Modeling Bounded and Skewed Item Response Data with the Multidimensional Beta Factor Model

2023 Western Psychological Association (WPA)

Alfonso J. Martinez

April 28, 2023

Department of Psychological and Quantitative Foundations
Department of Statistics and Actuarial Sciences

IOWA

Overview

- Purpose of research: develop a psychometric model (beta IFA) for analyzing interval-scale data from continuous rating scales (CRS)

- Existing approaches (e.g., normal-theory factor analysis) are not always appropriate for CRS data
- Features of beta IFA model: (1) can directly model skewness in the responses (no data transformations needed) and (2) respects bounds of data

Overview

- Purpose of research (part 2): derive an estimation routine for the beta IFA model (expectation maximization algorithm) and investigate performance of algorithm in finite sample settings
  - Simulation study I: parameter recovery across multiple R and I, where R is sample size and I is number of indicators
  - Simulation study II: comparative study between beta IFA and NTFA under different response distributions
  - Empirical application of beta IFA model to real dataset

Latent Variable Models

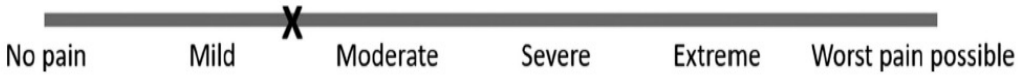
- Latent variable models (LVMs) are among the most widely used statistical methods in social and behavioral sciences for modeling latent constructs, e.g., personality, socioeconomic status, attitudes
- Latent constructs are not directly observable (or are difficult to observe)
- Researchers rely on indirect measurements via observable indicators (items) that serve as proxies of the construct
- Indicators are noisy manifestations (measurements) of latent construct

Latent Variable Models

- LVMs conceptualize latent construct(s) as latent variable(s) θ
- LVMs specify a functional relationship between latent variable θ and set of observed indicators y

Continuous Rating Scales

- Measurement of latent constructs with continuous rating scales (CRS) has been found to be more informative than discrete rating scales., e.g., Likert-scales
- Example of CRS format



Analysis of CRS Data

- CRS data is often analyzed with normal-theory factor analysis (NTFA)
- Issue 1: tendency for response distribution to be skewed
- Issue 2: NTFA assumes response range is unbounded; however CRS data typically have well defined endpoints
- Beta IFA model accounts for these two issues directly

The Beta Item Factor Analysis Model

- Assume that R respondents are each measured on I observed variables (indicators/items), and the total number of latent variables is $K < I$
- Let $\theta_r = (\theta_{r1}, \ldots, \theta_{rK})^\top$ be respondents r's vector of factor scores (latent variables)
- Assume that $\theta_r \sim N(0, \Sigma)$ where 0 is the $K \times 1$ zero vector and Σ is a $K \times K$ covariance matrix
- Collection of factor scores is $\Theta = (\theta_1^\top, \ldots, \theta_R^\top)^\top$
- The $R \times I$ observed response matrix is denoted $\mathcal{Y} = (y_1^\top, \ldots, y_R^\top)$ where $y_r = (y_{r1}, \ldots, y_{rI})^\top$ is the vector of responses produced by respondent r and y_{ri} is the observed response to the ith item

The Beta Item Factor Analysis Model

- Beta IFA model is constructed by assuming that

$$Y_{ri} \mid \theta_r \sim \text{MP-Beta}(\mu_{ri}, \phi_i) \tag{1}$$

where $0 < \mu_{ri} < 1$ and $\phi_i \in \mathbb{R}^+$ are the mean and precision parameters, respectively, with

$$\mu_{ri} \equiv E(Y_{ri} \mid \theta_r) = F(\beta_i + \lambda_i^\top \theta_r) \tag{2}$$

