2017 IACAT Conference

Advancing assessment through CAT

18-21 August 2017, Niigata, Japan

Benesse Corporation

PACIFIC METRICS™ CORPORATION

Graduate Management Admission Council®

Multistage Testing using Determinantal Point Processes

Jill-Jênn Vie

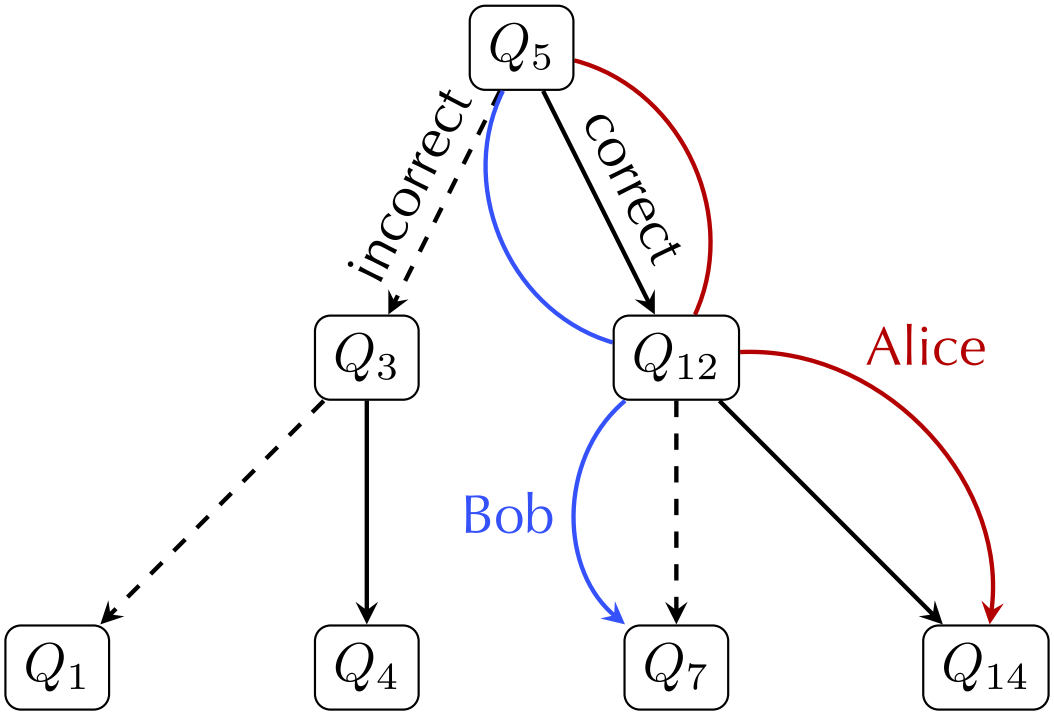
RIKEN Center for Advanced Intelligence Project

