Use factor-augmented regularized latent regression to analyze complex large-scale assessment

He Ren, Yijun Cheng, and Chun Wang
University of Washington, Seattle

IMPS 2025, Minneapolis, Minnesota

UNIVERSITY *of* WASHINGTON | COLLEGE OF EDUCATION

Introduction: Large-scale Assessment

- Large-scale assessment
  - Evaluates student proficiency and informs educational policy
  - Report group-level statistics rather than results for the individual participants
- Latent regression and plausible values
  - Optimal individual estimations may not result in optimal group-level estimations
  - Plausible values are introduced for optimal group-level estimates
  - The plausible values are randomly drawn based on the IRT-latent regression model (IRT-LR; Mislevy, 1984, 1985)

$$\textbf{IRT-LR} \quad d(\theta|\mathbf{Y}) \propto f(\mathbf{Y}|\theta)g(\theta|\mathbf{X})$$

Introduction: Large-scale Assessment

- Large-scale assessment
  - Evaluates student proficiency and informs educational policy
  - Report group-level statistics rather than results for the individual participants
- Latent regression and plausible values
  - Optimal individual estimations may not result in optimal group-level estimations
  - Plausible values are introduced for optimal group-level estimates
  - The plausible values are randomly drawn based on the IRT-latent regression model (IRT-LR; Mislevy, 1984, 1985)

$$\textbf{IRT-LR} \quad d(\theta|\mathbf{Y}) \propto \underbrace{f(\mathbf{Y}|\theta)}_{\textbf{IRT}}\underbrace{g(\theta|\mathbf{X})}_{\textbf{LR}}$$

Introduction: Latent Regression

> Latent regression $g(\theta|\mathbf{X})$

Student Background Information

$$\theta = \mathbf{X}\beta + \epsilon$$

- High-dimension, numbering in thousands
- Highly correlated

> Principal component analysis (PCA) solutions

$$\mathbf{PCs} = \mathbf{XW}$$

$$\theta = \mathbf{PCs}\beta + \epsilon$$

- The first several PCs that explain 80% of the total variance
- The first N/8 PCs, where N is the student sample size

Introduction: Concerns

- Interpretation
  - Coefficients on PCs are hard to interpret
- Congeniality
  - Compatible with secondary analysis
  - The LR and secondary analysis model should be derived from the same underlying joint model for the data
  - The LR model should include as many background information as possible
- Estimation stability
  - Large standard errors
  - Too many PCs may lead to unstable estimations due to the multi-collinearity

Methods: Factor-augmented regularized LR

> Factor-augmented regularized latent regression (FARLR)
  - Proposed by Fan et al. (2020)
  - FA on background information

$$\mathbf{X} = \boldsymbol{f}B + \boldsymbol{u}$$

Common Factors **Idiosyncratic Residuals**

  - LR with both factor components and idiosyncratic residuals

$$\boldsymbol{\theta} = \boldsymbol{f}\beta + \boldsymbol{u}\gamma + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

Methods: Factor-augmented regularized LR

> Factor-augmented regularized latent regression (FARLR)
  - Proposed by Fan et al. (2020)
  - FA on background information

$$\mathbf{X} = \boldsymbol{f}B + \boldsymbol{u}$$

Common Factors **Idiosyncratic Residuals**

  - LR with both factor components and idiosyncratic residuals

$$\boldsymbol{\theta} = \boldsymbol{f}\beta + \boldsymbol{u}\gamma + \epsilon \quad \textbf{Add Regularization}$$

$$\epsilon \sim N(0, \sigma^2)$$

  - Ensures both congeniality and stable estimation

Methods: Model Estimation

> Factor-augmented regularized latent regression (FARLR)

- Combined with IRT, the FARLR is

$$\mathbf{P}(\mathrm{Y}_i, \theta_i | f_i, u_i, \beta, \gamma, \sigma^2) = \mathbf{P}(\mathrm{Y}_i | \theta_i)\mathbf{P}(\theta_i | f_i, u_i, \beta, \gamma, \sigma^2)$$

IRT Probability **LR with Common Factors and Idiosyncratic Residuals**

- The log-likelihood is

$$l(\beta, \gamma, \sigma^2) = \log \mathbf{P}(\mathrm{Y}_i | \theta_i) + \log \mathbf{P}(\theta_i | f_i, u_i, \beta, \gamma, \sigma^2)$$

Methods: Model Estimation

> EM Algorithm
  - E-step

$$Q(\beta, \gamma, \sigma^2) = \sum_{i=1}^{N} \mathbf{E}_\theta[\log \mathbf{P}(\theta_i | f_i, u_i, \beta, \gamma, \sigma^2)] + \text{constant}$$

$$= \sum_{i=1}^{N} \mathbf{E}_\theta \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\theta_i - f_i\beta - u_i\gamma)^2}{2\sigma^2} \right] + \text{constant}$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N} \int (\theta_i - f_i\beta - u_i\gamma)^2 q(\theta_i)\, d\theta_i + \text{constant}$$