where (β_i, λ_i) are item parameters and F is a suitable link function

- We use the inverse logit link function is used so that

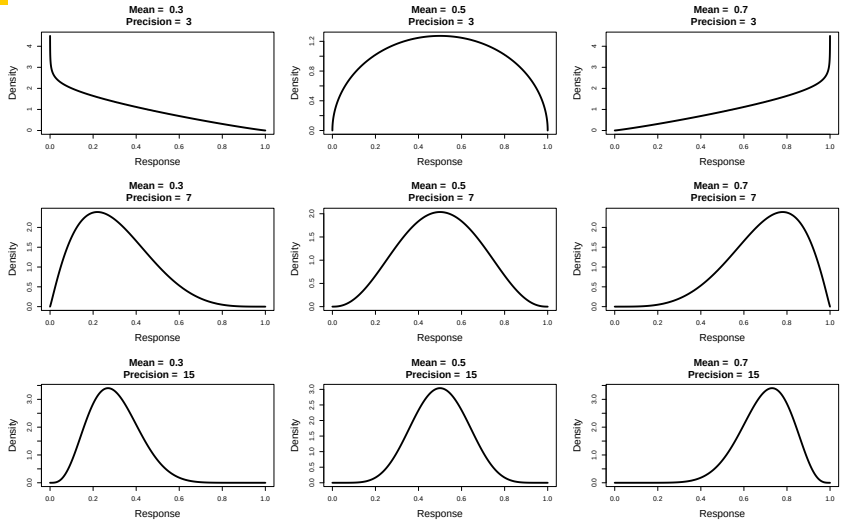
$$\mu_{ri} \equiv E(Y_{ri} \mid \theta_r) = \frac{\exp(\beta_i + \lambda_i^\top \theta_r)}{1 + \exp(\beta_i + \lambda_i^\top \theta_r)} \tag{3}$$

- θ_r can be thought of as unobserved/missing covariates

The Beta Item Factor Analysis Model with Logit Link

- β_i is item intercept (logit of the expected response when $\theta_r = 0$)
- $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{iK})^\top$ is a $K \times 1$ vector of factor loadings that relate the kth factor to the ith item
- ϕ_i is item precision (degree of separability in the responses)

Response Distributions Beta IFA Model can Accomodate



Maximum Likelihood Estimation via EM Algorithm

- We estimate beta IFA model parameters from observed response matrix $\mathcal{Y}$ via an EM algorithm
- The EM algorithm is an iterative procedure for MLE in the presence of missing, or incomplete, data
- Key in developing EM algorithm for beta IFA model is recognizing that Θ *is* the missing (incomplete) data
- The EM algorithm maximizes incomplete-data likelihood indirectly by maximizing the complete-data likelihood, where the complete data is $(\mathcal{Y}, \Theta)$

Complete Data Log-likelihood

The complete data log-likelihood for the beta IFA model is

$$\log L(\lambda, \beta, \phi \mid \mathcal{Y}, \Theta) = \sum_{r=1}^{R} \sum_{i=1}^{I} \Big\{ \log \Gamma(\phi_i) - \log \Gamma(\mu_{ri}\phi_i) - \log \Gamma(\phi_i(1-\mu_{ri})) + (\mu_{ri}\phi_i - 1) \log y_{ri} + (\phi_i(1-\mu_{ri}) - 1) \log(1-y_{ri}) \Big\}$$

Let s indicate the iteration number, $s = 1, \ldots, S$. To evolve to the $(s+1)$th iteration, first compute the expected complete-data log-likelihood

$$Q(\lambda, \beta, \phi \mid \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}) = \mathbb{E}\left(\log L(\lambda, \beta, \phi; \mathcal{Y}, \Theta) \mid \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}, \mathcal{Y}\right) \tag{4}$$

where expectation is taken with respect to the posterior of Θ:

$$h(\Theta \mid \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}, \mathcal{Y}) = \frac{L(\lambda^{(s)}, \beta^{(s)}, \phi^{(s)}; \mathcal{Y}, \Theta)}{\int_{\tilde{\Theta}} L(\lambda^{(s)}, \beta^{(s)}, \phi^{(s)}; \mathcal{Y}, \tilde{\Theta}) d\tilde{\Theta}} \tag{5}$$

Posterior expectation is not in closed form and must be evaluated numerically

The approximate Q function is

$$\hat{Q}\left(\lambda, \beta, \phi \mid \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}\right) = \sum_{r=1}^{R} \sum_{i=1}^{I} \sum_{\vartheta \in \mathcal{G}^K} \hat{h}\left(\vartheta \mid y_r, \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}\right) \Big\{ \log \Gamma(\phi_i) - \log \Gamma(\mu_{ri}\phi_i) - \log \Gamma(\phi_i(1-\mu_{ri})) + (\mu_{ri}\phi_i - 1) \log y_{ri} + (\phi_i(1-\mu_{ri}) - 1) \log(1-y_{ri}) \Big\}, \quad (6)$$

where (...see next slide)