Tokyo, Japan

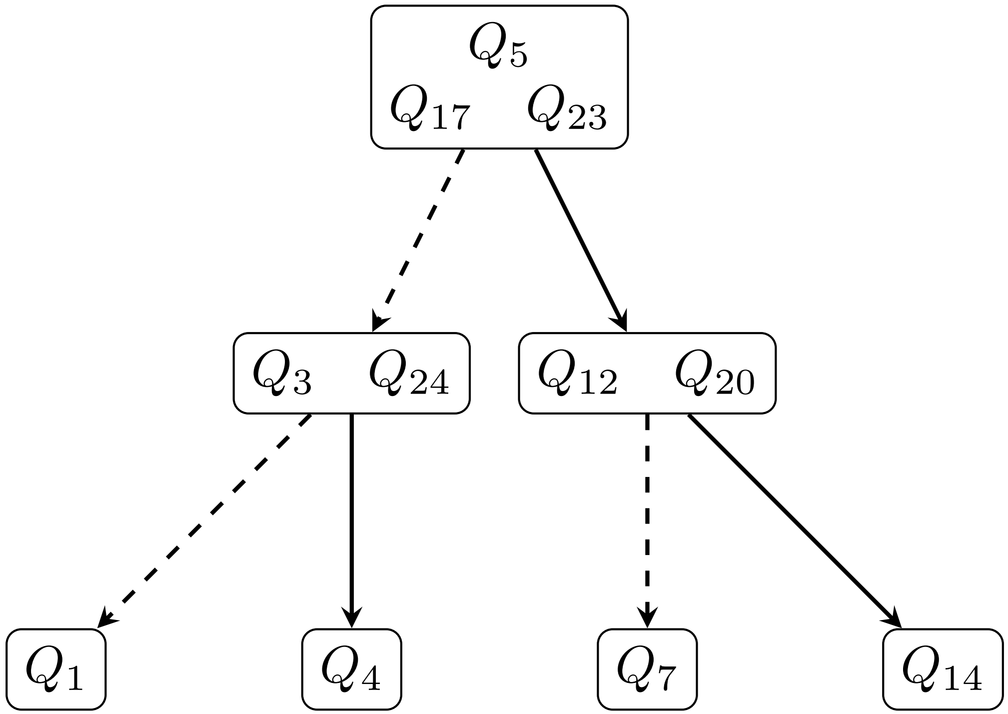
Context

- Many similarities between machine learning and psychometric models or concepts
  - MIRT ↔ Matrix factorization with a sigmoid link
  - Calibration ↔ **Feature extraction**
  - Control item exposure ↔ Exploitation-exploration
- Why not bring some ideas from ML to CAT? (also: the other way)

Adaptive Testing



Multistage Testing



Motivation

- In a CAT, we want to ask questions that minimize the uncertainty over the examinee parameters
- But the first MLE of the examinee parameters is hard to obtain
  - If $dim = 1$ (Rasch), examinee should at least fail one item and succeed at another item
  - If $dim > 1$ (MIRT) the MLE is less likely to exist
- Chalmers (2016) suggests doing a pre-CAT
- How to choose the **very first** items?

Talk

- Introduce MIRT
- Visualize the items
- Present a measure of diversity
- Apply it → selection of the first items of a MST
- Experiments and results

- Disclaimer: this presentation is not adaptive ☹
  - Sequence of slides will be the same for everyone

Assumptions

- Dichotomous data D: right/wrong answers
  - $D_{ij} = 1$: "Examinee i answers correctly item j"
- MIRT model
  - Examinee i : ability $\theta_i \in \mathrm{R}^d$
  - Item j : discrimination $a_j \in R^d$ and easiness $b_j \in R$