Methods: Model Estimation

> EM Algorithm

$\int g(\theta)\, q(\theta) d\theta$

- E-step

$$Q(\beta, \gamma, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\int (\theta_i - f_i\beta - u_i\gamma)^2 q(\theta_i)\, d\theta_i + \text{constant}$$

- Important Sampling

$$\int g(\theta)q(\theta)d\theta = \int \frac{g(\theta)q(\theta)}{h(\theta)} h(\theta)d\theta \approx \frac{1}{K}\sum_{k=1}^{K}\frac{g(\theta_k)q(\theta_k)}{h(\theta_k)}$$

Methods: Model Estimation

> EM Algorithm
  - Important Sampling for E-step

$$Q(\beta, \gamma, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\int(\theta_i - f_i\beta - u_i\gamma)^2 q(\theta_i)\, d\theta_i + \text{constant}$$

$$\approx -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2 K}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\tilde{\theta}_{ik} - f_i\beta - u_i\gamma\right)^2 w_{ik} + \text{constant}$$

Sampling Weight

Randomly Drawn Samples

Methods: Model Estimation

> EM Algorithm
- M-Step

Weighted Regression

$$Q(\beta, \gamma, \sigma^2) \approx \boxed{-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2 K}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\tilde{\theta}_{ik} - f_i\beta - u_i\gamma\right)^2 w_{ik}} + \text{constant}$$

- Weighted Regression with regularization (LASSO penalty) on u_i
- LASSO introduces bias on the coefficients and node-wise estimation (Van de Geer et al., 2014) is used to de-bias the estimations

Simulation: Design

> Generation of background variables X

$$X_i = f_i B^T + u_i$$

where f is d-dimension and X is p-dimension

- (d, p) has **three levels** with (2,60), (20,150), and (45, 900)
- Loading structure has **two levels**

Evenly load only on one dominant factor; and each covariate also loads on 20% of the non-dominant factors.

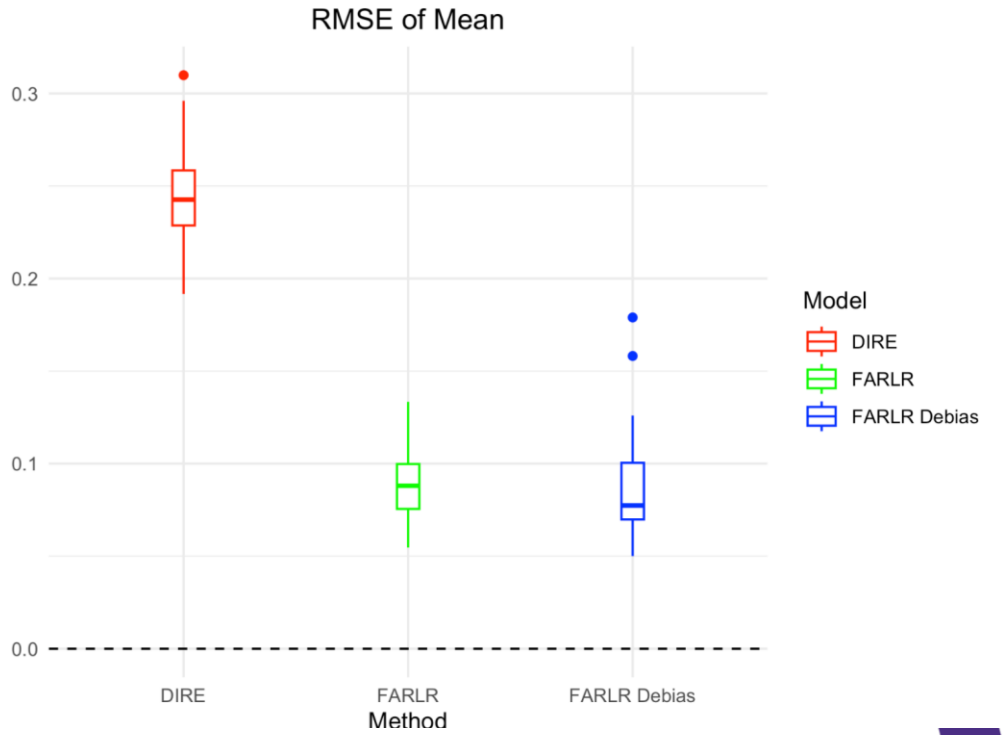
> Generation of θ

$$\theta_i = X_i \beta + \epsilon$$

- β has **two levels** on sparsity: 20% or 40% non-zero, drawn from U(0.75, 1.25)