$$\hat{h}\left(\vartheta \mid y_r, \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}\right) = \frac{\varphi(\vartheta)\prod_{i=1}^{I} B\left(\phi_i^{(s)}, \mu_{\bullet i}^{(s)}\right) y_{ri}^{\mu_{\bullet i}^{(s)}\phi_i^{(s)}-1}(1-y_{ri})^{(1-\mu_{\bullet i}^{(s)})\phi_i^{(s)}-1}}{\sum_{\vartheta' \in \mathcal{G}^K}\varphi(\vartheta')\prod_{i=1}^{I} B\left(\phi_i^{(s)}, \mu_{\bullet i}^{(s)}\right) y_{ri}^{\mu_{\bullet i}^{(s)}\phi_i^{(s)}-1}(1-y_{ri})^{(1-\mu_{\bullet i}^{(s)})\phi_i^{(s)}-1}}$$

is the approximate posterior density and

$$\mu_{\bullet i}^{(s)} = \frac{\exp(\beta_i^{(s)} + (\lambda_i^{(s)})^\top \vartheta)}{1 + \exp(\beta_i^{(s)} + (\lambda_i^{(s)})^\top \vartheta)}$$

is the expected response evaluated at the quadrature point ϑ.

M-step

Given provisional item estimates $\left(\lambda^{(s)}, \beta^{(s)}, \phi^{(s)}\right)$, the parameter estimates are updated at the $(s+1)$th iteration by solving the optimization problem:

$$\left(\lambda^{(s+1)}, \beta^{(s+1)}, \phi^{(s+1)}\right) = \underset{\lambda,\beta,\phi}{\arg\max}\ \hat{Q}\left(\lambda, \beta, \phi \mid \lambda^{(s)}, \beta^{(s)}, \phi^{(s)}\right). \tag{7}$$

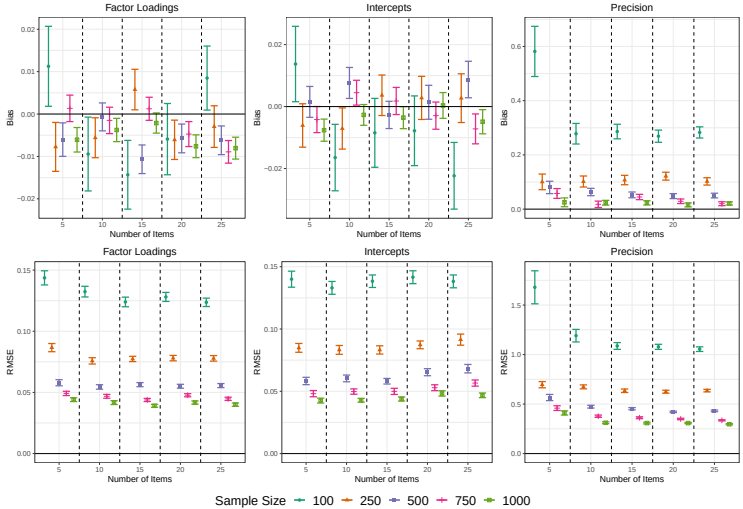
Repeat E- and M- step until convergence (i.e., parameters stabilize). BFGS algorithm was used in M-step.

Simulation Studies

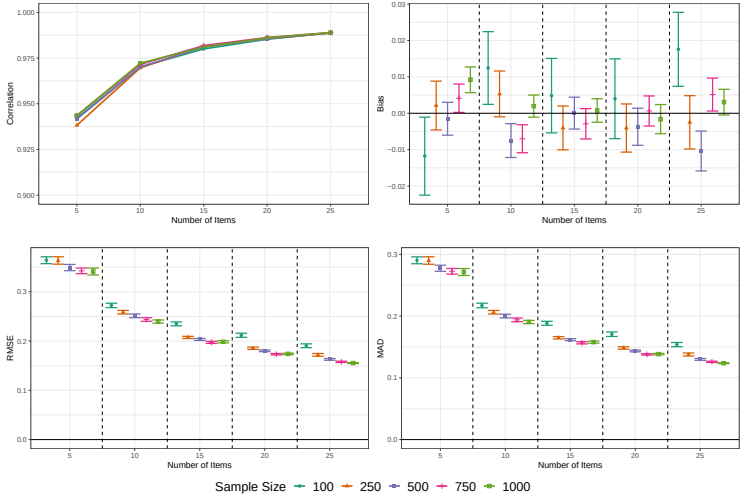
Simulation Study I: Parameter Recovery

- Purpose: evaluate performance of EM algorithm in recovering item parameters in small and large sample size settings
- 25 Conditions: $(R, I) \in \{100, 250, 500, 750, 1000\} \times \{5, 10, 15, 20, 25\}$
- Specifications: $K = 1$, $\lambda \sim U(0.3, 1.8)$, $\beta \sim U(-1.5, 1.5)$, $\phi \sim U(2, 10)$
- $J = 100$ independent datasets per condition
- Measures of estimation quality
  - Bias
  - RMSE
  - Absolute error
  - Correlation