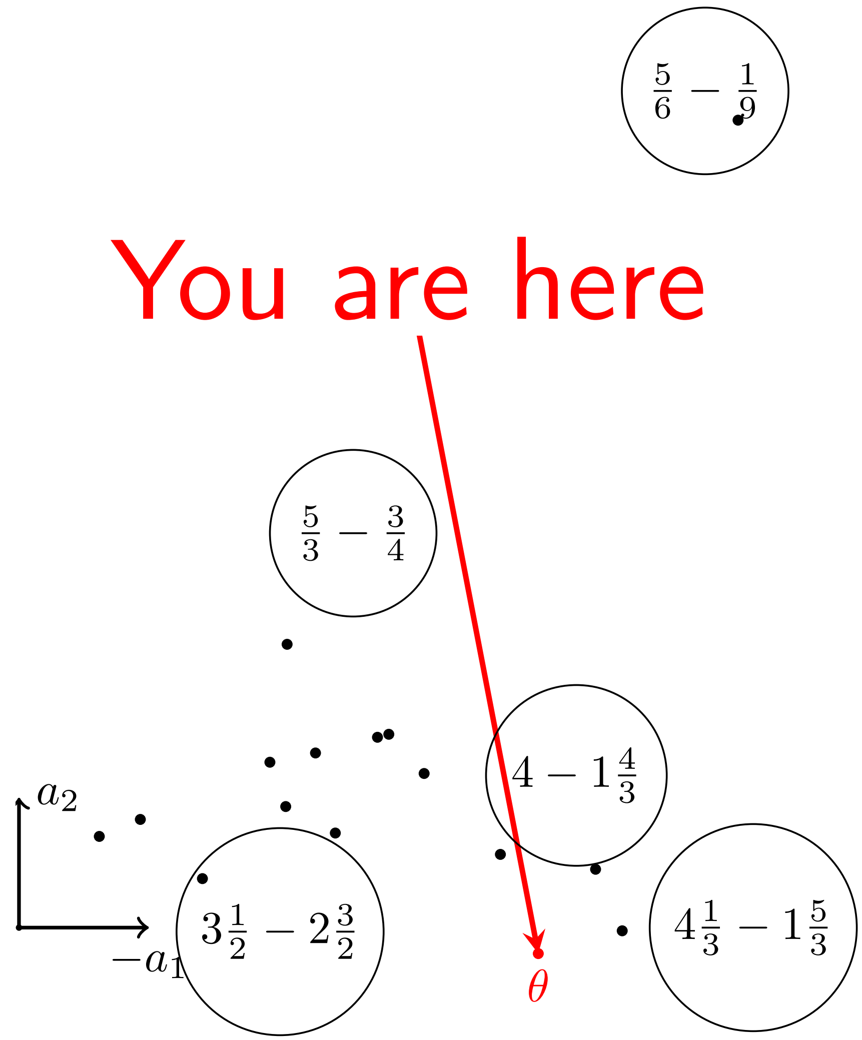
$$\Pr\left(D_{ij} = 1\right) = \frac{1}{1 + e^{-\left(a_j^T \theta_i + b_j\right)}}$$

Ex: Fraction Subtraction data

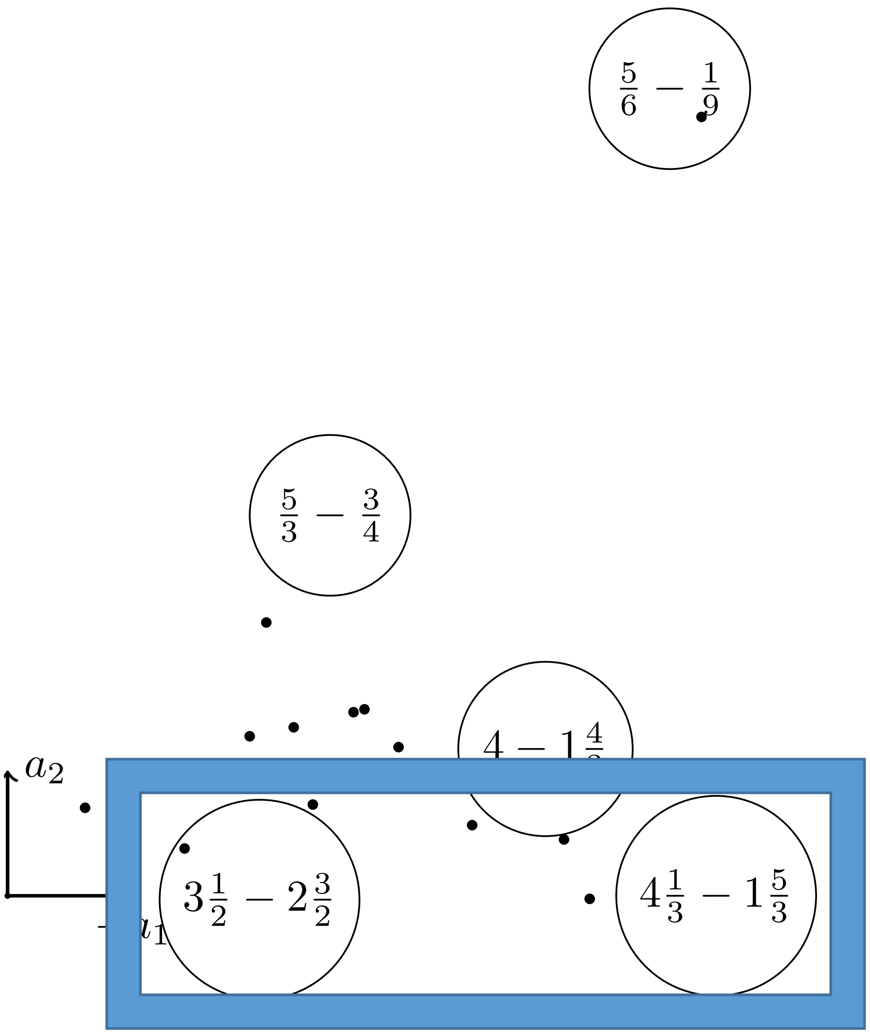
- 536 middle-school examinees
- 20 fraction subtraction items (DeCarlo, 2010)
- Calibrate 2-dim MIRT model: for each item,
  - a_1, a_2 discrimination along the 2 dimensions
  - b the easiness
- Population has a prior $\sim \mathcal{N}(0, I)$

```
library('mirt')
library('CDM')
fit = mirt(fraction.subtraction.data, 2)
coef(fit)
```

