行政院國家科學委員會專題研究計畫成果報告

廣義 LISREL 模式（V）
Generalized LISREL Models (V)

計畫類別：☑ 個別型計畫　☐ 整合型計畫

計畫編號：NSC 89－2118－M－002－008

執行期間：89 年 8 月 1 日至 90 年 7 月 31 日

個別型計畫：計畫主持人：胡賦強 助理教授
共同主持人：

整合型計畫：總計畫主持人：
子計畫主持人：

註：整合型計畫總報告與子計畫成果報告請分開編印各成一冊，彙整一起繳送國科會。

處理方式：（請打✓）
☐ 可立即對外提供參考
☑ 一年後可對外提供參考
☐ 兩年後可對外提供參考
（必要時，本會得展延發表時限）

執行單位：國立台灣大學 公共衛生學院 流行病學研究所
生物醫學統計組

中華民國　91　年　1　月　11　日

Draft

Generalized Factor Analysis: A Preliminary Result

Fu-Chang Hu and Shu-Hui Lai[1]

Division of Biostatistics
Graduate Institute of Epidemiology
College of Public Health
National Taiwan University
Taipei, Taiwan 100
R.O.C.

July 5, 2001

[1]Correspondence: Fu-Chang Hu, Division of Biostatistics, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, 1 Jen-Ai Road, Section 1, Room 1541, Taipei, Taiwan 100, R.O.C., Phone: +886 2 / 2394-2050 or +886 2 / 2312-3456 × 8349, Fax: +886 2 / 2351-1955, Email: fchu@ha.mc.ntu.edu.tw. This paper is based on the second part of Lai's dissertation under Hu's advice. We are very grateful to the Doctoral Oral Examination Committee members for their helpful discussions and comments. And, Hu wishes to thank Professor Karl Jöreskog for his interesting introduction of the statistical software *LISREL* (Version 8.0) in an invited workshop held in Taipei, Taiwan, R.O.C., August 29-31, 1995. This research project was financially supported by the National Science Council of the Republic of China (NSC 87-2118-M-002-003, NSC 88-2118-M-002-008, and NSC 89-2118-M-002-001).

Abstract

Factor analysis (FA) has been used widely in various areas of sciences to explore or examine the latent measurement structure from a set of observed indicator variables. Both the observed and the latent variables are usually assumed to be continuous and, at least, symmetrically distributed. In the past 30 years or so, several methods had been proposed to extend the FA method for discrete observed indicator variables and/or latent variables, which include latent structure analysis, latent profile analysis, latent class analysis, latent trait analysis, and factor analysis of categorical data. See, for example, Bartholomew (1987) and Basilevsky (1994, Chaps. 8 and 9, pp. 501-621) for details and the references therein. We are interested in developing a general framework for FA, called the "*generalized* factor analysis" (GFA), for continuous, discrete, or mixed observed indicator variables, as long as they belong to the *exponential family of distributions* such as Normal, Binomial, and Poisson distributions. Just like the *generalized* linear models (GLMs), which include analysis of variances (ANOVA), linear regression, logistic regression, and Poisson regression as the special cases, we hope that the GFA method extends the standard FA method to build a measurement structure of continuous latent variable(s) from observed continuous, binary, ordinal, count, or mixed indicator variables in a unified way. Yet, before doing that, we investigate the equivalence between *exploratory* factor analysis (EFA) and *confirmatory* factor analysis (CFA). To estimate the factor loadings in a GFA model, we apply the *iterative reweighted least squares* (IRLS) algorithm of GLMs to "linearize" the generalized factor model first, and then use the usual estimation methods of factor analysis to obtain the estimates of the factor loadings. Specifically, we develop independently a unified three-step estimation procedure for GFA, which is similar to the *E-M algorithm* discussed in Bartholomew (1987, Sec. 6.1, pp. 107-115). On the other hand, we treat the estimation of factor loadings in GFA models as an error-in-variable problem of GLMs, and then take econometricians' *instrumental variable* (IV) approach for *simultaneous equations model* (SiEM) to estimating factor loadings. We shall discuss the results of our simulation study and compare the performances of different estimators numerically.

Keywords:

EFA, CFA, Spectral decomposition, Identifiability, GLMs, IRLS algorithm, Categorical response, Discrete data, Response of a mixed type, Structural equation model, LISREL model, Error-in-variable, Instrumental variable, Simultaneous equations model.

Contents

1 Introduction

1.1 Development

The main purpose of factor analysis (FA) is to describe the covariance relationships among several observed variables in terms of fewer latent variables with the interpretation that the latent variables (called the *factors*), although unobserved, are manifested by these observed variables (called the *observed indicator variables*). Thus, the factor model specified for an FA is usually considered as a *measurement model*, which describes how the factors are "measured" by these observed indicator variables.

There are two major kinds of FA:

1. *Exploratory* factor analysis (EFA)
2. *Confirmatory* factor analysis (CFA)

The modern beginning of EFA lies in the early 20th century through the attempts of Karl Pearson, Charles Spearman, and some others to define and measure human's intelligence. At times, however, FA of a set of observed variables may have already been carried out in different samples, and thus the prior information about the measurement structure is available for use in further samples. For example, the number of common factors or the factor loading coefficients found to be insignificantly different from zero may be known to the researchers in the field. This method is CFA and its development is historically independent from EFA. As suggested by their names, EFA and CFA are usually used for different purposes – EFA for exploring, but CFA for verifying – a latent measurement structure among the observed indicator variables.

1.2 Motivation

Example: A Measurement Model for Child's Health Status.

Suppose that a pediatrician is interested in studying the effect of exercises on kid's health. He asks the following questions in order to measure a child's health status:

1. Body strength? e.g., What is her/his body mass index (BMI)? (*Ratio* data)
2. Physical activities? e.g., How fast can s/he run? (*Continuous* data)

3. Parents' judgment: 1, 2, $\cdots$, 7? (*Ordinal* data)

4. Number of times getting cold in the past two months? (*Counts* data)

Notice that the answer to each of these questions is of a unique data type. Although a variety of factor analysis techniques have been developed in the past, factor analysis of discrete data is still under development and most of factor analysis methods are designed to analyze the data with *all* continuous, ordinal, or categorical indicator variables. To handle the data of such a mixed type, factor analysis for non-homogeneous data is still a challenge to statisticians.

1.3 Focus

In this research, we shall focus on the following two settings:

- **Discrete indicator variables:** For example, binary, counts, or ordinal.
- **Mixed indicator variables:** For example, continuous, binary, and counts.

An outline of the related methods are listed below.

1.3.1 Discrete Indicator Variables

1. **All Binary Indicator Variables:**
   (a) Standard Factor Analysis:
      - Tetrachoric correlation
   (b) Factor Analysis of Categorical Data (Bartholomew 1987):
      - *Probit* link
      - *Logit* link
   (c) Other Factor Analyses:
      - Latent structure analysis
      - Latent profile analysis (LPA)
      - Latent class analysis (LCA)
      - Latent trait analysis (LTA)

2. **Polytomous Indicator Variables:**

   (a) Non-Ordinal Scale

   (b) All Ordinal Scale:

   - Factor Analysis: Polychoric correlation

1.3.2 Mixed Indicator Variables

- Continuous vs Ordinal Scales:

  – Factor Analysis: Polyserial correlation

We shall give some details in the review section.

2 Review #1: Factor Analysis (FA)

2.1 Continuous Data

2.1.1 Exploratory Factor Analysis (EFA)

A brief review is given below. See Hamilton (1992, Chap. 8, pp. 249-288), Johnson and Wichern (1998, Chap. 9, pp. 514-586), and Basilevsky (1994, Chap. 6, pp. 351-422) for more details. The **Table 8.16: Summary of factor analysis options** of Hamilton (1992, p. 282) provided an outline of the key elements of EFA.

1. **Model (with m Orthogonal Common Factors):**

   (a) Specification:

$$
\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\
X_3 - \mu_3 &= l_{31}F_1 + l_{32}F_2 + \cdots + l_{3m}F_m + \varepsilon_3 \\
\cdots \quad &\quad \cdots \ \cdots \ \cdots \ \cdots \ \cdots \ \cdots \\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

where $m \leq p$. Or, in matrix notation,

$$\boldsymbol{X}_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = \boldsymbol{L}_{p\times m}\boldsymbol{F}_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1}.$$

(b) Assumptions:

$$\begin{aligned} E(\boldsymbol{F}) &= \boldsymbol{0}, \qquad \mathrm{Var}(\boldsymbol{F}) = \boldsymbol{I}, \\ E(\boldsymbol{\varepsilon}) &= \boldsymbol{0}, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} \quad \text{(a diagonal matrix)}, \\ \mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{F}) &= \boldsymbol{0}. \end{aligned}$$

(c) Covariance Structure:

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X}) &= \mathrm{Var}(\boldsymbol{LF} + \boldsymbol{\varepsilon}) = \boldsymbol{L}\mathrm{Var}(\boldsymbol{F})\boldsymbol{L}^{\mathrm{T}} + \mathrm{Var}(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{LIL}^{\mathrm{T}} + \boldsymbol{\Psi} = \boldsymbol{LL}^{\mathrm{T}} + \boldsymbol{\Psi}. \\ \sigma_k^2 = \mathrm{Var}(X_k) &= (l_{k1}^2 + l_{k2}^2 + \cdots + l_{km}^2) + \psi_k \\ &= h_k^2 + \psi_k \qquad (k = 1, 2, \cdots, p) \\ &= \text{Communality} + \text{Specific Variance}. \\ \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{F}) &= \mathrm{Cov}(\boldsymbol{LF}, \boldsymbol{F}) = \boldsymbol{LI} = \boldsymbol{L}. \\ \mathrm{Cov}(X_k, F_j) &= l_{kj}. \end{aligned}$$

2. Estimation:

Since the m latent variables $\boldsymbol{F}$ have zero means and an identity variance-covariance matrix, the estimates of the factor loadings $\boldsymbol{L}$ can be obtained by decomposing the variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ of the centered $\boldsymbol{X}$ by the *spectral decomposition theorem* (for a square matrix) as shown below.

(a) Spectral Decomposition of A Square Matrix:

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathrm{T}}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues λ_i's of the square matrix $\boldsymbol{A}$ and $\boldsymbol{P}$ is an orthogonal matrix (i.e., $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{I}$ so that $\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{P}^{-1}$) of the eigenvectors corresponding to the eigenvalues λ_i's (see Johnson and Wichern 1998, pp. 67-68 and Basilevsky 1983, pp. 201-203).

(b) Application to the Orthogonal Factor Model:

If $\boldsymbol{\Psi} = \boldsymbol{0}$, then

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{LL}^{\mathrm{T}}$$

but

$$\boldsymbol{\Sigma_X} = \boldsymbol{P\Lambda P^{\mathrm{T}}} = (\boldsymbol{P\Lambda^{1/2}})(\boldsymbol{\Lambda^{1/2}P^{\mathrm{T}}})$$

so that

$$\boldsymbol{L} = \boldsymbol{P\Lambda^{1/2}}.$$

When the model has as many factors as the number of the indicator variables, (i.e., $m = p$), the covariance matrix $\boldsymbol{\Sigma_X}$ can be decomposed exactly as $\boldsymbol{LL^{\mathrm{T}}}$, since $\boldsymbol{\Psi}$ is a zero matrix. In EFA, it is requested that the set of eigenvectors in the orthogonal matrix $\boldsymbol{P}$ be *orthonormal*. Specifically, there are two major methods for obtaining the estimates of the factor loadings.

A. Principal Factor Method:

Principal factors are the principal components of the modified correlation matrix $\boldsymbol{R_r}$, in which the estimates of the communality replace the diagonal elements.

(a) Let $\boldsymbol{X}$ be standardized, if the factor model $\boldsymbol{\rho = LL^{\mathrm{T}} + \Psi}$ is correctly specified, then $\rho_{ii} = h_i^2 + \psi_i = 1$. Let $h_i^2 = 1 - \psi_i$, and thus the result matrix is $\boldsymbol{\rho - \Psi = LL^{\mathrm{T}}}$.

