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
  $\alpha_{ik} = 1$ 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

$$P(y_{ij} = 1) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \qquad \text{(DINA)}$$

$$P(y_{ij} = 1) = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}]^{q_{jk}} \qquad \text{(NIDA)}$$

$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = 1$ if student i has mastered all the skills necessary for item j.

$s_j, s_k, g_j, g_k \in [0,1]$: slip and guess parameters on items and skills.

Estimand

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

Estimation

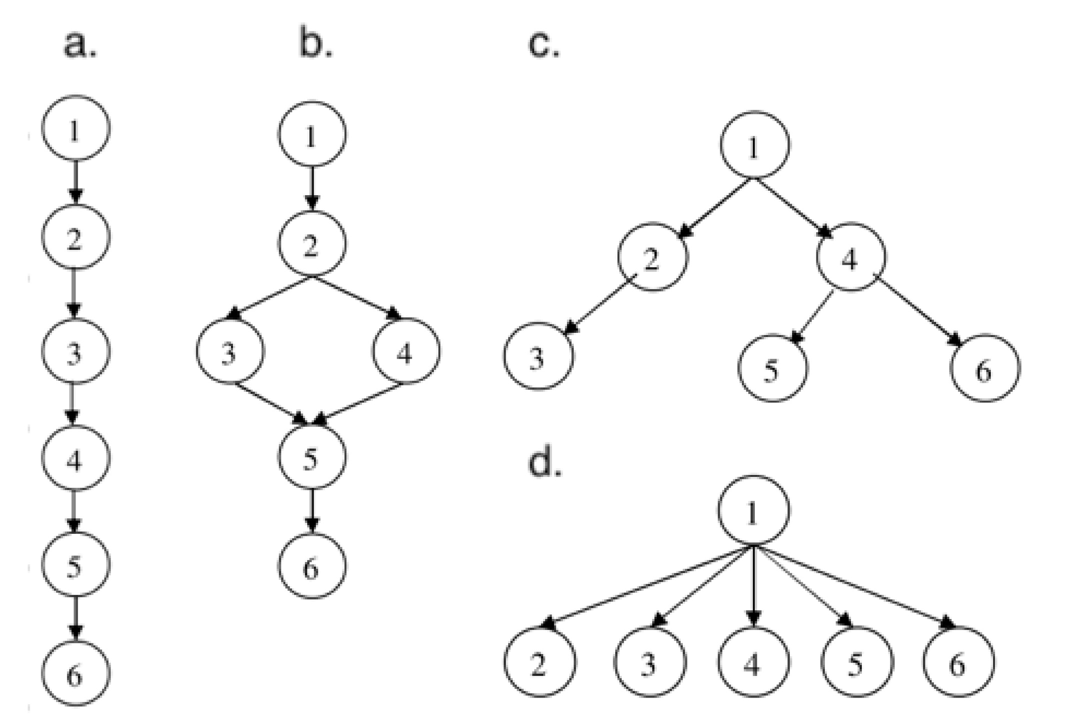
- Likelihood-based:
  consistent but **intractable for large K or N**
- Pseudo-profiles + clustering *[1, 2]*:
  fast, consistent under strong assumptions