Discrimination plot



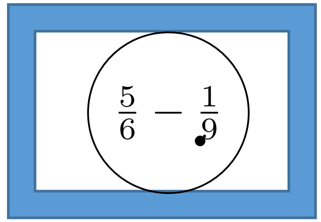
Interpreting components



Items that discriminate only over one dimension

$3\frac{1}{2} - 2\frac{3}{2}$	$4\frac{1}{3} - 2\frac{4}{3}$	$4\frac{1}{3} - 1\frac{5}{3}$
$b = 0.13$	$b = -0.46$	$b = -1.99$
$-a_1 = 2.01$	$-a_1 = 4.65$	$-a_1 = 5.66$
$a_2 = -0.03$	$a_2 = -0.02$	$a_2 = 0.00$

Interpreting components



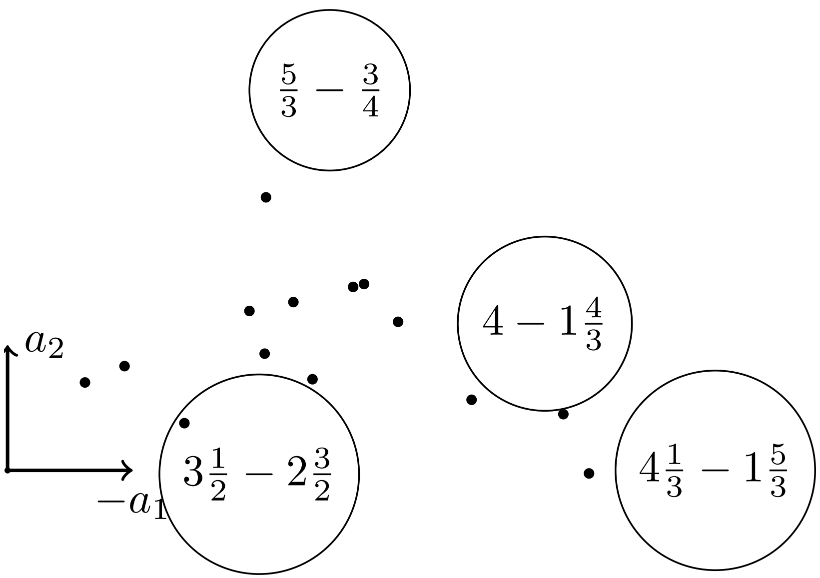
Items that highly discriminate over both dimensions