(b) Set the initial estimates ψ_i^*. The most popular choice is $\psi_i^* = 1/r^{ii}$, where r^{ii} is the ith diagonal element of $\boldsymbol{R^{-1}}$, or $\psi_i^* = 1/s^{ii}$, where s^{ii} is the ith diagonal element of $\boldsymbol{S^{-1}}$). Then, replace the ith diagonal element of the sample correlation matrix $\boldsymbol{R}$ by $h_i^{*2} = 1 - \psi_i^*$, we obtain a "reduced" sample correlation matrix

$$\boldsymbol{R_r} \cong \boldsymbol{L_r^* L_r^{*\mathrm{T}}}$$

where $\boldsymbol{L_r^*} = \{l_{ij}^*\}$ are the estimated loading. The *principal factor method* of factor analysis obtains the estimates

$$\boldsymbol{L^*}_r = [\widehat{\lambda}_1^* \widehat{\boldsymbol{e}}_1^*, \widehat{\lambda}_2^* \widehat{\boldsymbol{e}}_2^*, \cdots, \widehat{\lambda}_m^* \widehat{\boldsymbol{e}}_m^*].$$

(c) In turn,

$$h_i^{*2} = \sum_{j=1}^{m} l_{ij}^{*2}.$$

The communalities are then (re)estimated by

$$\psi_i^* = 1 - \sum_{j=1}^{m} l_{ij}^{*2}.$$

(d) The principal factor solution can be obtained iteratively with the communality estimates from (b).

See Johnson and Wichern (1998, pp. 529-530).

B. Maximum Likelihood Method:

If the common factors $\boldsymbol{F}$ and the specific factors ϵ can be assumed to be normally distributed, then the maximum likelihood estimates of the factor loadings and the specific variances may be obtained. When F_j and ε_j are joint normal, the observations $X_j - \mu = LF_j + \varepsilon_j$ are normal. Then the likelihood function is

$$\begin{aligned} L(\mu, \boldsymbol{\Sigma_X}) &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma_X}|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{\Sigma_X}^{-1}\left(\sum_{j=1}^{n}(X_j-\bar{X})(X_j-\bar{X})^{\mathrm{T}}+n(\bar{X}-\mu)(\bar{X}-\mu)^{\mathrm{T}}\right)\right]} \\ &= (2\pi)^{-\frac{(n-1)p}{2}} |\boldsymbol{\Sigma_X}|^{-\frac{(n-1)}{2}} e^{-\left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{\Sigma_X}^{-1}\left(\sum_{j=1}^{n}(X_j-\bar{X})(X_j-\bar{X})^{\mathrm{T}}+n(\bar{X}-\mu)(\bar{X}-\mu)^{\mathrm{T}}\right)\right]} \times \\ &\quad (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma_X}|^{-\frac{1}{2}} e^{-\left(\frac{n}{2}\right)(\bar{X}-\mu)^{\mathrm{T}}\boldsymbol{\Sigma_X}^{-1}(\bar{X}-\mu)} \end{aligned}$$

which depends on $\boldsymbol{L}$ and Ψ through $\boldsymbol{\Sigma_X} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} + \boldsymbol{\Psi}$. It is desirable to make $\boldsymbol{L}$ well defined by imposing the computationally convenient *uniqueness condition* that

$$\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{L} = \Delta \text{ be a diagonal matrix.}$$

The maximum likelihood estimates $\widehat{\boldsymbol{L}}$ and $\widehat{\boldsymbol{\Psi}}$ must be obtained by numerical maximization of $L(\mu, \boldsymbol{\Sigma_X})$. See Johnson and Wichern (1998, pp. 529-530).

3. Factor Rotation:

If $\widehat{\boldsymbol{L}}$ is the $p \times m$ matrix of the estimated factor loadings obtained by one of the available methods (e.g. principal component, maximum likelihood, and so forth), then $\widehat{\boldsymbol{L}}^{*} = \widehat{\boldsymbol{L}}\boldsymbol{T}$ is a $p \times m$ matrix of the "rotated" factor loadings, where $\boldsymbol{T}$ is a chosen orthogonal matrix (i.e. $\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}} = \boldsymbol{T}^{\mathrm{T}}\boldsymbol{T} = \boldsymbol{I}$). After the rotation, the estimated covariance (or correlation) matrix remains unchanged, since $\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\mathrm{T}} + \widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{L}}(\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}})\widehat{\boldsymbol{L}}^{\mathrm{T}} + \widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{L}}^{*}\widehat{\boldsymbol{L}}^{*\mathrm{T}} + \widehat{\boldsymbol{\Psi}}$. See Lawley and Maxwell (1971, pp. 79-83), Hamilton (1992, pp. 259-263), Johnson and Wichern (1998, pp. 540-550), and/or Basilevsky (1994, pp. 56-62 and pp. 258-275) for more details.

The purpose of factor rotation is to seek more interpretable factors. That is, we hope to obtain a new factor model that fit the data equally well but has a simpler factor structure than the initial factors. A "simple factor structure" means that each variable loads strongly (either positively or negatively) on only one factor, but near zero on the other factors. For example, for standardized variables X_1, X_2, X_3, and X_4, Hamilton (1992, p. 259) suggested that a simple factor structure might be of the following pattern.

Variable	Loadings after Rotation	
	Factor 1	**Factor 2**
X_1	**near ± 1**	near 0
X_2	**near ± 1**	near 0
X_3	near 0	**near ± 1**
X_4	near 0	**near ± 1**

A. The Orthogonal Rotations:

An orthogonal rotation assumes that the common factors are *independent.* As an example, If we assume the coordinate axes are rotated clockwise or counterclockwise through the angle *phi*, the new loadings after rotation

$$\widehat{L}^* = \widehat{L}T$$

can be obtained by choosing the "rotation" matrix

$$T = \begin{bmatrix} cos\phi & sin\phi \\ -sin\phi & cos\phi \end{bmatrix} \text{ for clockwise rotation}$$

or

$$T = \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \text{ for counterclockwise rotation.}$$

See Basilevsky (1994, pp. 56-60) for more details. The most commonly used orthogonal rotation methods are the *Varimax method* and the *Quartimax method.*

For example, the *Varimax rotation* seeks an orthogonal transformation matrix T such that

$$\widehat{L}^* = \widehat{L}T$$

maximizes

$$\sum_{k=1}^{m}\sum_{j=1}^{p}\left(l_{kj}^2 - d_j\right)^2$$

where l_{kj} is the loading of the kth variable on the jth factor and d_j is the mean squared loading of the m variables on the jth factor

$$d_j = \frac{\sum_{k=1}^{m} l_{kj}^2}{m}.$$

See Hamilton (1992, pp. 261-262) for details.

B. The Oblique Rotations:

An oblique rotation is a transformation of the *correlated* factors. For example, the oblique matrix G

$$\begin{aligned}\widehat{L}\widehat{L}^{\mathrm{T}} &= LG(G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}G(G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}L \\ &= B\Phi B^{\mathrm{T}}\end{aligned}$$

where

$$\begin{aligned}B^{\mathrm{T}} &= (G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}L^{\mathrm{T}} \\ \Phi &= G^{\mathrm{T}}G \text{ is the correlation matrix of the oblique factors.}\end{aligned}$$

See Basilevsky (1994, pp. 270-278) for more details. The most commonly used oblique rotation methods are the *Promax method* and the *Oblimin method.*

For example, the *Promax rotation* begins with the varimax-rotated loadings $\widehat{L}^{*}$. We seek a transformation matrix G with columns that minimize

$$\mathrm{tr}\left[\left(M-\widehat{L}^{*}G\right)^{\mathrm{T}}\left(M-\widehat{L}^{*}G\right)\right]$$

where tr( ) denotes the trace operator. Elements of M are the factor loadings (elements of $\widehat{L}^{*}$) raised to an arbitrary power greater then 1 but usually less then 4. The higher the power, the stronger the correlations allowed between factors. The Promax-rotated loadings $\widehat{L}^{**}$ result from the postmultiplication of the varimax loadings $\widehat{L}^{*}$ by the transformation matrix G (after scaling G for unit-variance factors):

$$\widehat{L}^{**} = \widehat{L}^{*}G.$$

See Hamilton (1992, p. 262) for details.

2.1.2 Confirmatory Factor Analysis (CFA)

A brief review is given below. See Johnson and Wichern (1998, Sec. 9.7, pp. 565-571) and Bollen (1989, Chap. 7, pp. 226-318) for more details.

1. **Model Specification and Assumptions:**

$$\begin{aligned}x_1 &= \lambda_{11}F_1+\lambda_{12}F_2+\cdots+\lambda_{1m}F_m+\delta_1 \\ x_2 &= \lambda_{21}F_1+\lambda_{22}F_2+\cdots+\lambda_{2m}F_m+\delta_2 \\ x_3 &= \lambda_{31}F_1+\lambda_{32}F_2+\cdots+\lambda_{3m}F_m+\delta_3 \\ \cdots &\quad \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ x_p &= \lambda_{p1}F_1+\lambda_{p2}F_2+\cdots+\lambda_{pm}F_m+\delta_p\end{aligned}$$

or, in matrix notation

$$\boldsymbol{x} = \boldsymbol{\Lambda}_{\boldsymbol{x}} \boldsymbol{F} + \boldsymbol{\delta}$$

where

$$\begin{aligned} E(\boldsymbol{F}) = \mathbf{0}, \qquad & Var(\boldsymbol{F}) = \boldsymbol{\Phi} = [\phi_{ij}] \\ \phi_{ij} &= \begin{cases} 1 & \text{for } i = j \\ \phi_{ij} & \text{for } i \neq j \end{cases} \\ E(\boldsymbol{\delta}) &= \mathbf{0} \\ \mathrm{Var}(\boldsymbol{\delta}) &= \boldsymbol{\Theta}_{\boldsymbol{\delta}} \\ \mathrm{Cov}(\boldsymbol{\delta}, \boldsymbol{F}) &= \mathbf{0}. \end{aligned}$$

We can open the parameter ϕ_{ii} to be estimated, but we must specify one of the factor loadings to be equal to one, e.g. $\lambda_{11} = 1$ for each factor instead due to the unknown scale of the factor.

2. **Identification:**

The "known" parameters are the parameters that are known to be identified. These parameters are the population characteristics of the distribution of the observed variables such as their variances and covariances for which consistent sample estimators are readily available and for which the identification is typically not an issue. The "unknown" parameters are the parameters whose identification status are not known. Identification is demonstrated by showing that the unknown parameters are functions only of the identified parameters *and* that these functions lead to unique solutions. If this can be done, the unknown parameters are *identified*; otherwise, one or several parameters are *unidentified.*

The known-to-be-identified parameters are the elements of the population variance-covariance matrix of the observed variables $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. The unknown parameters are in $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ contains the t free and (nonredundant) constrained parameters of $\boldsymbol{\lambda}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Phi}$. If an unknown parameter in $\boldsymbol{\theta}$ can be written as a function of one or more elements of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, then that parameters is identified. If all unknown parameters in $\boldsymbol{\theta}$ are identified, then the model is identified. The most important necessary condition for identification is that Number of parameters in $\boldsymbol{\theta} \leq \frac{1}{2}p(p+1)$, where p is the number of variables. See Bollen (1989, pp. 88-104) for more details.

3. **Estimation:**

(a) According to the correctly specified factor model, we hypothesize that the population variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ of the observed variables is equal to the model *implied* variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta}).$$

(b) Given a sample of size n, the population variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ of the observed variables can be consistently and unbiasedly estimated by the *sample* variance-covariance matrix $\mathbf{S}$ of the observed variables, i.e.