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

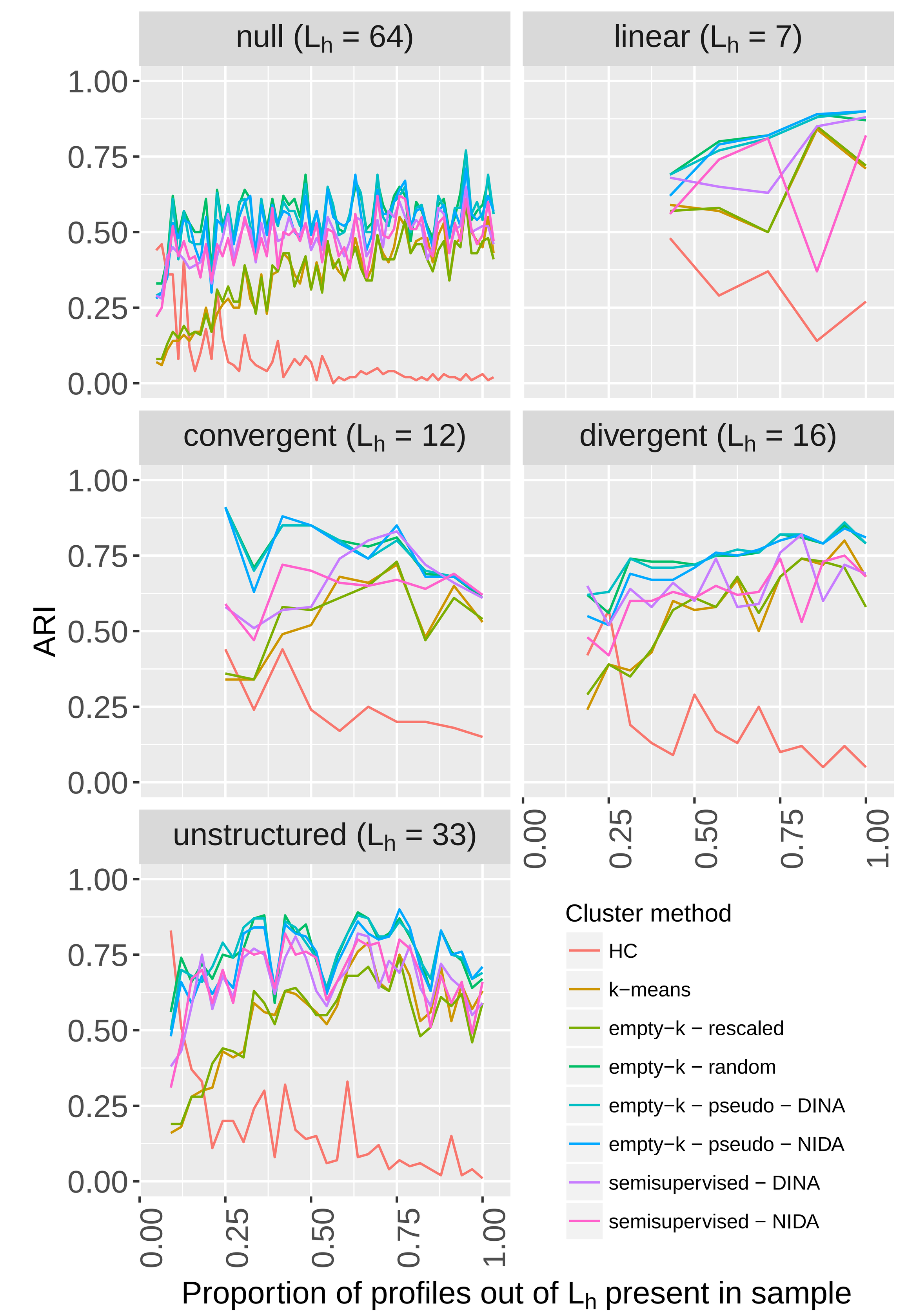
2. Compute pseudo-profiles (capability scores):

- For student i, score $\mathbf{W}_i = (W_{i1}, W_{i2}, \ldots, W_{iK})$, where $W_{ik} = \sum_{j=1}^{J} Y_{ij} Q_{jk}$.