$\frac{3}{4} - \frac{3}{8}$ $\quad$ $\frac{5}{6} - \frac{1}{9}$

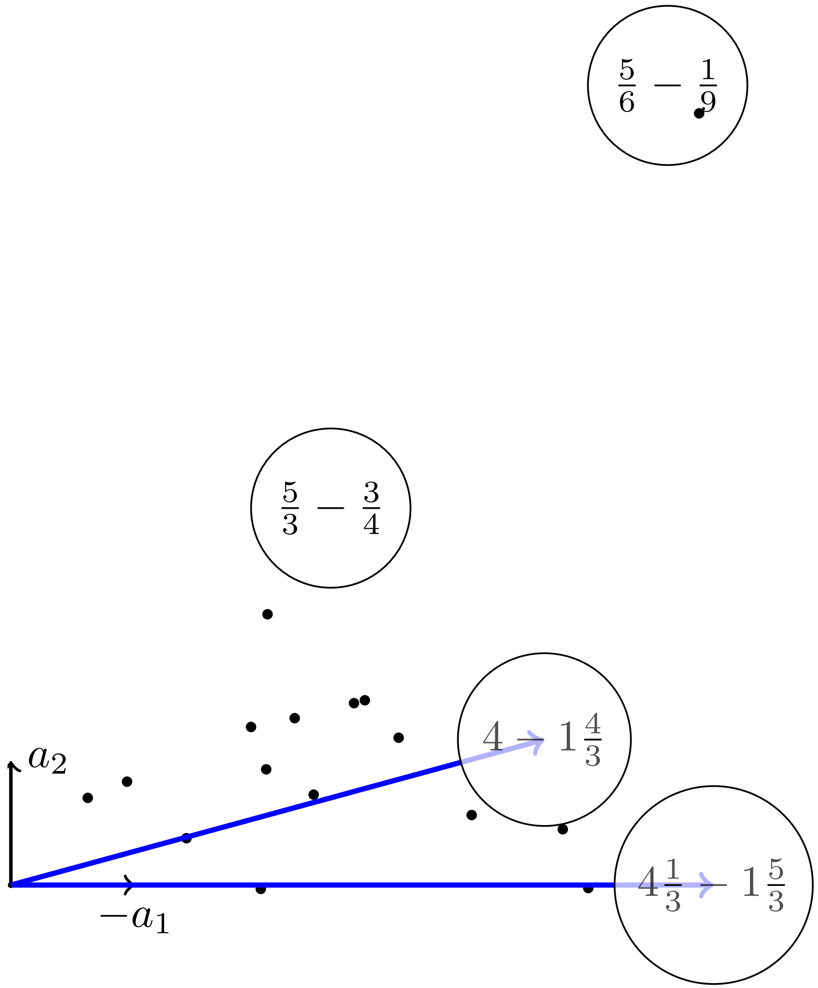
$b = 1.09$ $\quad$ $b = -0.28$

$-a_1 = 5.54$ $\quad$ $-a_1 = 5.29$

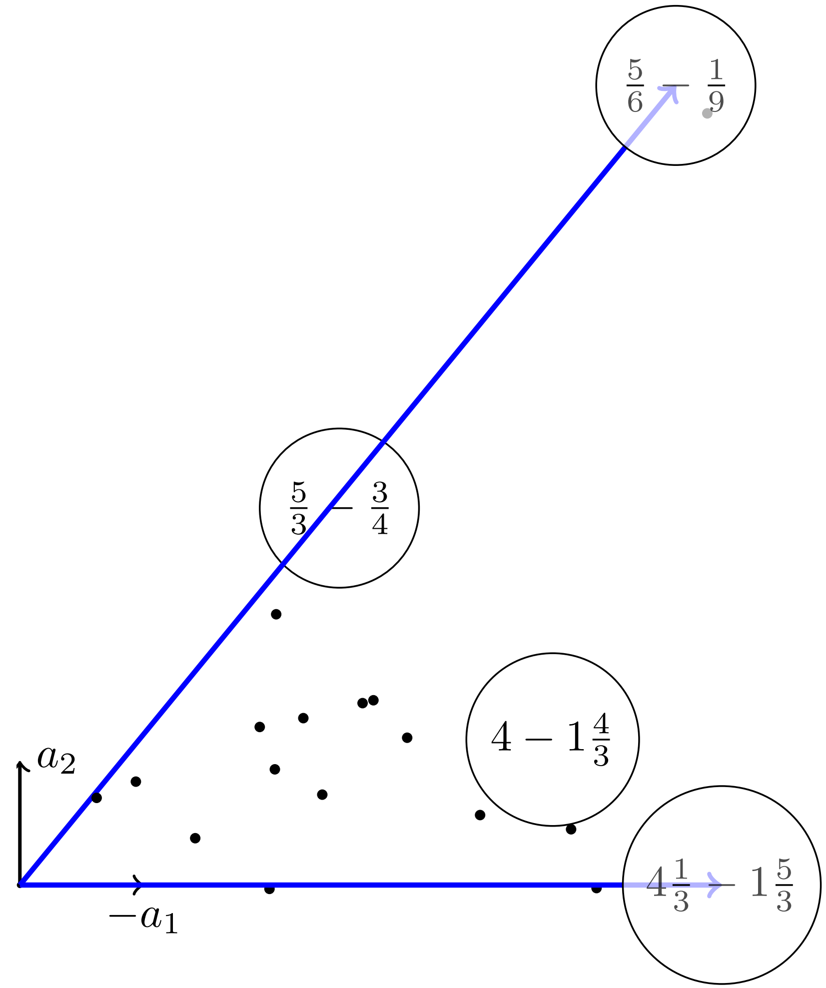
$a_2 = 6.22$ $\quad$ $a_2 = 6.44$



Is this a good choice of items?



Is this a good choice of items?



Intuition

- Items with close parameters receive similar response patterns (columns)
- In order to maximize response pattern diversity, we should present items of which parameters are least correlated to each other
  - = of which the volume spanned is high

Geometry

- If we have n vectors $V_1, \ldots, V_n$
- And I is a subset of $\{1, \ldots, n\}$
- Let V_I denote the matrix of rows $\{V_i | i \in I\}$
- Then the volume spanned by rows of V_I is:

$$Vol(V_I) = (\det V_I V_I^T)^2$$

Application:

$V_j = \left(a_{j1}, \ldots, a_{jd}, b_j\right)$ parameters of item j

Determinantal Point Processes

- Stochastic process that samples subsets
  - Diverse subsets have higher probability to be drawn
  - A DPP samples a subset S such that for all set I:

$$\Pr(I \subset S) \propto \det V_I^T V_I = \sqrt{Vol(V_I)}$$

- Benefits:
  - k items of n can be drawn with probability $O(k^3 n)$
    - (after a diagonalization of $O(n^3)$ computed once)
  - Random process so lower item exposure
- Applied to machine learning problems (Kulesza & Taskar, 2012)