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \mathbf{S}_n.$$

(c) Then, we choose the values of $\boldsymbol{\theta}$ such that $\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})$ is as close to $\mathbf{S}_n$ as possible. That is, find a $\widehat{\boldsymbol{\theta}}$ such that

$$\boldsymbol{\Sigma}_{\boldsymbol{X}}(\widehat{\boldsymbol{\theta}}) \cong \mathbf{S}_n.$$

In practice, the estimate $\widehat{\boldsymbol{\theta}}$ can be obtained by minimizing one of the following *fitting* functions, $F\left(\mathbf{S}, \boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})\right)$:

$$\begin{aligned} F_{\text{ULS}} &= \frac{1}{2} tr\left[\left(\mathbf{S} - \boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})\right)^2\right], \\ F_{\text{GLS}} &= \frac{1}{2} tr\left[\left(\mathbf{S} - \boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})\right)\mathbf{W}^{-1}\right]^2, \\ F_{\text{ML}} &= \log\left|\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})\right| + tr\left(\mathbf{S}\,\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{\theta})\right) - \log|\mathbf{S}| - p \end{aligned}$$

where $\mathbf{W}$ is a weight matrix for the residual matrix and p is the dimension of $\boldsymbol{X}$. In fact, they are special cases of the following more general fitting function

$$F_{\text{WLS}} = \left[\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})\right]^{\mathrm{T}} \mathbf{W}^{*-1} \left[\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})\right]$$

where

$$\begin{aligned} \boldsymbol{s} &= \frac{1}{2}\,p(p+1) \times 1 \text{ vector} \\ &\quad \text{(obtained by placing the nonduplicated elements of } \mathbf{S} \text{ in a vector)}, \\ \boldsymbol{\sigma}(\boldsymbol{\theta}) &= \frac{1}{2}\,p(p+1) \times 1 \text{ vector} \\ &\quad \text{(obtained by placing the nonduplicated elements of } \boldsymbol{\Sigma} \text{ in a vector)}, \\ \boldsymbol{\theta} &= t \times 1 \text{ vector of the free parameters}, \\ \mathbf{W}^* &= \frac{1}{2}\,p(p+1) \times \frac{1}{2}\,p(p+1) \text{ positive definite weighted matrix.} \end{aligned}$$

See Bollen (1989, pp. 425-432) for more details.

(d) The minimization

$$\min_{\widehat{\boldsymbol{\theta}}} F\left[\mathbf{S}_n, \boldsymbol{\Sigma}_{\boldsymbol{X}}(\widehat{\boldsymbol{\theta}})\right]$$

can be done by solving a set of the *second-moment estimating functions*:

$$\frac{\partial F\left[\mathbf{S}_n, \boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

for $\widehat{\theta}$. The global minimum reaches **0** when in fact

$$\mathbf{S}_n - \mathbf{\Sigma}_{\boldsymbol{X}}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$$

elementwise.

2.2 Discrete Data

A classification of various factor analysis methods is listed in **Table 1**. See Basilevsky (1994, p. 608) and Bartholomew (1987, Table 1.1, p. 4) for details.

1. Latent Structure Analysis
2. Latent Profile Analysis (LPA)
3. Latent Class Analysis (LCA)
4. Latent Trait Analysis (LTA)
5. Factor Analysis of Categorical Data:
   (a) *Probit* link (Muthén 1978)
   (b) *Logit* link (Bartholomew 1980)

If underlying normal distributions can be assumed for all the categorical observed indicators, then we can use

1. All Binary Case: Factor analysis with *probit* link.
2. All Ordinal Case: Factor analysis with *polychoric* correlation.
3. Mixed Case: Factor analysis with *polyserial* correlation.

Specifically, Bartholomew (1987) discusses the following two different approaches to factor analysis of categorical data:

1. **Response Function (RF) Approach:**

This approach is achieved by choosing a suitable response function

$$G^{-1}\{\pi_i(y)\} = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} H^{-1}(y_i) \qquad (i = 1, \cdots, p)$$

where $\pi_i(y) = Pr\{x_i = 1|y\}$ is the response function and y_i $(j = 1, 2, \cdots, q)$ are independently and uniformly distributed on $(0, 1)$. The function G^{-1} and H^{-1} were arbitrary but such that their inverses G and H were distribution functions of random variables symmetrically distributed about zero.

2. **Underlying Variable (UV) Approach:**

This approach is achieved by supposing that underlying the ith dichotomy there is a continuous variable, F_i say. The observed binary variable x_i is then an indicator of whether F_i is above or below some critical level τ_i. Thus, we may define

$$\begin{aligned} x_i &= 1, \text{ if } F_i \leq \tau_i \\ &= 0, \text{ otherwise} \end{aligned}$$

where

$$\boldsymbol{F} = \boldsymbol{\mu} + \boldsymbol{\Lambda Z} + \boldsymbol{e}.$$

And, Bartholomew (1987, pp. 107-115) discusses an *E-M algorithm* for estimating the parameters as listed follows:

- **Step 1:** Assume arbitrary starting values for the parameters.
- **Step 2:** Using these values, predict Z for each individual using the posterior expectation of Z given X.
- **Step 3:** Treating these expected values as if they were true values, estimate the parameters by maximum likelihood.
- **Step 4:** Return to **Step 2** and repeat the cycle until convergence is attained.

3 Review #2: Generalized Linear Models (GLMs)

See Dobson (1990), McCullagh and Nelder (1989), and Fahrmeir and Tutz (1994) for details.

3.1 Model Specification and Interpretation

A generalized linear model (GLM) is

$$g(\mu_i) = \mathbf{\Lambda_Y} \boldsymbol{F}_i$$

where i indexes observation, $\mu_i = E(Y_i)$, and

- $g(\cdot)$: A *link* function, which is monotonic and differentiable,
- $Y_i \sim$ Exponential family of distributions,
- $\mathbf{\Lambda_Y}$: Unknown regression parameters,
  $\boldsymbol{F}_i$: Covariates.

Examples:

1. **Linear regression model:**

$$E[Y_i] = \mathbf{\Lambda_Y} \boldsymbol{F}_i$$

which is a GLM because

- $Y_i \sim N(\mu_i, \sigma^2)$,
- g is the *identity* function, i.e.

$$g(\mu_i) = \mu_i.$$

2. **Logistic regression model:**

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{\Lambda_Y} \boldsymbol{F}_i$$

which is a GLM because

- $Y_i \sim Bernoulli(p_i)$,
- g is the *logit* function, i.e.

$$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right),$$

and thus

$$\mu_i = p_i = Pr(Y_i = 1) = \frac{1}{1 + exp^{\mathbf{\Lambda_Y} \boldsymbol{F}_i}}.$$

Then,

$$\mu_i = E(Y_i) = -\frac{c'(\theta_i)}{b'(\theta_i)}.$$

since $E(U) = 0$.

The jth score equation is

$$U_j = \frac{\partial l}{\partial \lambda_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0,$$

and thus

$$\boldsymbol{U} = \frac{\partial l}{\partial \boldsymbol{\Lambda_Y}} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0},$$

where

$$\eta_i = g(\mu_i) = \boldsymbol{\Lambda_Y} \boldsymbol{F}_i$$

and

$$\boldsymbol{D} = \begin{bmatrix} \ddots & & 0 \\ & \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} & \\ 0 & & \ddots \end{bmatrix} \boldsymbol{F}.$$

And, the information matrix $\boldsymbol{I}$ has the elements

$$I_{jk} = E[U_j U_k] = E\left[\frac{\partial l}{\partial \lambda_j} \frac{\partial l}{\partial \lambda_k}\right] = -E\left[\frac{\partial^2 l}{\partial \lambda_j \partial \lambda_k}\right].$$

When applying the *Newton-Raphson method* to solve for the maximum likelihood estimates (MLEs) $\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}$ of $\boldsymbol{\Lambda_Y}$, at the mth iteration

$$\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m)} = \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} - \left[\frac{\partial^2 l}{\partial \lambda_j \partial \lambda_k}\right]^{-1}_{\boldsymbol{\Lambda_Y} = \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m-1)}} \boldsymbol{U}^{(m-1)}.$$

If the *method of scoring* is used to obtain the MLEs $\widehat{\boldsymbol{\beta}}$, then at the mth iteration

$$\begin{aligned} \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m)} &= \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} + \left[\boldsymbol{I}^{(m-1)}\right]^{-1} \boldsymbol{U}^{(m-1)}, \\ \boldsymbol{I}^{(m-1)} \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m)} &= \boldsymbol{I}^{(m-1)} \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} + \boldsymbol{U}^{(m-1)} \end{aligned} \tag{3.1}$$

where

$$I_{jk} = E[U_j U_k] = E\left[\frac{\partial l}{\partial \lambda_j} \frac{\partial l}{\partial \lambda_k}\right] = \sum_{i=1}^{n} \frac{F_{ij} F_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Notice that we can write

$$\boldsymbol{I} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{F}$$

where $\boldsymbol{W}$ is an $n \times n$ diagonal matrix with elements

$$W_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Define a *pseudo*-response variable

$$Z_i = \sum_k \widehat{\lambda}_k^{(m-1)} F_{ik} + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i)$$

or in matrix notation

$$\boldsymbol{Z} = \boldsymbol{\Lambda}_{\boldsymbol{Y}}\boldsymbol{F} + g'(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}).$$

Then, the right-hand side of Eq. (3.1), which has the elements

$$\sum_k \sum_i \frac{F_{ij}F_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \widehat{\lambda}_k^{(m-1)} + \sum_i \frac{(y_i - mu_i)F_{ij}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right),$$

can be written as

$$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}.$$

It is straightforward to show that

1. $E(\boldsymbol{Z}) = \boldsymbol{\Lambda}_{\boldsymbol{Y}}\boldsymbol{F}$,
2. $Var(\boldsymbol{Z}) = \boldsymbol{W}^{-1}$.

Hence, the iterative equation for the method of scoring can be written as a *normal* equation

$$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{W}\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{Y}}^{(m)}\boldsymbol{F} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}$$

or

$$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{Z} - \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{(m)}\boldsymbol{F}\right) = \boldsymbol{0}$$

for a weighted least squares (WLS) estimation of the regression coefficients $\boldsymbol{\Lambda}_{\boldsymbol{Y}}$ in the above linear model of the pseudo-response variable $\boldsymbol{Z}$. This is the *iteratively reweighted least squares* (IRLS) algorithm. Thus, MLEs of the regression coefficients $\boldsymbol{\Lambda}_{\boldsymbol{Y}}$ in GLMs can be obtained by applying the unified IRLS procedure.

3.3 Statistical Inference

The statistical inference on $\mathbf{\Lambda_Y}$ of a GLM is based on

1. the asymptotic distribution of the score function $U(\mathbf{\Lambda_Y})$ (by a *central limit theorem*) and
2. a first-order Taylor expansion of $U(\mathbf{\Lambda_Y})$ for $\widehat{\mathbf{\Lambda}}_{\mathbf{Y}}$.

Dobson (1990, Chap. 5, pp. 49-67) gives the details.

3.4 Model-Fitting Techniques

When fitting a GLM to data in practice, the techniques for accomplishing the following tasks are needed:

1. **Goodness of fit:** See Dobson (1990, Secs. 5.5 and 5.8, pp. 56 and 60-61) for a nice discussion.
2. **Model selection:** Fahrmeir and Tutz (1994, Sec. 4.1, pp. 119-124) discuss several methods for variable selection.
3. **Regression diagnostics and remedies:** One may consult McCullagh and Nelder (1989, Chap. 12, pp. 391-418) and Fahrmeir and Tutz (1994, Secs. 4.2-4.3, pp. 124-149).

4 Research Problem

Factor analysis (FA) has been used widely in various branches of sciences to discover the latent measurement structure from a set of observed indicator variables. Both the observed and the latent variables are usually assumed to be continuous and, at least, symmetrically distributed. In the past 30 years or so, several methods had been proposed to extend the FA method for categorical observed indicator variables and/or latent variables, which include latent structure analysis, latent profile analysis, latent class analysis, latent trait analysis, and factor analysis of categorical data. See, for example, the books written by Bartholomew (1987) and Basilevsky (1994) and the references therein.

4.1 Framework

We are interested in developing a general framework for FA, called the *generalized* factor analysis (GFA), for continuous, discrete, or mixed observed indicator variables, as long as they are from the exponential family of distributions such as Normal, Binomial, and Poisson distributions. Just like the *generalized* linear models (GLMs), which include analysis of variance (ANOVA), linear regression, logistic regression, and Poisson regression as the special cases, we hope that the GFA method extends the standard FA method to explore or verify the measurement structure of continuous latent variables from observed continuous, binary, ordinal, count, or mixed indicator variables in a unified way.