Simulation Study I Selected Results: Item Parameter Recovery



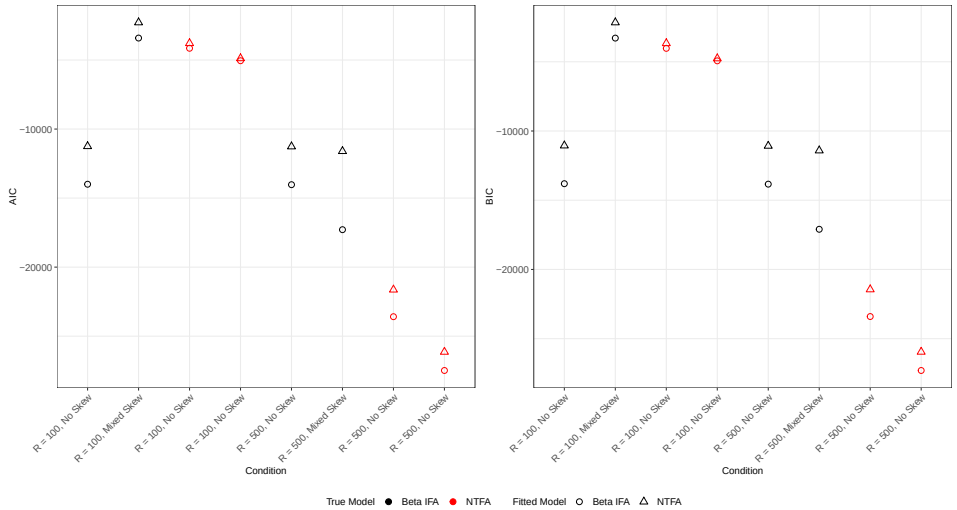
Simulation Study I Selected Results: Factor Score Recovery



Simulation Study II: Comparison with Normal-Theory Factor Analysis

- Idea: How does beta IFA perform relative to normal-theory factor analysis under different data-generating models (DGM)?
- NTFM: $Y_{ri} \mid \theta_r \sim N(\tau_i + \lambda_i \theta_r, \psi_i)$
- Specifications: $K = 1$, $I = 15$, $\lambda_i = 1$ for all $i = 1, \ldots, 15$ Conditions:
  - DGM: beta IFA with $R = 100$ (500) and $\beta_i = 0$ for all $i = 1, \ldots, 15$ (**No Skewness**)
  - DGM: beta IFA with $R = 100$ (500) and $\beta_i = -1.5$ for all $i = 1, \ldots, 5$, $\beta_i = 0$ for all $i = 6, \ldots, 10$, and $\beta_i = 1.5$ for all $i = 11, \ldots, 5$ (**Mixed Skewness**)
  - DGM: NTFM with $R = 100$ (500) and $\tau_i = 0$ for all $i = 1, \ldots, 15$ (**No Skewness**)
  - DGM: NTFM with $R = 100$ (500) and $\tau_i = -1.5$ for all $i = 1, \ldots, 5$, $\tau_i = 0$ for all $i = 6, \ldots, 10$, and $\tau_i = 1.5$ for all $i = 11, \ldots, 5$ (**No Skewness**)

