000)$ requires ca. 6GB

Our goal: scale down the input while preserving the accuracy of learned models.

IRT models 2PL / 3PL



Item Characteristic Curve (ICC): Probability of solving item i, with ability θ_j

$$p_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp(-a_i\theta_j + b_i)}$$

Item parameters in 3PL model:

- a_i: discrimination ($a_i > 0$)
- b_i: difficulty
- c_i: guessing ($c_i = 0$ in 2PL)

IRT learning framework

General algorithmic framework for non-convex IRT problem:

Alternating (convex) optimization conditioned on one parameter set:

1. Initialize all latent parameters.
2. While termination condition not satisfied:
   a) Learn the ability parameters, given fixed item characteristics.
   b) Learn the item characteristics, given fixed ability parameters.

Coresets

Given:
- an input set $X \in \mathbb{R}^{n \times d}$
- a variable $\eta \in \mathbb{R}^d$, and
- loss function $f(X\eta) : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ to minimize.

A significantly smaller $K \subseteq X$ (with corresponding weights $w \in \mathbb{R}^{|K|}$) is an ε-coreset of X, if

$$|f(X\eta) - f_w(K\eta)| \leq \varepsilon f(X\eta)$$

for all $\eta \in \mathbb{R}^d$.

Beyond worst case

Problem: IRT problems $\supseteq$ logistic regression. No sublinear coresets for logistic regression in general [MSSW18].

Solution: We extend a data-dependent μ parameter from [MSSW18]: for $p \in \{0, 1\}$, the input $X = Y \odot A^T$ is μ-complex if $\exists \mu < \infty$, s.t.:

$$\mu_p(X) = \sup_{\eta \in \mathbb{R}^d \setminus \{0\}} \frac{\sum_{x_i \eta \geq 0} |x_i \eta|^p}{\sum_{x_i \eta < 0} |x_i \eta|^p} \leq \mu.$$

Typically $\mu \in O(1)$. When model specifications are violated, μ becomes large.

Common coreset construction: sensitivity framework [LS10]

Problem: X changes in each iteration, since labels Y are different for each item.

Lemma: For any $X \in \mathbb{R}^{m \times n}$, the sensitivity scores of X and DX are the same, where $D \in \mathbb{R}^{m \times m}$ is an arbitrary diagonal matrix with ± 1 entries.

$\Rightarrow$ We can construct only one coreset, valid for all items – within one iteration!

[LS10] Langberg, Schulman, SODA 2010

[MSSW18] Munteanu, Schwiegelshohn, Sohler, Woodruff, NeurIPS 2018

Coresets for 2PL (reduction to multiple logistic regression)

$A = [\alpha_i]_{i\in[m]}$, $B = [\beta_j]_{j\in[n]}$: item, ability param. $Y = (y_{ij})_{\substack{i\in[m]\\ j\in[n]}}$ labels ($\in \{\pm 1\}$).

We fix one set and minimize the other in:

$$f(A \mid B) = \sum\nolimits_{i\in[m],j\in[n]} \ln(1+\exp(-Y_{ij}\alpha_i^T\beta_j)) = f(B \mid A),$$

i.e. given fixed B, we write $x_j = -Y_{ij}\beta_j^T$, $X_{(i)} = (x_j)_{[n]}$, $\forall i \in [m]$:

$$\min\nolimits_{\alpha_i\in\mathbb{R}^2} f(\alpha_i \mid B) = \min\nolimits_{\alpha_i\in\mathbb{R}^2} \sum\nolimits_{j\in[n]} \ln(1+\exp(x_j\alpha_i)).$$

Theorem: 2PL Coresets

Let $X_{(i)}$be μ-complex, for each $i \in [m]$. There exists $K \subseteq X$ of size

$$|K| \in \tilde{O}((\mu^3/\varepsilon^4)\cdot(\log(n)^4 + \log(m)),$$

that is a $(1+\varepsilon)$-coreset simultaneously for all $X_{(i)}$, $i \in [m]$ for the 2PL IRT problem. The construction time is $\tilde{O}(n)$.

Theorem holds verbatim for fixed A, minimizing $f(\beta_j \mid A)$, $\forall j \in [n]$.

Coresets for 3PL (multiple non-convex sub-problems)

A, B, C: item, ability, guessing parameters. $Y = (y_{ij})_{i\in[m],j\in[n]}$ labels ($\in \{\pm 1\}$). Given fixed A and C, we write $x_i = -Y_{ij}\alpha_i^T$, and minimize $\forall j \in [n]$:

$$f(\beta_j \mid A, C) = \sum_{i\in[m]}^{[Y_{ij}=-1]} \ln\left(\frac{1+e^{x_i\beta_j}}{1-c_i}\right) + \sum_{i\in[m]}^{[Y_{ij}=1]} \ln\left(\frac{1+e^{-x_i\beta_j}}{1+c_i e^{-x_i\beta_j}}\right).$$

First $\sum$ logistic: X', second $\sum$ sigmoid: X''. Additional μ-assumption needed.