For example, if an investigator is interested in studying **child's health status**, which is an abstract construct, a set of indicator variables may be used to measure the latent variable. Let's consider the following potentially useful indicator variables:

1. **Body strength:** What is her/his body mass index (BMI)? (*Ratio* data)
2. **Physical activity:** How fast can the kid run for a 50 meter distance? (*Continuous* data)
3. **Parents' judgment:** Which one, $1, 2, \cdots, 7$ (from the weakest to the strongest) do you think suitable for ranking your kid's health status among all the kids of the same age? (*Ordinal* data)
4. **Disease history:** How many times had the kid got cold in the past two months? (*Counts* data)

Obviously, these measurements consist of different types of data – ratio, continuous, ordinal, and counts. The standard FA usually assumes that both the observed and latent variables are continuous and, at least, symmetrically distributed. Although several past researches had tried to extend the FA to be used with categorical observed and/or latent variables, it is still difficult to handle indicator variables of such a mixed type. Thus, two questions arise:

1. How can we use the FA method to analyze some special types of noncontinuous data, e.g., *counts* data?
2. Can FA be generalized to analyze indicator variables of a *mixed* type?

4.2 Outline

What we plan to do is to put the standard FA in a more general framework so that different types of indicator variables can be put together for constructing a "generalized" factor model (see below). Just like the *generalized*

linear models (GLMs), which include analysis of variance (ANOVA), linear regression, logistic regression, and Poisson regression as special cases, we hope that the *generalized* factor analysis (GFA) extends the standard FA to discover a latent measurement structure from the observed continuous, binary, ordinal, or count indicator variables in a unified way, as long as they belong to the *exponential family of distributions* such as Normal, Binomial, and Poisson distributions.

Yet, since we have investigated the equivalence between EFA and CFA in the previous chapter, we shall focus on CFA in this chapter. To estimate the factor loadings in a GFA model, we apply the *iterative reweighted least squares* (IRLS) algorithm of GLMs to "linearize" the generalized factor model first, and then use the usual estimation methods of factor analysis to obtain the estimates of the factor loadings. Specifically, we develop independently a unified three-step estimation procedure for GFA, which is similar to the *E-M algorithm* discussed in Bartholomew (1987, Sec. 6.1, pp. 107-115).

On the other hand, we treat the estimation of factor loadings in GFA models as an error-in-variable problem of GLMs, and then take an econometricians' *instrumental variable* (IV) approach for *simultaneous equations model* (SiEM) to estimating factor loadings.

The objective of this project is to develop the estimation methods for GFA models. An outline is given below. We shall conduct simulations to examine and compare numerically the performances of different estimators.

1. **Model Specification:**

   For each indicator variable Y_i, a generalized factor model (GFA) is:

   $$g(\mu_{\boldsymbol{Y}}) = \Lambda_{\boldsymbol{Y}} \boldsymbol{F}$$

   (a) $Y_i \sim$ Exponential family of distributions, $i = 1, \cdots, p$. (p indicator variables)

   (b) g_i: a suitable *link* function for the mean μ_{Y_i} of the indicator variable Y_i.

   (c) $\Lambda_{Y_i} \boldsymbol{F}$: a linear combination of the unobserved m latent variables $\boldsymbol{F}$ and the factor loadings Λ_{Y_i}.

2. **Estimation Methods:**

   (a) The EM-CFA Estimator:

   A main task is to find a feasible way to estimating the factor loadings $\Lambda_{\boldsymbol{Y}}$ for all indicator variables. What is new in our approach is that we apply the *iterative reweighted least squares* (IRLS) algorithm of GLMs to "linearize" the generalized factor model first, and then use the usual estimation method(s) of factor analysis to obtain the estimates of the factor loadings. Specifically, we develop independently a unified three-step estimation procedure for GFA:

- **Step 0:** Set the initial estimates of the factor loadings $\Lambda_{\boldsymbol{Y}}$.
- **Step 1:** Obtain the factor scores $\boldsymbol{F}$.
- **Step 2:** Treat the pseudo-response variables[1] Z_i's derived from the indicator variables Y_i's in the IRLS algorithm of GLM as observed continuous indicator variables to obtain the updated estimates of the factor loadings $\Lambda_{\boldsymbol{Y}}$.
- **Step 3:** Iterate between **Steps 1 and 2 until** the estimates of the factor loadings converge.

This estimation method is similar to the *E-M algorithm* discussed in Bartholomew (1987, Sec. 6.1, pp. 107-115) except that numerical integration in the *E-Step* is avoided on purpose. The details will be given in Section 5.4.

(b) The Instrumental Variable #1 (IV1) Estimator:

- **Step 1:** Obtain the predicted values of $Y_{1|2}$ and $Y_{1|3}$, $\widehat{Y}_{1|2}$ and $\widehat{Y}_{1|3}$, using the ordinary linear model (LM).
- **Step 2:** Fit the GLM for the response variable Y_2 on the covariate $\widehat{Y}_{1|3}$ and Y_3 on the covariate $\widehat{Y}_{1|2}$ using the IRLS algorithm to obtain the estimates of the factor loadings $\Lambda_{\boldsymbol{Y}} = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$ from the GLM's parameters $\Lambda_{\boldsymbol{Y}_2} = [\lambda_{20}, \lambda_2]^{\mathrm{T}}$ and $\Lambda_{\boldsymbol{Y}_3} = [\lambda_{30}, \lambda_3]^{\mathrm{T}}$ respectively. where λ_{20} and λ_{30} are the intercept coefficients, and λ_2 and λ_3 are the slope coefficients.

Notice that this method is the same as that of Carroll and Stefanski (1994). See Carroll and Stefanski (1994) and Carroll, Ruppert, and Stefanski (1995) for the technical details.

(c) The Instrumental Variable #2 (IV2) Estimator:

- **Step 0:** Fit the GLM for the response variable Y_2 on the covariate Y_1 using the IRLS algorithm to obtain the initial estimate of the factor loadings λ_2 which is the GLM's slope coefficient, and fit the GLM for the response variable Y_3 on the covariate Y_1 using the IRLS algorithm to obtain the initial estimate of the factor loadings λ_3 which is the GLM's slope coefficient, which are the naive estimates of $\Lambda_{\boldsymbol{Y}} = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$.
- **Step 1:** Obtain the predicted values of $Y_{1|2}$ and $Y_{1|3}$, $\widehat{Y}_{1|2}$ and $\widehat{Y}_{1|3}$, as the IV using the ordinary linear model (LM).
- **Step 2:** Treat the pseudo-response variables Z_i's derived from the indicator variables Y_i's in the IRLS algorithm of GLM as observed continuous indicator variables to obtain the updated estimates of the factor loadings $\Lambda_{\boldsymbol{Y}} = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$ using the IRLS algorithm through the following

[1] They are also called the *transformed* or *adjusted* response variables in the literatures.

formula, and iterate Z_i and Λ_Y until Λ_Y converge:

$$\begin{aligned}
\Lambda_{Y_2} = [\lambda_{20}, \lambda_2]^{\mathrm{T}} &= \left[Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|3} (\widehat{Y}_{1|3}^{\mathrm{T}} V \widehat{Y}_{1|3})^{-1} \widehat{Y}_{1|3}^{\mathrm{T}} g'(\mu)^{-1} Y_1\right]^{-1} \times \\
&\quad Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|3} (\widehat{Y}_{1|3}^{\mathrm{T}} V \widehat{Y}_{1|3})^{-1} \widehat{Y}_{1|3}^{\mathrm{T}} g'(\mu)^{-1} Z_2, \text{ where } V = Var(Y_2) \\
\Lambda_{Y_3} = [\lambda_{30}, \lambda_3]^{\mathrm{T}} &= \left[Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|2} (\widehat{Y}_{1|2}^{\mathrm{T}} V \widehat{Y}_{1|2})^{-1} \widehat{Y}_{1|2}^{\mathrm{T}} g'(\mu)^{-1} Y_1\right]^{-1} \times \\
&\quad Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|2} (\widehat{Y}_{1|2}^{\mathrm{T}} V \widehat{Y}_{1|2})^{-1} \widehat{Y}_{1|2}^{\mathrm{T}} g'(\mu)^{-1} Z_3, \text{ where } V = Var(Y_3).
\end{aligned}$$

Then, the statistical properties of our estimators will be examined numerically in simulations.

5 Model Specification, Assumptions, and Interpretation

5.1 Factor Analysis (FA) Model

$$\begin{aligned}
x_1 &= \lambda_{11} F_1 + \lambda_{12} F_2 + \cdots + \lambda_{1m} F_m + \delta_1 \\
x_2 &= \lambda_{21} F_1 + \lambda_{22} F_2 + \cdots + \lambda_{2m} F_m + \delta_2 \\
x_3 &= \lambda_{31} F_1 + \lambda_{32} F_2 + \cdots + \lambda_{3m} F_m + \delta_3 \\
\cdots &\quad \cdots \ \cdots \ \cdots \ \cdots \ \cdots \ \cdots \\
x_p &= \lambda_{p1} F_1 + \lambda_{p2} F_2 + \cdots + \lambda_{pm} F_m + \delta_p
\end{aligned}$$

or, in matrix notation,

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{\lambda}_{\boldsymbol{x}} \boldsymbol{F} + \boldsymbol{\delta} \\
E(\boldsymbol{F}) = \mathbf{0}, &\qquad Var(\boldsymbol{F}) = \Phi = [\phi_{ij}]
\end{aligned}$$

where

$$\phi_{ij} = \begin{cases} 1 & \text{for } i = j \\ \phi_{ij} & \text{for } i \neq j \end{cases}$$

$$\begin{aligned}
E(\delta) &= \mathbf{0} \\
Var(\delta) &= \Theta_\delta \text{ is a diagonal matrix} \\
Cov(\delta, F) &= \mathbf{0}.
\end{aligned}$$

5.2 Generalized Factor Analysis (GFA) Model

$$\begin{aligned}
g_{X_1}(\mu_{X_1}) &= \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1m}F_m \\
g_{X_2}(\mu_{X_2}) &= \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2m}F_m \\
g_{X_3}(\mu_{X_3}) &= \lambda_{31}F_1 + \lambda_{32}F_2 + \cdots + \lambda_{3m}F_m \\
\cdots \quad & \quad \cdots \ \cdots \ \cdots \ \cdots \ \cdots \ \cdots \\
g_{X_p}(\mu_{X_p}) &= \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pm}F_m
\end{aligned}$$

or, in matrix notation,

$$\begin{aligned}
g_{\boldsymbol{X}}(\mu_{\boldsymbol{X}}) &= \Lambda_{\boldsymbol{X}}\boldsymbol{F} \\
E(\boldsymbol{F}) = \mathbf{0}, \qquad & Var(\boldsymbol{F}) = \Phi = [\phi_{ij}] \\
Cov(\Lambda_{\boldsymbol{X}}, \boldsymbol{F}) &= \mathbf{0}.
\end{aligned}$$

1. $X_i \sim$ Exponential family of distributions with mean μ_{x_i}, $i = 1, \cdots, p$. (p indicator variables).
2. g_i: A suitable *link* function for the mean μ_{x_i} of the indicator variable X_i.
3. $\Lambda_{x_i}\boldsymbol{F}$: A linear combination of the unobserved m latent variables $\boldsymbol{F}$ and the factor loadings Λ_{x_i}.

6 Estimation

6.1 Tools

Our original ideas about two important tools for estimation are discussed below.

Tool #1: The *IRLS algorithm* linearizes GLMs!

1. **Equivalence:** The IRLS estimator is equivalent to the MLE in GLMs.
2. **Transformation:** At each iteration indexed by (m), the IRLS algorithm takes a special transformation on the original response variable Y_i to obtain a *pseudo*-response variable Z_i, no matter what type of Y_i is, to modify the property of the original response variable such that the original GLM

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$$

becomes a linear regression model

$$E\left[Z_i^{(m)}\right] = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(m)}$$

until the convergence of $\widehat{\boldsymbol{\beta}}$. Most importantly, the derived linear regression model for the pseudo-response variable Z_i at the *last* iteration of the IRLS algorithm provides an "equivalent" linear model for the original GLM in the sense that they have the same values of the regression coefficients.

Yet, instead of writing specific programs for a variety of GLMs according to their likelihood functions to obtain MLEs, one can just apply one single unified IRLS algorithm for all kinds of GLMs.

