Latent Class Models for Multivariate Categorical Data

Siyuan Tang

STATS 607 Final Project Presentation

December 2025

Motivation: High-Dimensional Categorical Data

- We observe categorical vectors:

$$X = \left(X^{(1)}, \dots, X^{(m)}\right), \quad X^{(r)} \in \{1, \dots, C_r\}.$$

- The joint distribution has $\prod_{r=1}^{m} C_r - 1$ parameters (exponential in m).
- How do we model high-dimensional categorical data efficiently, without estimating an enormous joint table?

Latent Class Model

Introduce a discrete latent variable $H \in \{1, \ldots, K\}$ and assume

$$X^{(1)}, \ldots, X^{(m)} \perp\!\!\!\perp \mid H.$$

Then

$$P(X = x) = \sum_{k=1}^{K} P(H = k) \prod_{r=1}^{m} P\Big(X^{(r)} = x^{(r)} \mid H = k\Big), \quad \forall x \in \mathcal{X}.$$

$$\text{Parameters: } \pi_k = P(H = k), \qquad \theta_{rkc} = P(X^{(r)} = c \mid H = k).$$

$$\text{Dimension reduction: } (K - 1) + \sum_{r=1}^{m} K(C_r - 1) \ll \prod_{r=1}^{m} C_r - 1.$$

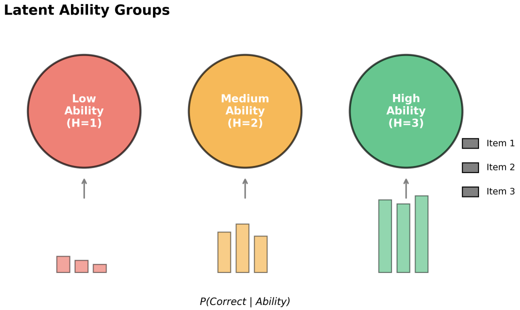
Fit the model using EM algorithm.

Psychometrics / Educational Testing

Observed Responses (3 binary items)

Student	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
1	1	1	0
2	1	0	1
3	0	0	0
4	1	1	1

$$X = (X^{(1)}, X^{(2)}, X^{(3)}), \; X^{(r)} \in \{0, 1\}.$$



Latent ability groups ($H = 1, 2, 3$):
Low / Medium / High ability

$$P(X = x) = \sum_{k=1}^{K} P(H = k) \prod_{r=1}^{3} P\left(X^{(r)} = x^{(r)} \mid H = k\right).$$

Statistical Computing Principles

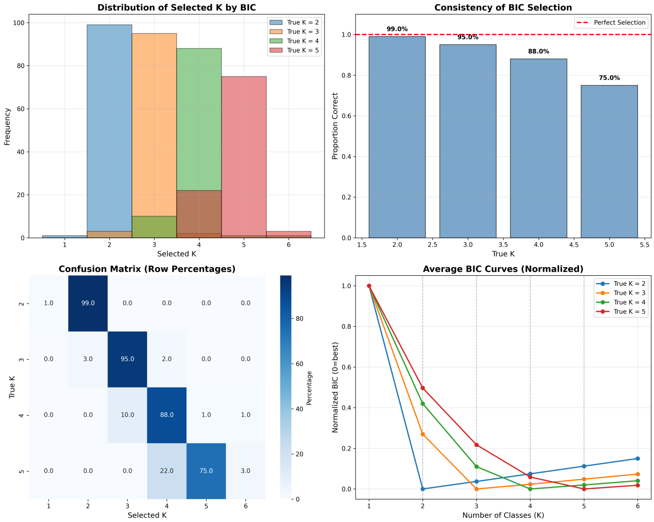
Techniques from the course applied in this project:

- Vectorization of probability updates.
- Parallelization for Monte Carlo simulation studies.
- Log-sum-exp for numerical stability.
- Modular Python structure (initialization, fit, predict).
- Makefile.

Structure of the Solution

1. **Define the latent class likelihood** (complete-data + observed-data versions).
2. **Implement EM algorithm** – Functions for E-step, M-step, log-likelihood, convergence check.
3. **Random initialization strategy** – Multiple starts to avoid poor local optima. – Label-switching mitigation via class ordering.
4. **Model selection loop over K** – Fit model for $K = 1, \ldots, K_{\max}$. – Track log-likelihood paths and BIC values.

Progress So Far



(BIC vs K from simulation. The jupyter notebook is here.)

Remaining Work

The altimate goal is to develop an end-to-end statistical software.

- **Simulation Studies**
  - Parameter estimation in different configurations.
- **Software Packaging**
  - Create a clean module + README.
  - Makefile.
  - Unit tests for key steps.