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)

Results: DINA



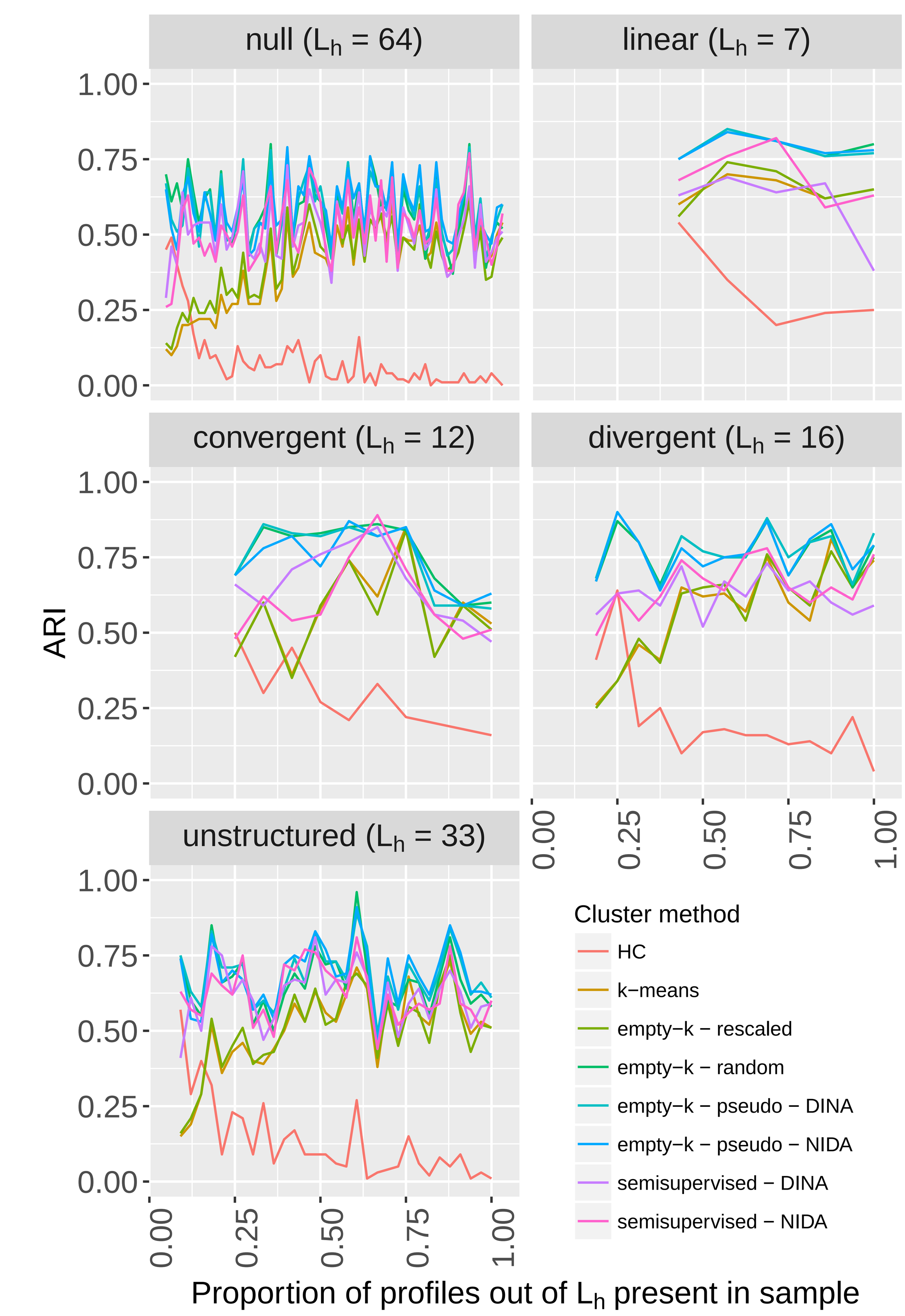
4. Evaluate clusters

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

Conclusions

- Winner: **Empty k-means** with **pseudocenters**
- ...even when different models used for data vs. starting centers! (DINA vs. NIDA)
- HC performs poorly, unlike when all profiles present *[1]*
- Fluctuations in null and unstructured hierarchies, possibly due to random sampling of profiles
- Performance non-monotonic in prop. of profiles

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

[1] Chiu, Douglas, & Li. (2009).
[2] Nugent, Dean, & Ayers. (2010).
[3] Su, Yu-Lan. (2013).