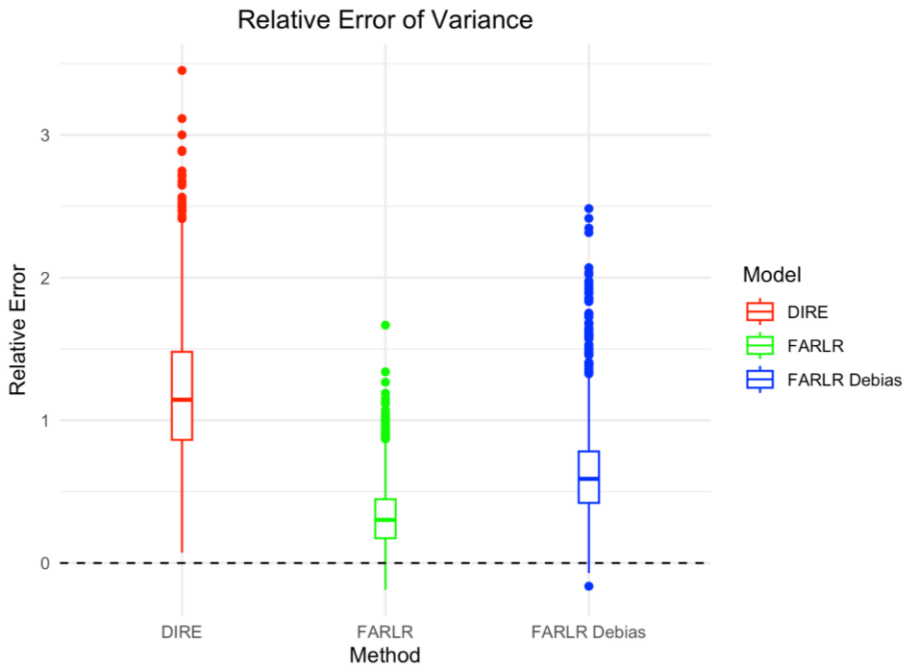
Simulation: Results

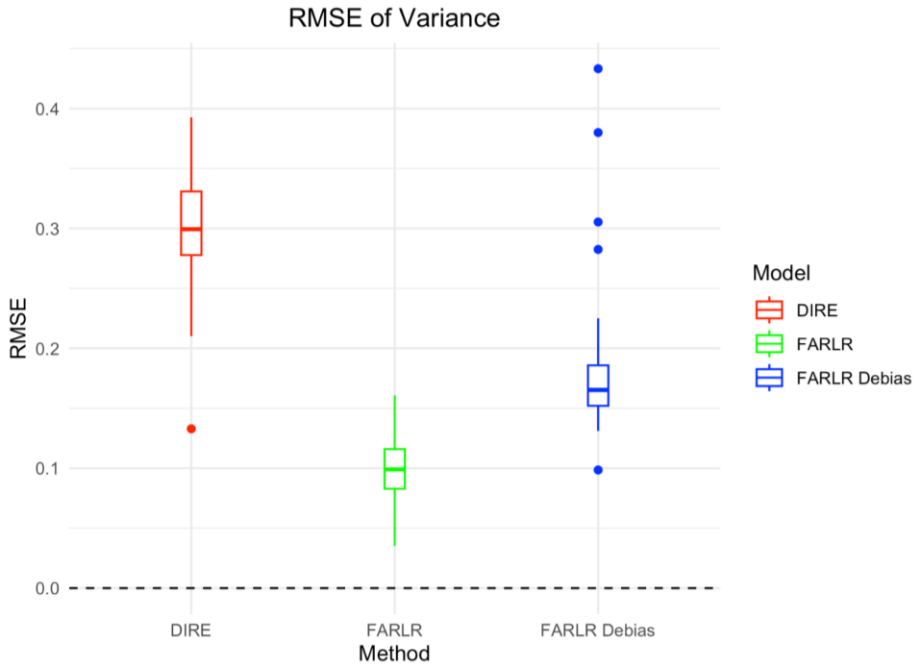
> Group Mean



Simulation: Results

> Group Variance





Takeaway

- We proposed a factor-augmented regularized latent regression for LSA
- The proposed model balance between the congeniality and stable estimation
- Simulation studies show that our methods can result in better group-mean and group-variance estimation

Thanks!

heren@uw.edu
chengxb@uw.edu

Personal Homepage

He Ren, Yijun Cheng, Chun Wang and **Gongjun Xu**

UNIVERSITY *of* WASHINGTON | COLLEGE OF EDUCATION