Clustering Students and Inferring Skill Set Profiles with Skill Hierarchies

Alan Mishler (amishler@stat.cmu.edu), Rebecca Nugent (rnugent@stat.cmu.edu)

Department of Statistics & Data Science, Carnegie Mellon University

Cognitive Diagnosis Models (CDMs)

Class of models used to estimate students’ mastery of target skills in a learning environment.

Parameters

- $N$ students, $K$ skills, $J$ items
- $Q: J \times K$ binary skill coding matrix $Q_{jk} = 1$ if item $j$ requires skill $k$
- $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik}) \in \{0, 1\}^K$: latent skill profile $\alpha_i = 1$ if student $i$ has mastered skill $k$
- Others, depending on model.

Data

- $Y$: $N \times J$ binary response matrix $Y_{ij} = 1$ if Student $i$ got item $j$ correct

Typical models

- $P(y_{ij} = 1) = (1 - s_j)^{\eta_i} k_{ij}^{1-\eta_i}$ (DINA)
- $P(y_{ij} = 1) = \prod_{k=1}^{K} (1 - s_k)^{\eta_i} g_{jk}^{1-\eta_i}$ (NIDA)

Estimand

- Skill profile $\alpha_i$, for student $i = 1, \ldots, N$.

Estimation

- Likelihood-based: consistent but intractable for large $K$ or $N$
- Pseudo-profiles + clustering $[1, 2]$; fast, consistent under strong assumptions

Typical clustering assumption: All $2^K$ skill profiles are present $[1]$ or possible $[2]$ in the sample.

Research question: How can we optimally perform clustering when
- Some profiles are known to be impossible?
- Not all possible profiles occur in the sample?

Skill Hierarchies

(a) Linear; (b) Convergent; (c) Divergent; (d) Unstructured $[3]$

- Upstream skills must be learned first.
- Number of possible profiles $L_h$ varies by hierarchy.

Simulations

1. Generate data:
   - 30 items, 6 skills, 250 students
   - $Q$-matrix: 30-60-10% of items requiring 1, 2, 3 skills
   - Models: DINA and NIDA
   - Hierarchy types: (a)-(d) + unstructured (no hierarchy)
   - Profiles in sample: 0-100% of possible profiles

2. Compute pseudo-profiles (capability scores):
   - For student $i$, score $W_i = (W_{i1}, W_{i2}, \ldots, W_{ik})$;
     where $W_k = \sum_{j=1}^{J} Y_{ij} Q_{jk}$.

3. Cluster, with $L_h$ clusters
   - Algorithms: (1) Hierarchical clustering (HC) with complete linkage; (2) k-means; (3) empty k-means (up to $L_h$ clusters) $[2]$; (4) semisupervised clustering
   - Starting centers: (1) Random; (2) rescaled $[2]$; (3) pseudocenters (mean capability scores for each possible profile from a separate pseudosample, generated via a DINA or NIDA)

4. Evaluate clusters
   - Adjusted Rand Index (ARI) of best assignment of clusters to profiles.

Results: DINA

Results: NIDA

Future Directions

- Further investigate pseudocenters
- Robust to “misspecification” with other CDMs?
- Investigate fluctuations in null and unstructured hierarchies
- What kinds of profiles are easy to distinguish?
- Soft hierarchical constraints
- Ways to infer skill hierarchy when unknown or partially known