Theorem: 3PL Coresets

Let $X_{(j)}$ be given, and X', X'' be its submatrices, both μ-complex, for each $j \in [n]$. There exists $K \subseteq X$ of size

$$|K| \in O((\mu^2/\varepsilon^2) \cdot \sqrt{m} \cdot (\log(m)^2 + \log(n))),$$

that is a $(1+\varepsilon)$-coreset simultaneously for all $X_{(j)}$, $j \in [n]$ for the 3PL IRT problem. The construction time is $O(m)$.

Theorem holds similarly for fixed B, minimizing $f(\alpha_i, c_i \mid B)$, $\forall i \in [m]$.

Experiments I

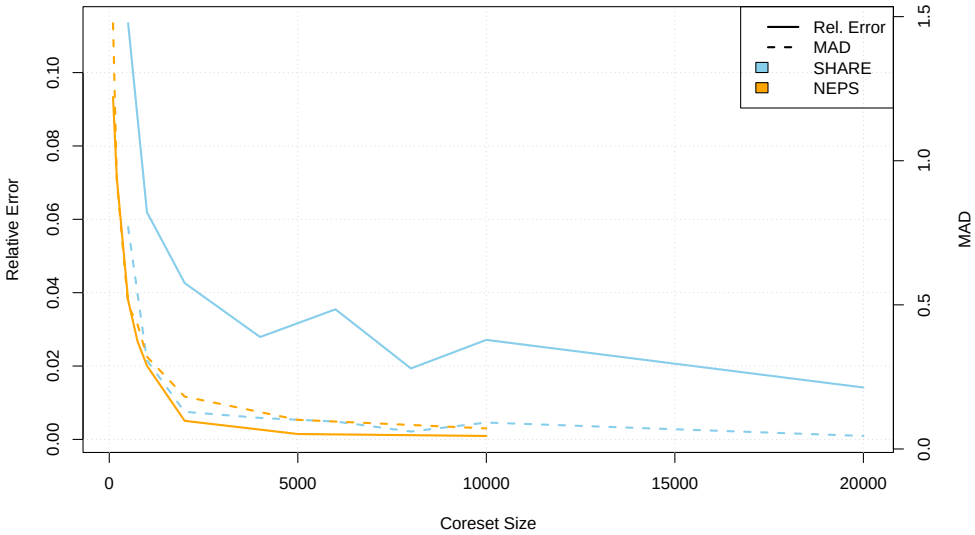
Parameter estimation guarantee

Let

- X: any matrix that satisfies the conditions of main coreset theorems.
- K: an ε-coreset for X.
- η_{opt}, η_{core}: argmin $f(X\eta)$, $f_w(K\eta)$.

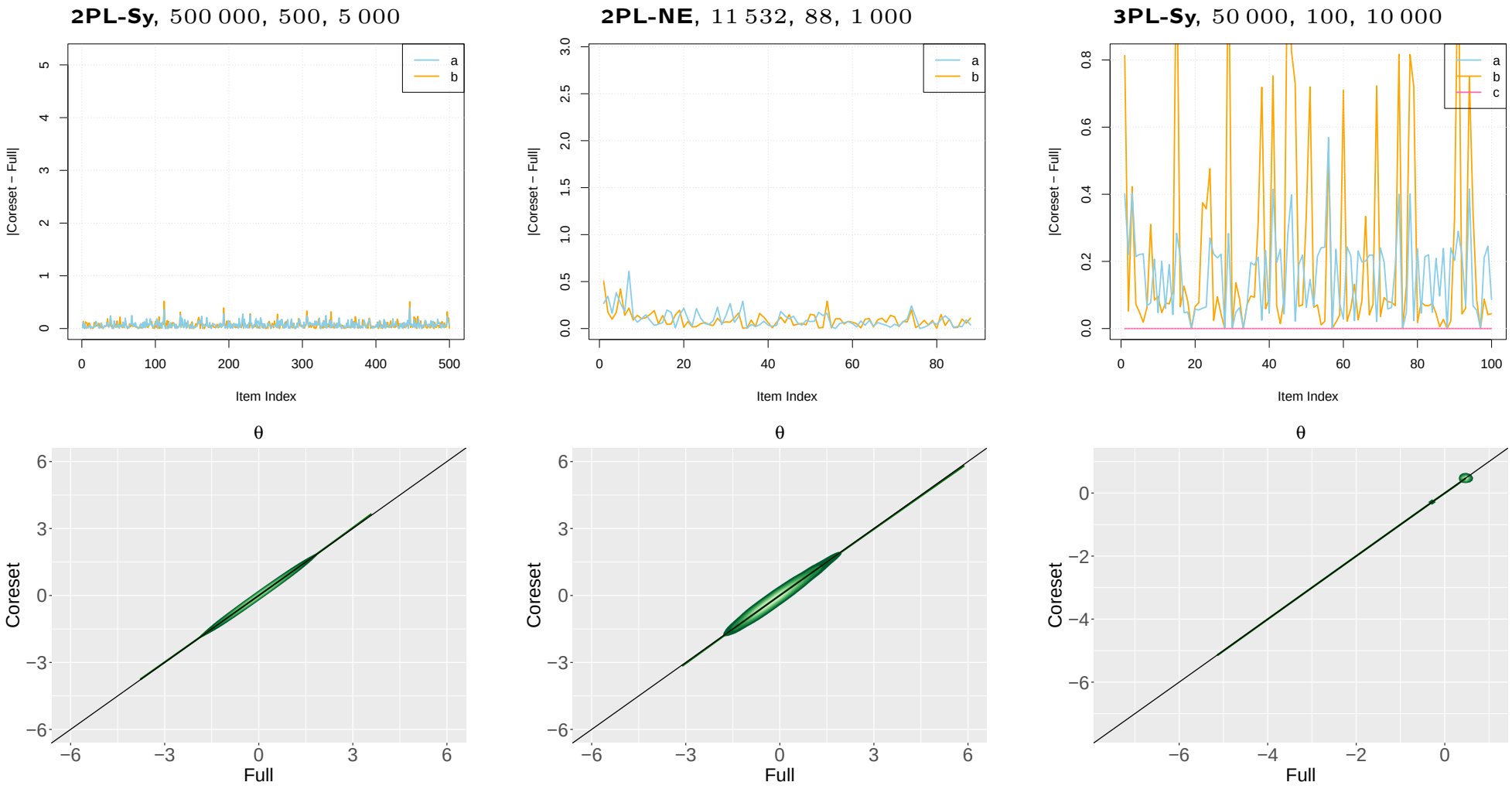
Then
the optimal solutions for the τ-PL problem, for $\tau \in \{2, 3\}$, satisfy