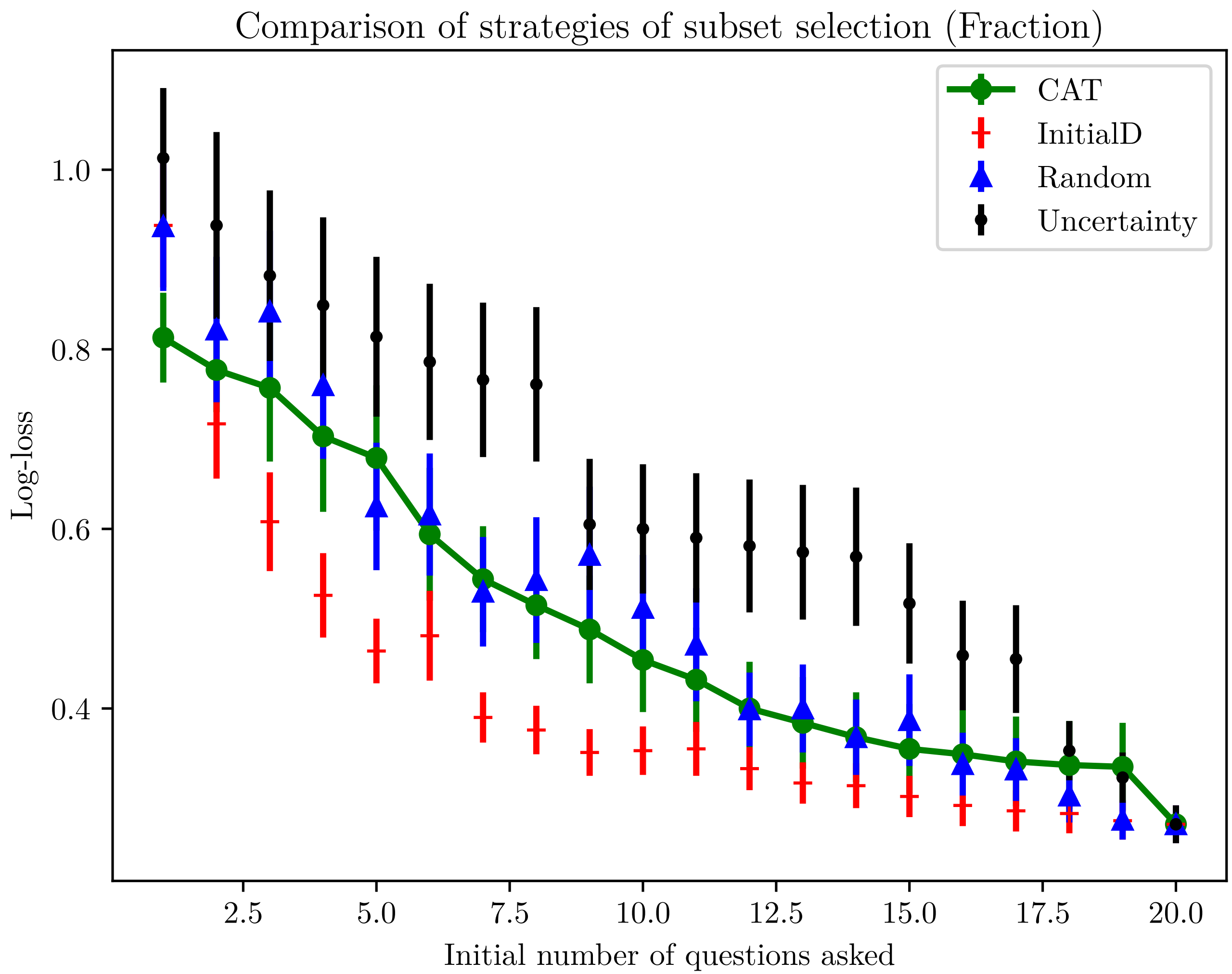
Study

- Compare:
  - CAT with D-optimality criterion
  - Random selection in MST
  - InitialD: DPP selection in MST
- Performance at predicting examinee responses
  - Metric: log-loss (negative log-likelihood)
- On two datasets:
  - Fraction: 536 students, 20 items, 8 skills
  - TIMSS: 757 8th-graders, 23 Maths items, 13 skills