Tool #2: The estimation methods previously developed for linear factor models may be applied to the derived linear regression model for the pseudo-response variable Z_i at each iteration, including the last one, of the IRLS algorithm.

1. Treat the pseudo-response variable Z_i constructed at the last iteration of the IRLS algorithm for a GLM as if it is a "continuous" response variable of a linear model.
2. Then, apply the usual estimation methods for a linear factor model to obtain estimates of the structural parameters.

6.2 Methods

We now introduce the EM-CFA, IV1, and IV2 methods for estimating the factor loadings of a GFA model.

1. **The EM-CFA Estimator:**
   - **Step 0:** Set the initial estimates of the factor loadings Λ_{x_i}'s, Φ, and Θ_δ. And, obtain Z_i's

$$Z_i = \Lambda_{x_i}\boldsymbol{F} + g_i'(\mu_{x_i})(X_i - \mu_{x_i}) = \Lambda_{x_i}\boldsymbol{F} + \delta_i^*.$$

   - **Step 1:** Treat the pseudo-response variable Z_i's derived from the indicator variables X_i's in the IRLS algorithm (as in GLMs) as observed continuous indicator variables to obtain the updated estimates of the factor loadings Λ_{x_i}, Φ, and Θ_δ, respectively by CFA.
   - **Step 2:** Update the factor scores $\boldsymbol{F}$ for **Step 1**. There are two methods to compute:

(a) Method #1: The regression method

$$\widehat{F}^{(1)} = \widehat{\Lambda}_X^{(1)} \widehat{\Sigma}^{(1)-1} \widehat{Z}^{(0)}.$$

(b) Method #2: The "EFA" method (by calling the EFA function in statistical software, e.g., S-Plus 2000 for Windows).

- **Step 3:** Update the values of Z_i's, then iterate between **Steps 1** and **2** until the estimates of the factor loadings converge.

Notice that as shown in **Table 2**, Rubin's E-M algorithm obtains the same results as the standard CFA method in estimating the factor loadings of a CFA model. See Rubin and Thayer (1982, 1983) for the technical details. According to our experience, the convergence of Rubin's E-M algorithm is very slow so that we choose the standard CFA method to be extended for GFA models.

2. **The Instrumental Variable #1 (IV1) Estimator:**

- **Step 1:** Obtain the predicted values of $Y_{1|2}$ and $Y_{1|3}$, $\widehat{Y}_{1|2}$ and $\widehat{Y}_{1|3}$, using the ordinary linear model (LM).
- **Step 2:** Fit the GLM for the response variable Y_2 on the covariate $\widehat{Y}_{1|3}$ and Y_3 on the covariate $\widehat{Y}_{1|2}$ using the IRLS algorithm to obtain the estimates of the factor loadings $\Lambda_Y = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$ from the GLM's parameters $\Lambda_{Y_2} = [\lambda_{20}, \lambda_2]^{\mathrm{T}}$ and $\Lambda_{Y_3} = [\lambda_{30}, \lambda_3]^{\mathrm{T}}$ respectively. where λ_{20} and λ_{30} are the intercept coefficients, and λ_2 and λ_3 are the slope coefficients.

Notice that this method is the same as that of Carroll and Stefanski (1994). See Carroll and Stefanski (1994) and Carroll, Ruppert, and Stefanski (1995) for the technical details.

3. **The Instrumental Variable #2 (IV2) Estimator:**

- **Step 0:** Fit the GLM for the response variable Y_2 on the covariate Y_1 using the IRLS algorithm to obtain the initial estimate of the factor loadings λ_2 which is the GLM's slope coefficient, and fit the GLM for the response variable Y_3 on the covariate Y_1 using the IRLS algorithm to obtain the initial estimate of the factor loadings λ_3 which is the GLM's slope coefficient, which are the naive estimates of $\Lambda_Y = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$.
- **Step 1:** Obtain the predicted values of $Y_{1|2}$ and $Y_{1|3}$, $\widehat{Y}_{1|2}$ and $\widehat{Y}_{1|3}$, as the IV using the ordinary linear model (LM).
- **Step 2:** Treat the pseudo-response variables Z_i's derived from the indicator variables Y_i's in the IRLS algorithm of GLM as observed continuous indicator variables to obtain the updated estimates

of the factor loadings $\Lambda_Y = [1, \lambda_2, \lambda_3]^T$ using the IRLS algorithm through the following formula, and iterate Z_i and Λ_Y until Λ_Y converge:

$$\begin{aligned}
\Lambda_{Y_2} = [\lambda_{20}, \lambda_2]^T &= \left[Y_1^T g'(\mu)^{-1}\widehat{Y}_{1|3}(\widehat{Y}_{1|3}^T V \widehat{Y}_{1|3})^{-1}\widehat{Y}_{1|3}^T g'(\mu)^{-1} Y_1\right]^{-1} \times \\
&\quad Y_1^T g'(\mu)^{-1}\widehat{Y}_{1|3}(\widehat{Y}_{1|3}^T V \widehat{Y}_{1|3})^{-1}\widehat{Y}_{1|3}^T g'(\mu)^{-1} Z_2, \text{ where } V = Var(Y_2) \\
\Lambda_{Y_3} = [\lambda_{30}, \lambda_3]^T &= \left[Y_1^T g'(\mu)^{-1}\widehat{Y}_{1|2}(\widehat{Y}_{1|2}^T V \widehat{Y}_{1|2})^{-1}\widehat{Y}_{1|2}^T g'(\mu)^{-1} Y_1\right]^{-1} \times \\
&\quad Y_1^T g'(\mu)^{-1}\widehat{Y}_{1|2}(\widehat{Y}_{1|2}^T V \widehat{Y}_{1|2})^{-1}\widehat{Y}_{1|2}^T g'(\mu)^{-1} Z_3, \text{ where } V = Var(Y_3).
\end{aligned}$$

The details of the estimation procedures will be given in the following section.

7 An Illustration: A One-Factor Three-Indicator GFA Model

Without loss of generalizability, we consider the following one-factor three-indicator GFA model

$$\begin{aligned}
Y_{i1} &= \alpha_1 + 1 \cdot F_i + \delta_{Y_{i1}} \\
g_2(\mu_{2i}) &= \text{logit}(\mu_{2i}) = \alpha_2 + \lambda_2 F_i \\
g_3(\mu_{3i}) &= \log(\mu_{3i}) = \alpha_3 + \lambda_3 F_i
\end{aligned}$$

where μ_{2i} and μ_{3i} are the *means* of the indicator variables Y_2 and Y_3, g_2 and g_3 are the *link* functions for μ_{2i} and μ_{3i}, $\delta_{Y_{i1}}$ is the measurement error of the indicator variable Y_1, and the latent variable F is a continuous variable measured by three indicator variables Y_1, Y_2, and Y_3 of a mixed type – *continuous*, *binary*, and *counts* – respectively. For example,

$$F \sim N(0,1) \quad \text{and} \quad \delta_{Y_1} \sim N(0, 0.25).$$

Then, at each iteration in the IRLS algorithm, we construct the following *pseudo*-response variables, Z_1, Z_2, and Z_3, and the corresponding error terms, $\delta_{Y_1}, \delta_{Z_2}$, and δ_{Z_3} as defined below:

$$\begin{aligned}
Z_{i1} &= Y_{i1} = \alpha_1 + 1 \cdot F_i + \delta_{Y_{i1}} \\
Z_{2i} &= \alpha_2 + \lambda_2 F_i + g'_{2i}(\mu_{2i})(Y_{2i} - \mu_{2i}) \\
&= \alpha_2 + \lambda_2 F_i + \frac{Y_{2i} - \mu_{2i}}{\mu_{2i}(1 - \mu_{2i})} \\
&= \alpha_2 + \lambda_2 F_i + \delta_{Z_{2i}} \\
Z_{3i} &= \alpha_3 + \lambda_3 F_i + g'_{3i}(\mu_{3i})(Y_{3i} - \mu_{3i}) \\
&= \alpha_{3i} + \lambda_{3i} F_i + \frac{Y_{3i} - \mu_{3i}}{\mu_{3i}} \\
&= \alpha_{3i} + \lambda_{3i} F_i + \delta_{Z_{3i}}
\end{aligned}$$

where $\alpha_j (j = 1, 2, \text{and } 3)$ are intercepts, $\phi = Var(F)$, $\boldsymbol{\delta} = [\delta_{Y_1}, \delta_{Z_2}, \delta_{Z_3}]^{\mathrm{T}}$, and $\Theta_\delta = Var(\boldsymbol{\delta})$ is a symmetric positive definite matrix.

1. **The EM-CFA Estimator:**

   First, we report the following interesting findings.

   (a) The covariance of the derived residuals δ_{Z_2} and δ_{Z_3} for the pseudo-response variables Z_2 and Z_3 are not zero (see **Tables 3-1** and **3-2**).

   (b) The means of the estimated covariances between δ_{Y_1} and δ_{Z_2}, δ_{Y_1} and δ_{Z_3}, and δ_{Z_2} and δ_{Z_3}, i.e., $\widehat{Cov}(\delta_{Y_1}, \delta_{Z_2})$, $\widehat{Cov}(\delta_{Y_1}, \delta_{Z_3})$, $\widehat{Cov}(\delta_{Z_2}, \delta_{Z_3})$, at the convergence of GLM or LM over 100 repetitions are *not* equal to zero (see **Tables 3-3, 3-4, 3-5** and **3-6**).

   Then, we obtain

$$\begin{aligned}
Var(Z_1) &= Var(\alpha_1 + F + \delta_{Y_1}) = \phi + Var(\delta_{Y_1}) \\
Var(Z_2) &= Var(\alpha_2 + \lambda_2 F + \delta_{Z_2}) = \lambda_2^2 \phi + Var(\delta_{Z_2}) \\
Var(Z_3) &= Var(\alpha_3 + \lambda_3 F + \delta_{Z_3}) = \lambda_3^2 \phi + Var(\delta_{Z_3}) \\
Cov(Z_1, Z_2) &= Cov(\alpha_1 + F + \delta_{Y_1}, \alpha_2 + \lambda_2 F + \delta_{Z_2}) = \lambda_2 \phi + Cov(\delta_{Y_1}, \delta_{Z_2}) \\
Cov(Z_1, Z_3) &= Cov(\alpha_1 + F + \delta_{Y_1}, \alpha_3 + \lambda_3 F + \delta_{Z_3}) = \lambda_3 \phi + Cov(\delta_{Y_1}, \delta_{Z_3}) \\
Cov(Z_2, Z_3) &= Cov(\alpha_2 + \lambda_2 F + \delta_{Z_2}, \alpha_3 + \lambda_3 F + \delta_{Z_3}) = \lambda_2 \lambda_3 \phi + Cov(\delta_{Z_2}, \delta_{Z_3})
\end{aligned}$$

which can be put into a matrix

$$Var(\boldsymbol{Z}) = \begin{bmatrix} \phi + Var(\delta_{Y_1}) & \lambda_2\phi + Cov(\delta_{Y_1}, \delta_{Z_2}) & \lambda_3\phi + Cov(\delta_{Y_1}, \delta_{Z_3}) \\ \lambda_2\phi + Cov(\delta_{Y_1}, \delta_{Z_2}) & \lambda_2^2\phi + Var(\delta_2) & \lambda_2\lambda_3\phi + Cov(\delta_{Z_2}, \delta_{Z_3}) \\ \lambda_3\phi + Cov(\delta_{Y_1}, \delta_{Z_3}) & \lambda_2\lambda_3\phi + Cov(\delta_{Z_2}, \delta_{Z_3}) & \lambda_3^2\phi + Var(\delta_{Z_3}) \end{bmatrix}.$$