Simulation Study II Results: Model Fit



Empirical Application

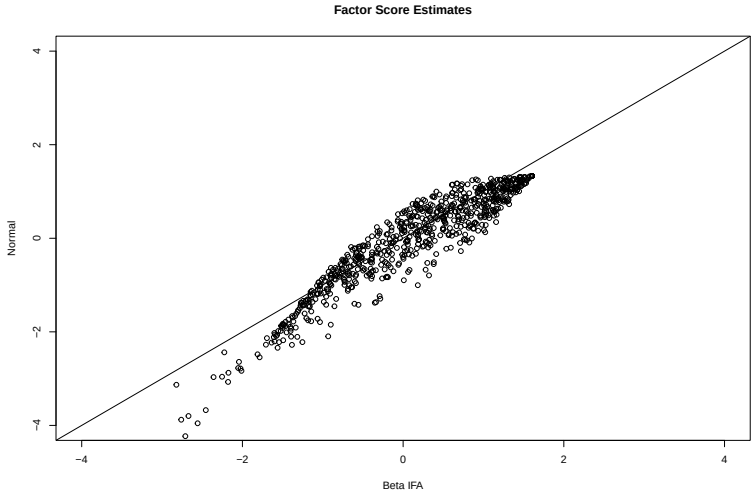
Empirical Application: Self-determination

- Self-Determination Inventory (SDI) is a 21-item measure of self-determination developed to document change in the self-determination of adolescents with and without disabilities
- Ratings are made on a slider scale with anchors of “Disagree” and “Agree” with numeric ratings represented on a scale ranging from 0 to 99
- Analysis with data from $R = 739$ students who participated in a randomized control trial in which the SDI was used as an outcome measure
- SDI response distribution are highly asymmetric

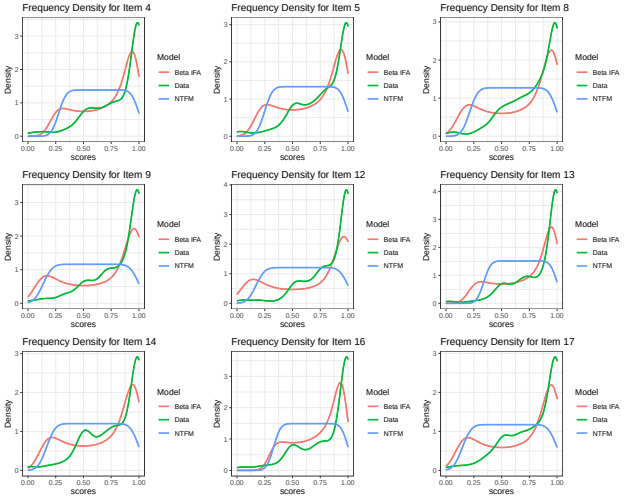
Empirical Application: Self-determination

- SDI was estimated with the beta IFA model assuming a one-factor structure ($K = 1$) using the EM algorithm
- For comparative purposes, SDI was also estimated under a normal-theory factor model (lavaan was used for estimation)
- Models compared via AIC and BIC (not shown here)
- Compute correlation of factor score estimates
- Superimposed plots of marginal density functions produced by both models

Empirical Application: Self-determination



Empirical Application: Self-determination



Wrapping Up

- Propose and develop the beta item factor analytic model (beta IFA) for modeling interval item response data
- Derive a maximum likelihood-based estimation algorithm

Simulation study I: Algorithm was capable of accurately recovering model parameters in relatively small sample size settings ($R = 250$, $I = 10$)

Simulation study II: Beta IFA model performed similar to the NTFA under symmetric response distributions but outperformed NTFA under asymmetric response distributions

Empirical applications provide preliminary evidence that model may be useful for interval scale data

Wrapping Up

Future directions:

- Estimation algorithm will be made publicly available
- Development of diagnostic procedures for evaluating model fit
- Further investigation of conditions where beta IFA model is more appropriate than NTFA

Thank You!

Email: alfonso-martinez@uiowa.edu

Academic website: https://alfonso-martinez.netlify.app/

Preprint available at https://psyarxiv.com/tp8sx

Selected References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach. John Wiley Sons.

Bock, R. D., & Gibbons, R. D. (2021). Item response theory. John Wiley Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1–22.

Ellerby, Z., Wagner, C., & Broomell, S. B. (2021). Capturing richer information: On establishing the validity of an interval-valued survey response mode. Behavior Research Methods, 1–23.

Selected References

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. Journal of applied statistics, 31(7), 799–815.

McLachlan, G. J., & Krishnan, T. (2007). The EM algorithm and extensions. John Wiley & Sons.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. Journal of statistical software, 48, 1–36.

Shogren, K. A., Little, T. D., Grandfield, E., Raley, S., Wehmeyer, M. L., Lang, K. M., & Shaw, L. A. (2020). The self-determination inventory–student report: Confirming the factor structure of a new measure. Assessment for Effective Intervention, 45(2), 110–120.