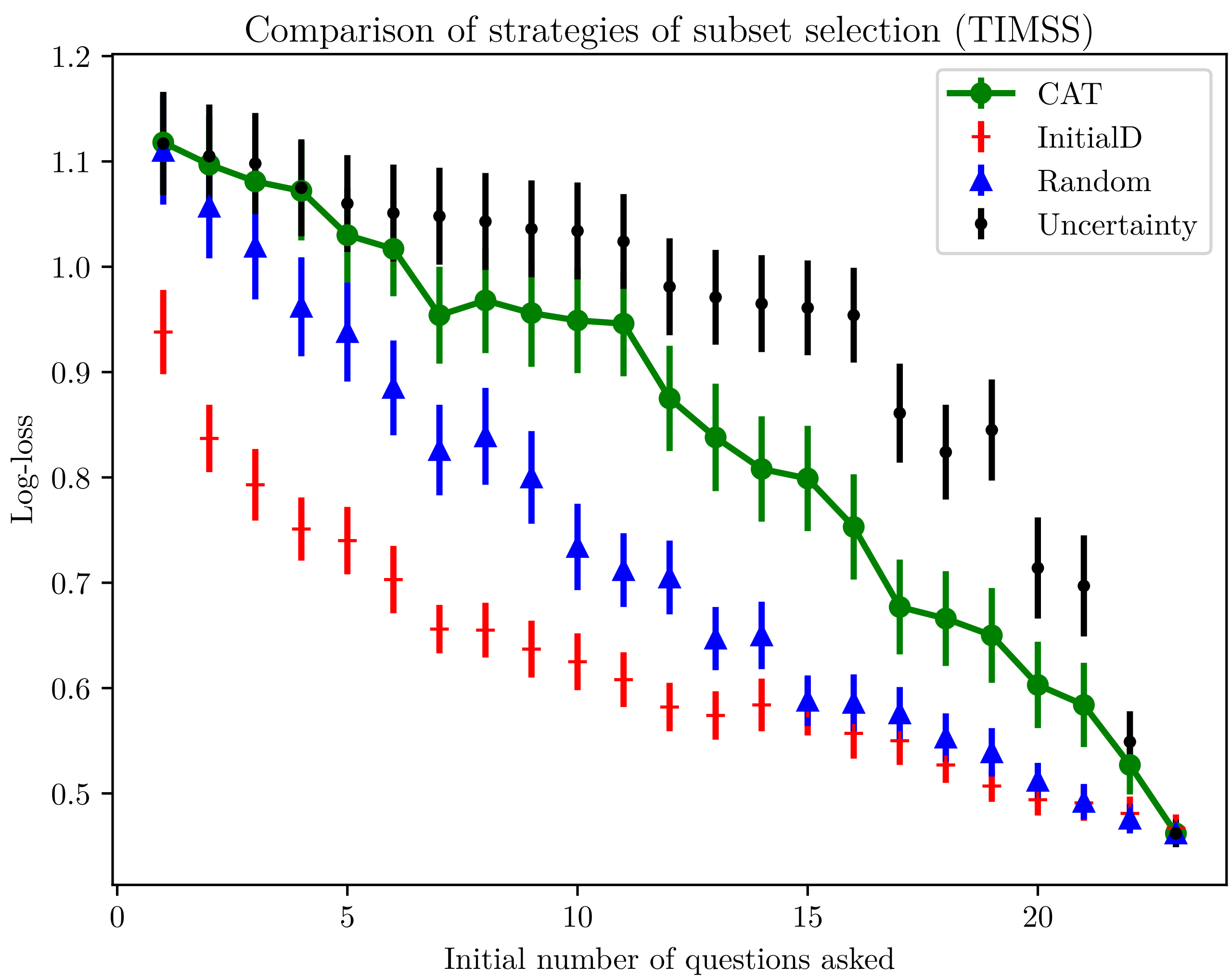
Experiment

- Q-matrix:
  - $q_{jk} = 1$ iff item j involves skill k
- For each dataset with q-matrix:
  - Train a MIRT model on 80% data + extra constraint:
    - If item j does not require skill k ($q_{jk} = 0$) then $a_{jk} = 0$
    - Equivalent to the General Diagnostic Model (Davier, 2005)
  - Get item parameters $V_j = \left(a_{j1}, \dots, a_{jd}, b_j\right)$ of item j
  - For each user in the remaining 20% data:
    - Sample items according to each criterion
    - Get MAP estimate of examinee parameters
    - Compute error (log-loss) of predictions

Results on Fraction data



Results on TIMSS data



Conclusion

- Volume is a measure of diversity that can be used for the first stage of MST
  - Determinantal point processes can sample diverse items efficiently
- Further work: can it be useful for later steps?
  - Can it help teachers build their modules?

Thank you!

Jill-Jênn Vie jilljenn.github.io (+ code)

- Kulesza and Taskar (2012). "Determinantal Point Processes for Machine Learning." *Foundations and Trends® in Machine Learning* 5.2–3: 123-286.
- Vie et al. (2016) "Adaptive Testing with a General Diagnostic Model". In : *11th European Conference on Technology Enhanced Learning (EC-TEL 2016).* Springer, 331–339.
- Chalmers (2016). "Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications." *Journal of Statistical Software* 71.5: 1-38.