Hence, we use the following CFA formulas for this one-factor three-indicator GFA model to obtain the estimates of the unknown parameters λ_2, λ_3, ϕ, $Var(\widehat{\delta}_{Y_1})$, $Var(\widehat{\delta}_{Z_2})$ and $Var(\widehat{\delta}_{Z_3})$ respectively:

$$\begin{aligned}
\widehat{\lambda}_2 &= \frac{Cov(Z_2, Z_3) - Cov(\delta_{Z_2}, \delta_{Z_3})}{Cov(Z_1, Z_3) - Cov(\delta_{Y_1}, \delta_{Z_3})} \\
\widehat{\lambda}_3 &= \frac{Cov(Z_2, Z_3) - Cov(\delta_{Z_2}, \delta_{Z_3})}{Cov(Z_1, Z_2) - Cov(\delta_{Y_1}, \delta_{Z_2})} \\
\widehat{\phi} &= \frac{[Cov(Z_1, Z_2) - Cov(\delta_{Y_1}, \delta_{Z_2})] \times [Cov(Z_1, Z_3) - Cov(\delta_{Y_1}, \delta_{Z_3})]}{Cov(Z_2, Z_3) - Cov(\delta_{Z_2}, \delta_{Z_3})} \\
Var(\widehat{\delta}_{Y_1}) &= Var(Z_1) - \frac{[Cov(Z_1, Z_2) - Cov(\delta_{Y_1}, \delta_{Z_2})] \times [Cov(Z_1, Z_3) - Cov(\delta_{Y_1}, \delta_{Z_2})]}{Cov(Z_2, Z_3) - Cov(\delta_{Z_2}, \delta_{Z_3})} \\
Var(\widehat{\delta}_{Z_2}) &= Var(Z_2) - \frac{[Cov(Z_1, Z_2) - Cov(\delta_{Y_1}, \delta_{Z_2})] \times [Cov(Z_2, Z_3) - Cov(\delta_{Z_2}, \delta_{Z_3})]}{Cov(Z_1, Z_3) - Cov(\delta_{Y_1}, \delta_{Z_3})} \\
Var(\widehat{\delta}_{Z_3}) &= Var(Z_3) - \frac{[Cov(Z_1, Z_3) - Cov(\delta_{Y_1}, \delta_{Z_3})] \times [Cov(Z_2, Z_3) - Cov(\delta_{Z_2}, \delta_{Z_3})]}{Cov(Z_1, Z_2) - Cov(\delta_{Y_1}, \delta_{Z_2})}.
\end{aligned}$$

However, note that the conditional variances of the pseudo-response variables Z_2 and Z_3 for each subject i

$$\begin{aligned}
Var(\delta_{Z_{2i}}) &= Var\left[\frac{Y_{2i} - \mu_{2i}}{\mu_{2i}(1-\mu_{2i})}\right] = \frac{\mu_{2i}(1-\mu_{2i})}{[\mu_{2i}(1-\mu_{2i})]^2} = \frac{1}{\mu_{2i}(1-\mu_{2i})} \\
Var(\delta_{Z_{3i}}) &= Var\left[\frac{Y_{3i} - \mu_{3i}}{\mu_{3i}}\right] = \frac{\mu_{3i}}{\mu_{3i}^2} = \frac{1}{\mu_{3i}}
\end{aligned}$$

are *not* constant, which depend on each subject's response means μ_{2i} and μ_{3i}. Accordingly, the above CFA formulas for the *variances* of the pseudo-response variables Z_2 and Z_3 are questionable due to the heteroscedasticity problem. This finding actually motivates us to also consider the second approach – fitting the GLMs for μ_{2i} and μ_{3i} directly to the data with an imputed $\widehat{F}_i$ or instrumental variable.

Specifically, the six unknown parameters, λ_2, λ_3, ϕ, $Var(\delta_{Y_1})$, $Var(\delta_{Z_2})$, and $Var(\delta_{Z_3})$ in this one-factor three-indicator GFA model are estimated by the following iterative three-step procedure:

- **Step 0:** Set $\widehat{\lambda}_2^{(0)} = \widehat{\lambda}_3^{(0)} = 0$, $\widehat{\mu}_{2i}^{(0)} = \widehat{p}_2 =$ Sample proportion of Y_2, $\widehat{\mu}_{3i}^{(0)} = \widehat{p}_3 =$ Sample proportion

of Y_3, $\widehat{\alpha}_{2i}^{(0)} = \log\left(\frac{\widehat{p}_2}{1-\widehat{p}_2}\right)$, and $\widehat{\alpha}_{3i}^{(0)} = \log\left(\frac{\widehat{p}_3}{1-\widehat{p}_3}\right)$. Then, construct the pseudo-response variables

$$\begin{aligned}
\widehat{Z}_{2i}^{(0)} &= \widehat{\alpha}_{2i}^{(0)} + \widehat{\delta}_{Z_{2i}}^{(0)} = \widehat{\alpha}_{2i}^{(0)} + g_2'\left(\widehat{\mu}_{2i}^{(0)}\right)\left(Y_{2i} - \widehat{\mu}_{2i}^{(0)}\right) = \widehat{\alpha}_{2i}^{(0)} + \frac{Y_{2i} - \widehat{\mu}_{2i}^{(0)}}{\widehat{\mu}_{2i}^{(0)}\left(1 - \widehat{\mu}_{2i}^{(0)}\right)} \\
&= \widehat{\alpha}_{2i}^{(0)} + \widehat{\delta}_{Z_{2i}}^{(0)} \\
\widehat{Z}_{3i}^{(0)} &= \widehat{\alpha}_{3i}^{(0)} + \widehat{\delta}_{Z_{3i}}^{(0)} = \widehat{\alpha}_{3i}^{(0)} + g_3'\left(\widehat{\mu}_{3i}^{(0)}\right)\left(Y_{3i} - \widehat{\mu}_{3i}^{(0)}\right) = \widehat{\alpha}_{3i}^{(0)} + \frac{Y_{3i} - \widehat{\mu}_{3i}^{(0)}}{\widehat{\mu}_{3i}^{(0)}} \\
&= \widehat{\alpha}_{3i}^{(0)} + \widehat{\delta}_{Z_{3i}}^{(0)}
\end{aligned}$$

- **Step 1:** Treat the pseudo-response variables $\boldsymbol{Z} = [Z_1,\ Z_2,\ Z_3]^{\mathrm{T}}$ in the IRLS algorithm as "continuous" variables to conduct a usual factor analysis. That is, treat $Z_1 (= Y_1)$, $\widehat{Z}_2^{(0)}$, and $\widehat{Z}_3^{(0)}$ as continuous response variables; and then, find the estimates of $\alpha_1, \alpha_2, \alpha_3, \lambda_2,\ \lambda_3,\ \phi$, $Var(\delta_{Y_1})$, $Var(\delta_{Z_2})$, and $Var(\delta_{Z_3})$, denoted by $\widehat{\alpha}_1^{(1)}$, $\widehat{\alpha}_2^{(1)}$, $\widehat{\alpha}_3^{(1)}$, $\widehat{\lambda}_2^{(1)}$, $\widehat{\lambda}_3^{(1)}$, $\widehat{\phi}^{(1)}$, $Var(\widehat{\delta}_{Y_1}^{(1)})$, $Var(\widehat{\delta}_{Z_2}^{(1)})$, and $Var(\widehat{\delta}_{Z_3}^{(1)})$ respectively by the CFA method using the estimates of the *marginal* variances of the centered pseudo-response variable $\boldsymbol{Z} = [Z_1,\ Z_2,\ Z_3]^{\mathrm{T}}$, $\widehat{Var}(\boldsymbol{Z})$. Specifically, we center these pseudo-response variables first.

  Since the variance-covariance of $\widehat{\boldsymbol{Z}}^{(0)}$ is equal to the variance-covariance of $\widehat{\boldsymbol{\delta}}_{\boldsymbol{Z}}^{(0)}$, we use the following non-corrected CFA formulas for this one-factor three-indicator GFA model to obtain the following estimates at the first iteration:

$$\begin{aligned}
\widehat{\lambda}_2^{(1)} &= \frac{Cov(Z_2, Z_3)}{Cov(Z_1, Z_3)} \\
\widehat{\lambda}_3^{(1)} &= \frac{Cov(Z_2, Z_3)}{Cov(Z_1, Z_2)} \\
\widehat{\phi}^{(1)} &= \frac{[Cov(Z_1, Z_2)] \times [Cov(Z_1, Z_3)]}{Cov(Z_2, Z_3)} \\
Var(\widehat{\delta}_{Y_1}^{(1)}) &= Var(Z_1) - \frac{[Cov(Z_1, Z_2)] \times [Cov(Z_1, Z_3)]}{Cov(Z_2, Z_3)} \\
Var(\widehat{\delta}_{Z_2}^{(1)}) &= Var(Z_2) - \frac{[Cov(Z_1, Z_2)] \times [Cov(Z_2, Z_3)]}{Cov(Z_1, Z_3)} \\
Var(\widehat{\delta}_{Z_3}^{(1)}) &= Var(Z_3) - \frac{[Cov(Z_1, Z_3)] \times [Cov(Z_2, Z_3)]}{Cov(Z_1, Z_2)}.
\end{aligned}$$

For the further iterations, we use the following CFA formulas for this one-factor three-indicator GFA

- **Step 3:** Since

$$\begin{aligned}\widehat{\mu}_2^{(1)} &= \frac{1}{1+e^{-\widehat{\lambda}_2^{(1)}\widehat{F}^{(1)}}} \\ \widehat{\mu}_3^{(1)} &= e^{\widehat{\lambda}_3^{(1)}\widehat{F}^{(1)}}\end{aligned}$$

we have

$$\begin{aligned}\widehat{\delta}_{Z_2}^{(1)} &= g_2'\left(\widehat{\mu}_2^{(1)}\right)\left(Y_2-\widehat{\mu}_2^{(1)}\right) = \frac{Y_2-\widehat{\mu}_2^{(1)}}{\widehat{\mu}_2^{(1)}\left(1-\widehat{\mu}_2^{(1)}\right)} \\ \widehat{\delta}_{Z_3}^{(1)} &= g_3'\left(\widehat{\mu}_3^{(1)}\right)\left(Y_3-\widehat{\mu}_3^{(1)}\right) = \frac{Y_3-\widehat{\mu}_3^{(1)}}{\widehat{\mu}_3^{(1)}}\end{aligned}$$

Thus, we update the pseudo-response variables

$$\begin{aligned}Z_1 &= Y_1 \\ \widehat{Z}_2^{(1)} &= \widehat{\lambda}_2^{(1)}\widehat{F}^{(1)}+\widehat{\delta}_{Z_2}^{(1)} \\ \widehat{Z}_3^{(1)} &= \widehat{\lambda}_3^{(1)}\widehat{F}^{(1)}+\widehat{\delta}_{Z_3}^{(1)}.\end{aligned}$$

And, go back to **Step 2**. Iterate between **Steps 2** and **3** until $\widehat{\lambda}_2$ and $\widehat{\lambda}_3$ converge.

2. **The Instrumental Variable #1 (IV1) Estimator:**

   - **Step 1:** Obtain the predicted values of $Y_{1|2}$ and $Y_{1|3}$, $\widehat{Y}_{1|2}$ and $\widehat{Y}_{1|3}$, using the ordinary linear model (LM).
   - **Step 2:** Fit the GLM for the response variable Y_2 on the covariate $\widehat{Y}_{1|3}$ and Y_3 on the covariate $\widehat{Y}_{1|2}$ using the IRLS algorithm to obtain the estimates of the factor loadings $\Lambda_{\boldsymbol{Y}} = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$ from the GLM's parameters $\Lambda_{\boldsymbol{Y}_2} = [\lambda_{20}, \lambda_2]^{\mathrm{T}}$ and $\Lambda_{\boldsymbol{Y}_3} = [\lambda_{30}, \lambda_3]^{\mathrm{T}}$ respectively. where λ_{20} and λ_{30} are the intercept coefficients, and λ_2 and λ_3 are the slope coefficients.

   Notice that this method is the same as that of Carroll and Stefanski (1994). See Carroll and Stefanski (1994) and Carroll, Ruppert, and Stefanski (1995) for the technical details.

3. **The Instrumental Variable #2 (IV2) Estimator:**

   - **Step 0:** Fit the GLM for the response variable Y_2 on the covariate Y_1 using the IRLS algorithm to obtain the initial estimate of the factor loadings λ_2 which is the GLM's slope coefficient, and fit the GLM for the response variable Y_3 on the covariate Y_1 using the IRLS algorithm to obtain the initial estimate of the factor loadings λ_3 which is the GLM's slope coefficient, which are the naive estimates of $\Lambda_{\boldsymbol{Y}} = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$.