$$\|\eta_{\text{opt}} - \eta_{\text{core}}\|_1 \leq O(\mu^{\tau-1}) \cdot f(X\eta_{\text{opt}}).$$



2PL Experiments on real world SHARE and NEPS data: Coreset sizes vs. relative error and mean absolute deviation (MAD)

Experiments II



Parameter estimates for the coresets vs. the full data sets. The first row: the bias for the item parameters a, b and c. The second row: a kernel density estimate for the ability parameters θ, standardized to zero mean and unit variance, with a LOESS regression line in dark green.

Experiments III

data, n, m, k	$\text{mean}_{\text{full}}$	$\text{mean}_{\text{core}}$	gain (%)	r.err. $\hat{\varepsilon}$	mad(α)	mad(θ)
2PL-Sy, 50 000, 500, 500	136.98	45.55	66.749	0.04803	0.525	0.008
2PL-Sy, 100 000, 200, 1 000	122.25	61.46	49.727	0.03404	0.379	0.008
2PL-Sy, 500 000, 500, 5 000	1 278.85	591.88	53.718	0.01445	0.171	0.001
2PL-Sy, 500 000, 5 000, 5 000	9 363.75	5 536.68	40.871	0.00076	0.120	0.013
2PL-SH, 138 997, 10, 8 000	28.85	27.64	4.216	0.01935	0.061	0.007
2PL-NE, 11 532, 88, 1 000	5.97	4.01	32.829	0.02007	0.320	0.045
3PL-Sy, 50 000, 100, 10 000	211.47	93.78	55.653	0.00212	0.384	0.010
3PL-Sy, 50 000, 200, 10 000	369.82	145.67	60.609	0.02186	0.488	0.001
3PL-Sy, 200 000, 100, 10 000	893.18	196.80	77.966	0.01789	0.524	0.003

Mean running times (in minutes), across 20 repetitions (of 50 iterations of the main loop) per data set 2-/3-PL, (Sy)nthetic, SH(ARE), NE(PS). m: number of items, n: number of examinees, k: coreset size. The (relative) gain: $(1 - \text{mean}_{\text{core}}/\text{mean}_{\text{full}}) \cdot 100\,\%$.

Let f_{full} and f_{core} be optimal objective values. Relative error: r.err. $\hat{\varepsilon} = |f_{\text{core}} - f_{\text{full}}|/f_{\text{full}}$.

Mean Absolute Deviation: $\text{mad}(\alpha) = \frac{1}{n}\sum(|a_{\text{full}} - a_{\text{core}}| + |b_{\text{full}} - b_{\text{core}}| + |c_{\text{full}} - c_{\text{core}}|)$;

$\text{mad}(\theta) = \frac{1}{m}\sum|\theta_{\text{full}} - \theta_{\text{core}}|$, evaluated on the parameters attaining the optimal f_{full} and f_{core}.

Some open questions

- How to extend our approach to other IRT models, e.g. (ordered) categorical, continuous, multidimensional, and multilevel IRT models?
- Is it possible to sketch the stream of items/examinees?
- Incorporate coresets into state-of-the-art IRT solvers (e.g. `mirt`)?
- How to deal with known numerical issues with 3PL models?