- **Step 1:** Obtain the predicted values of $Y_{1|2}$ and $Y_{1|3}$, $\widehat{Y}_{1|2}$ and $\widehat{Y}_{1|3}$, as the IV using the ordinary linear model (LM).
- **Step 2:** Treat the pseudo-response variables Z_i's derived from the indicator variables Y_i's in the IRLS algorithm of GLM as observed continuous indicator variables to obtain the updated estimates of the factor loadings $\Lambda_Y = [1, \lambda_2, \lambda_3]^{\mathrm{T}}$ using the IRLS algorithm through the following formula, and iterate Z_i and Λ_Y until Λ_Y converge:

$$\begin{aligned}
\Lambda_{Y_2} = [\lambda_{20}, \lambda_2]^{\mathrm{T}} &= \left[Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|3} (\widehat{Y}_{1|3}^{\mathrm{T}} V \widehat{Y}_{1|3})^{-1} \widehat{Y}_{1|3}^{\mathrm{T}} g'(\mu)^{-1} Y_1\right]^{-1} \times \\
&\quad Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|3} (\widehat{Y}_{1|3}^{\mathrm{T}} V \widehat{Y}_{1|3})^{-1} \widehat{Y}_{1|3}^{\mathrm{T}} g'(\mu)^{-1} Z_2, \text{ where } V = Var(Y_2) \\
\Lambda_{Y_3} = [\lambda_{30}, \lambda_3]^{\mathrm{T}} &= \left[Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|2} (\widehat{Y}_{1|2}^{\mathrm{T}} V \widehat{Y}_{1|2})^{-1} \widehat{Y}_{1|2}^{\mathrm{T}} g'(\mu)^{-1} Y_1\right]^{-1} \times \\
&\quad Y_1^{\mathrm{T}} g'(\mu)^{-1} \widehat{Y}_{1|2} (\widehat{Y}_{1|2}^{\mathrm{T}} V \widehat{Y}_{1|2})^{-1} \widehat{Y}_{1|2}^{\mathrm{T}} g'(\mu)^{-1} Z_3, \text{ where } V = Var(Y_3).
\end{aligned}$$

4. **The Naive Estimator:**

   For the purpose of comparison, we also compute the *naive* estimates of $\widetilde{\lambda}_2$ and $\widetilde{\lambda}_3$ by directly fitting the GLMs with the proxy of the latent variable F to the data, which are also taken as the initial values of the IV2 method.

8 Simulations

8.1 Design

Recall that the previously specified one-factor three-indicator GFA model is

$$\begin{aligned}
Y_1 &= \alpha_1 + 1 \cdot F + \delta_{Y_1} \\
\text{logit}(\mu_2) &= \alpha_2 + \lambda_2 F \\
\log(\mu_3) &= \alpha_3 + \lambda_3 F
\end{aligned}$$

where μ_2 and μ_3 are the *means* of the indicator variables Y_2 and Y_3, the commonly used "logit" and "log" *link* functions are chosen for μ_{2i} and μ_{3i}, δ_{Y_1} is the measurement error of the indicator variable Y_1, and the latent variable F is a continuous variable measured by three indicator variables Y_1, Y_2, and Y_3 of a mixed type – *continuous*, *binary*, and *counts* – respectively.

In the following simulations, the sample size (n) is 1000 and the number of repetition (m) is 500. We follow the following steps to generate the simulated data.

1. Firstly, we set the true parameter values.

   (a) For simplicity, we set all the intercepts to be zero, i.e., $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

   (b) For fixing the scale of F, let $\lambda_1 = 1.0$.

   (c) For comparisons, we choose: $(\lambda_2, \lambda_3) = (0.25, 0.25), (0.5, 0.5), (0.75, 0.75), (1.0, 1.0), (1.25, 1.25), (0.5, 1.0)$, and $(1.0, 0.5)$, respectively.

2. Next, we generate the random variables

$$\begin{aligned} F &\sim Normal(0, 1), \\ \delta_{Y_1} &\sim Normal(0, 0.25). \end{aligned}$$

3. Thirdly, we compute

$$\begin{aligned} \mu_2 &= \frac{1}{1 + e^{-(\alpha_2 + \lambda_2 F)}} \\ \mu_3 &= e^{(\alpha_3 + \lambda_3 F)} \end{aligned}$$

   respectively.

4. Finally, we have

$$\begin{aligned} Y_1 &= \alpha_2 + 1 \cdot F + \delta_{Y_1} \\ Y_2 &\sim Binomial(1, \mu_2) \\ Y_3 &\sim Poisson(\mu_3) \end{aligned}$$

   as the simulated data.

Then, we compute the EM-CFA, IV1, and IV2 estimators in each repetition. For the purpose of comparison, we also compute the *naive* estimates of $\widetilde{\lambda}_2$ and $\widetilde{\lambda}_3$ by directly fitting the GLMs with the proxy of the latent variable F to the data, which are also taken as the initial values of the IV2 method.

8.2 Results

The simulation results are listed in **Table 4**. We find the following interesting results.

1. The naive estimator always seriously underestimates the true values of (λ_2, λ_3).

2. General speaking, the EM-CFA, IV1, and IV2 estimators perform well for an "identity" or a "log" link function.

3. For logistic link,

   (a) EM-CFA estimator performs well except when the true values of $\lambda_2 = 1.25$:

   $$|\widehat{\lambda}_2 - \lambda_2| = |1.4022 - 1.25| = 0.1522.$$

   (b) IV1 estimator performs well except when the true values of $\lambda_2 = 1.0$ and 1.25:

   - $\lambda_2 = 1.0$: $|1.2453 - 1.0| = 0.2453$
   - $\lambda_2 = 1.25$: $|2.2607 - 1.25| = 1.0107$

   (c) IV2 estimator performs well except when the true values of $\lambda_2 = 0.25$:

   $$|\widehat{\lambda}_2 - \lambda_2| = |0.1882 - 0.25| = 0.0618.$$

   Among them, EM-CFA estimator always has a larger variance.

4. When the values of λ_2 and λ_3 are different such as $(0.5, 1.0)$ and $(1.0, 0.5)$, the performance of IV1 has a larger bias.

5. In all these values of (λ_2, λ_3), IV2 always has less bias and smaller variance except when $(\lambda_2, \lambda_3) = (0.25, 0.25)$,

   - $\lambda_2 = 0.25$: $|0.1882 - 0.25| = 0.0618$
   - $\lambda_3 = 0.25$: $|0.1830 - 0.25| = 0.067$

9 Discussions

9.1 Summary

1. Our EM-CFA and IV2 estimators aim to extend the standard FA method for exploring and/or verifying a latent measurement structure from a set of observed *continuous*, *binary*, *ordinal*, *count*, or *mixed* indicator variables in a unified way, as long as they belong to the *exponential family of distributions*, which include Normal, Binomial, and Poisson distributions.

2. Bartholomew (1987) has proposed a class of *logit* FA models for homogeneous *binary* indicator variables and discussed an *EM algorithm* for estimation. Both the model and the estimation method may be considered as special cases of GFA, but, in addition, our EM-CFA estimator avoids the hassle of numerical integration.

3. Yet, as shown in **Table 4**, the EM-CFA and IV1 estimators are away from the true values of (λ_2, λ_3), when the true values are too large (e.g., > 1.0), unless the link function is an "identity" or a "log" function. The explanation for this unsatisfactory result is that when the true values of (λ_2, λ_3) are too large (e.g., > 1.0), the *linear approximation* used by those estimators are not very good.

   See: Buzas and Stefanski (1994) and Liang and Liu (1991) ...

   Note that we have not gone through intensive simulations to find out the exact problematic regions for the values of (λ_2, λ_3) in our settings so that the values of 0.25 and 1.0 are really tentative cutting points. Also, it is worth mentioning that since the latent variable F is a continuous variable, the true values of (λ_2, λ_3) between 0.25 and 1.25 are reasonably useful for logistic and Poisson regression models.

4. Finally, notice that if there are more observed indicator variables available in the data set, then the IV2 estimator can easily take the extra IV's to improve its efficiency in estimation.

5. To sum up, our IV2 estimator is tentatively the best one among those three in terms of *biasedness* and *mean square error* (MSE).

9.2 Future Work

1. First, we shall improve our EM-CFA and IV1 estimators for the GFA models with relatively large values of λ_2 by correcting the bias caused by the poor linear approximation in such situations.

2. To deal with the correlated factors and the correlations among measurement errors, we will further develop the following two methods:

   (a) Oblique rotations for GFA

   (b) $\widehat{F}$ as an IV

   respectively.

3. Then, we should investigate analytically the *statistical properties* of our estimators such as

   (a) consistency,

   (b) asymptotic unbiasedness,

(c) variance estimator, and

(d) asymptotic distribution

for making statistical inference by conducting

- simulations and/or
- a large sample study.

4. And, continue to develop the GFA as a statistical method with future efforts

(a) to improve estimation methods for multiple-factor GFA models,

(b) to develop model-fitting techniques for

- Goodness of fit?
- Model selection?
- Regression diagnostics and Remedies?

(c) to analyze real data,

(d) to design measurement instruments using indicator variables of a mixed type, and

(e) to compare with the other FA methods for non-continuous indicator variables.

10 Appendices

10.1 Appendix 1: The Iteratively Reweighted Least Squares (IRLS) Algorithm for the Two-Stage Least Square Estimator

For the GLM, let

$$Y \sim \text{Exponential family of distributions (with one parameter } \theta)$$

and its log-likelihood function be

$$l(\theta; y) = \log f(y; \theta).$$

And, define

$$U = \frac{\partial l}{\partial \theta} \quad \text{(Score)}.$$

Then, it can be shown (see, e.g., Dobson 1990, **Appendix A, pp. 142-144**) that

$$\begin{aligned} E(U) &= 0, \\ Var(U) &= E(U^2) = E(-U') \quad \textbf{(Information)} \end{aligned}$$

where

$$U' = \frac{\partial U}{\partial \theta}.$$

The joint log-likelihood function for the independent response variables $Y_1, Y_2, \cdots, Y_n$ from a member of the exponential family of distributions is of the following form

$$l(\theta; \mathbf{y}) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i).$$

Then,

$$\mu_i = E(Y_i) = -\frac{c'(\theta_i)}{b'(\theta_i)}.$$

since $E(U) = 0$.

The jth score equation is

$$U_j = \frac{\partial l}{\partial \lambda_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0,$$

and thus

$$\boldsymbol{U} = \frac{\partial l}{\partial \boldsymbol{\Lambda_Y}} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) = \mathbf{0},$$

where

$$\eta_i = g(\mu_i) = \boldsymbol{\Lambda_Y} \boldsymbol{F}_i$$

and

$$\boldsymbol{D} = \begin{bmatrix} \ddots & & 0 \\ & \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} & \\ 0 & & \ddots \end{bmatrix} \boldsymbol{F}.$$

Let $\boldsymbol{A}$ be the IV, which insteed the covariate $\boldsymbol{F}$, then

$$\boldsymbol{U} = \boldsymbol{D}^{\mathrm{T}} \left[g'(\mu)^{-1} A\right]^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) = \mathbf{0},$$

And, the information matrix I has the elements

$$\begin{aligned} I = -E(U') = Var(U) &= \left\{ \left[g'(\mu)^{-1} A \right]^{\mathrm{T}} V^{-1} \right\} Var(Y) \left\{ \left[g'(\mu)^{-1} A \right]^{\mathrm{T}} V^{-1} \right\}^{\mathrm{T}} \\ &= A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} g'(\mu)^{-1} A \end{aligned}$$

When applying the *Newton-Raphson method* to solve for the maximum likelihood estimates (MLEs) $\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}$ of $\mathbf{\Lambda}_{\boldsymbol{Y}}$, at the mth iteration

$$\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m)} = \widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} + \left[I^{(m-1)} \right]^{-1}_{\mathbf{\Lambda}_{\boldsymbol{Y}} = \widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)}} U^{(m-1)}.$$

then

$$\begin{aligned} A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} g'(\mu)^{-1} A \widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m)} &= A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} g'(\mu)^{-1} A \widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} + A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} (Y - \mu(\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)})) \\ &= A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} g'(\mu)^{-1} \left[A \widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} + g'(\mu)(Y - \mu(\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)})) \right] \end{aligned}$$

$$\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m)} = \left[A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} g'(\mu)^{-1} A \right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} V^{-1} g'(\mu)^{-1} \left[A \widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)} + g'(\mu)(Y - \mu(\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}^{(m-1)})) \right]$$

10.2 Appendix 2: The Iteratively Reweighted Least Squares (IRLS) Algorithm for the IV Estimator

Let A be the IV, and we obtain the IV stimator through

$$min_{\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}} Q = min_{\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}} \left[A^{\mathrm{T}} \left(Y - \mu(\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}) \right) \right]^{\mathrm{T}} V^{*-1} \left[A^{\mathrm{T}} \left(Y - \mu(\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}) \right) \right].$$

where

$$V^* = Var \left(A^{\mathrm{T}} \left(Y - \mu(\widehat{\mathbf{\Lambda}}_{\boldsymbol{Y}}) \right) \right) = A^{\mathrm{T}} V A$$

The score equation is

$$U = \frac{\partial l}{\partial \mathbf{\Lambda}_{\boldsymbol{Y}}} = \left[A^{\mathrm{T}} D \right]^{\mathrm{T}} V^{*-1} (Y - \mu) = 0,$$

where

$$\eta_i = g(\mu_i) = \mathbf{\Lambda}_{\boldsymbol{Y}} Y_{1i},$$

and

$$D = \begin{bmatrix} \ddots & & 0 \\ & \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} & \\ 0 & & \ddots \end{bmatrix} Y_1 = g'(\mu)^{-1} Y_1.$$

then

$$U = Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} (Y - \mu)$$

And, the information matrix I

$$I = -E(U') = Var(U) = Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} Y_1$$

When applying the *Newton-Raphson method* to solve for the maximum likelihood estimates (MLEs) $\widehat{\Lambda}_Y$ of Λ_Y, at the mth iteration

$$\widehat{\Lambda}_Y^{(m)} = \widehat{\Lambda}_Y^{(m-1)} + \left[I^{(m-1)}\right]^{-1}_{\Lambda_Y = \widehat{\Lambda}_Y^{(m-1)}} U^{(m-1)}.$$

then

$$\begin{aligned} Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} Y_1 \widehat{\Lambda}_Y^{(m)} &= Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} Y_1 \widehat{\Lambda}_Y^{(m-1)} + \\ & \quad Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} (Y - \mu(\widehat{\Lambda}_Y^{(m-1)})) \end{aligned}$$

thus

$$\begin{aligned} \widehat{\Lambda}_Y^{(m)} &= \left[Y_1^{\mathrm{T}} g'(\mu)^{-1} \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} Y_1\right]^{-1} Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} \\ & \quad \left[Y_1 \widehat{\Lambda}_Y^{(m-1)} + g'(\mu)(Y - \mu(\widehat{\Lambda}_Y^{(m-1)}))\right] \\ &= \left[Y_1^{\mathrm{T}} g'(\mu)^{-1} \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} Y_1\right]^{-1} Y_1^{\mathrm{T}} g'(\mu)^{-1} A \left[A^{\mathrm{T}} V A\right]^{-1} A^{\mathrm{T}} g'(\mu)^{-1} Z^{(m-1)} \end{aligned}$$

where

$$Z^{(m-1)} = Y_1 \widehat{\Lambda}_Y^{(m-1)} + g'(\mu)(Y - \mu(\widehat{\Lambda}_Y^{(m-1)}))$$

11 References

1. Anderson, T. W. (1956). Statistical inference in factor analysis. *The Third Berkeley Symposium in Mathematical Statistics and Probability*, 5, pp. 111-150.

2. Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. New York, NY: John Wiley & Sons.

3. Bartholomew, D. J. (1984). The foundations of factor analysis. *Biometrika*, 71, pp. 221-232.

4. Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. London: Charles Griffin & Company.

5. Basilevsky, A. (1983). *Applied Matrix Algebra in the Statistical Sciences*. New York, NY: North-Holland.

6. Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York, NY: John Wiley & Sons.

7. Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, pp. 443-459.

8. Bock, R. D. and Lieberman, M (1970). Fitting a response model for n dichotomously score items. *Psychometrika*, 35, pp. 179-197.

9. Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons.

10. Bowden, R. J. and Turkington, D. A. (1984). *Instrumental Variables*. New York, NY: Cambridge University Press.

11. Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman & Hall.

12. Carroll, R. J. and Stefanski, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analysis. *Statistics in Medicine*, 13, pp. 1265-1282.

13. Cassella, G. and Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Duxbury Press.

14. Chambers, J. M. and Hastie, T. J. (1993). *Statistical Models in S*. London: Chapman & Hall.

15. Cox, D. R. and Hinkly, D. V. etc. (1991). *Statistical Theory and Modelling*. London: Chapman & Hall.

16. Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*. London: Chapman & Hall.

17. Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*. New York, NY: Springer-Verlag.

18. Godambe, V. P. (1991). *Estimating Functions*. New York, NY: Oxford University Press.

19. Hamilton, L. C. (1992). *Regression with Graphics: A Second Course in Applied Statistics*. Belmont, CA: Duxbury Press.

20. Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute.

21. Jöreskog, K. G. (1993). *Testing Structural Equation Models*. In: K. A. Bollen and J. S. Long (Eds), *Testing Structural Equation Models*. Newbury Park, CA: Sage, pp. 294-316.

22. Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, 4th Ed. London: Prentice-Hall International.

23. Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method.* New York, NY: American Elsevier Publishing Company.

24. McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Ed. London: Chapman & Hall.

25. Mooijaart, A. (1983). Two kinds of factor analysis for ordered categorical variables. *Multivariate Behavioral Research*, 18, pp. 423-441.

26. Mooijaart, A. (1985). Factor analysis for non-normal variables. *Psychometrika*, 50, pp. 323-342.

27. Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. New York, NY: Mcgraw-Hall.

28. Mulaik, S. A. (1986). Factor analysis and Psychometrika: Major developments. *Psychometrika*, 51, pp. 23-33.

29. Mulaik, S. A. and McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, 43, pp. 177-192.

30. Noble, B. and Daniel, J. W. (1977). *Applied Linear Algebra*, 4th Ed. Englewood Cliffs, NJ: Prentice-Hall.

31. Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*, 2nd Ed. New York, NY: Springer-Verlag.

32. Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47, pp. 69-76.

33. Rubin, D. B. and Thayer, D. T. (1983). More on EM for ML factor analysis. *Psychometrika*, 48, pp. 253-257.

34. Schumacker, R. E. and Lomax, R. G. (1996). *A Beginner's Guide to Structural Equation Modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

35. White, H. (1984). *Asymptotic Theory for Econometricians.* Orlando, FL: Academic Press.

12 Tables

Table 1: Classification of Various Factor Analysis Methods

***Latent* Variables (Factors)**	***Observed* Variables**	
	Metrical (Continuous, Ranks)	**Categorical** (Nominal)
Metrical (Continuous, Ranks)	Standard factor analysis (FA)	Latent trait analysis (LTA); FA of multivariate multinomial data
Categorical (Nominal)	Latent profile analysis (LPA)	Latent class analysis (LCA)

Source:

1. Basilevsky (1994, p. 608).
2. Bartholomew (1987, Table 1.1, p. 4).

Table 2: The Simulation Result for the Standard CFA Method and Rubin's E-M Algorithm

Sample size $(n) = 200$. Number of repetition $(m) = 100$.

	λ_1	λ_2	λ_3
True Value	1.0 (-)	0.5 (-)	1.5 (-)
GLM[a]	1.007 (0.07659)	0.5017 (0.07536)	1.493 (0.06809)
CFA	1.0[b] (-)	0.5044 (0.1004)	1.545 (0.2823)
EM	1.0[b] (-)	0.5044 (0.1004)	1.545 (0.2954)
True Value	1.0 (-)	1.0 (-)	1.0 (-)
GLM[a]	1.00718461323299 (0.07659218952307)	1.00165958427255 (0.075359320015064)	0.992911007057482 (0.0680889066046947)
CFA	1.0[b] (-)	1.00642332159071 (0.146486136744037)	1.00390032515117 (0.127779493535428)
EM	1.0[b] (-)	1.00642332159073 (0.146486136744016)	1.0039003251512 (0.127779493535412)
True Value	1.0 (-)	-0.5 (-)	1.5 (-)
GLM[a]	1.01080951875209 (0.0787165950331392)	-0.499908664997619 (0.0620798552717672)	1.50316648639313 (0.0709859641717377)
CFA	1.0[b] (-)	-0.497106238392904 (0.100612888335183)	1.5274259107822 (0.301106425945521)
EM	1.0[b] (-)	-0.497105912444716 (0.100612640991489)	1.52420771991646 (0.2925925393515)

	$Var(\delta_1)$	$Var(\delta_2)$	$Var(\delta_3)$
True Value	1 (-)	1 (-)	1 (-)
GLM[a]	- (-)	- (-)	- (-)
CFA	0.9876 (0.2185)	0.9917 (0.1123)	0.9026 (0.3636)
EM	0.9876 (0.2209)	0.9917 (0.1123)	0.9026 (0.3824)
True Value	1 (-)	1 (-)	1 (-)
GLM[a]	- (-)	- (-)	- (-)
CFA	0.984060934093806 (0.172361031177362)	0.987879423883293 (0.161391795214473)	0.965065743683386 (0.129557416686403)
EM	0.984060934093876 (0.172361031177244)	0.98787942388328 (0.161391795214454)	0.965065743683358 (0.129557416686419)
True Value	1 (-)	1 (-)	1 (-)
GLM[a]	- (-)	- (-)	- (-)
CFA	0.972624637396587 (0.208174371958415)	0.993004872439881 (0.106320764031874)	0.945644141370921 (0.405803006037516)
EM	0.971666938163725 (0.206280358720579)	0.99268945056702 (0.106270157252494)	0.95070315377262 (0.390448600722036)

	ϕ
True Value	1.1.013 (0.1043)
GLM[a]	- (-)
CFA	1.034 (0.281)
EM	1.034 (0.2829)
True Value	1.01284464109456 (0.104256888544777)
GLM[a]	- (-)
CFA	1.03797823394451 (0.246365505074508)
EM	1.03797823394444 (0.246365505074388)
True Value	1.0069160832841 (0.10342782462389)
GLM[a]	- (-)
CFA	1.04607330864509 (0.263226799961674)
EM	1.04703100845262 (0.261897241882429)

a: Take the *true factor scores* F as the observed data, then estimate the parameters by directly fitting the GLM models for comparison.

b: $\lambda_1 = 1$ (to fix the scale of F).

Table 3-1: The Covariance Estimates of the Derived Residuals δ_{Y_1}, δ_{Z_2}, and δ_{Z_3} for the Pseudo-Response Variables Y_1, Z_2, and Z_3 from the Raw Data

$$\begin{aligned} Y_1 &= F + \delta_{Y_1} \\ logit(\mu_2) &= 0.5F \\ log(\mu_3) &= 0.5F \\ \text{Sample size } n = 1000, \quad F &\sim N(0,1), \quad \delta_{Y_1} \sim N(0, 0.25) \end{aligned}$$

Estimators	Cov(F, δ_{Y_1})	Cov(F, δ_{Z_2})	Cov(F, δ_{Z_3})
Mean	0.00145622575741779	-0.000284173424563074	0.00147076032658248
SD	0.014568840760596	0.00797413291220497	0.00592956130792931
Estimators	Cov(δ_{Y_1}, δ_{Z_2})	Cov(δ_{Y_1}, δ_{Z_3})	Cov(δ_{Z_2}, δ_{Z_3})
Mean	0.000725764225032145	0.00395229243467163	**0.00600479275260526**
SD	$\doteq 0$	0.00266820752352165	0.000510252226763834
Estimators	Cov(F, $Y_2 - \mu_2$)	Cov(F, $Y_3 - \mu_3$)	Cov(δ_{Y_1}, $Y_2 - \mu_2$)
Mean	-9.00240386790204e-006	0.00243583555551094	0.000191641208935875
SD	0.00321754784298773	$\doteq 0$	$\doteq 0$
Estimators	Cov(δ_{Y_1}, $Y_3 - \mu_3$)	Cov($Y_2 - \mu_2$, $Y_3 - \mu_3$)	Cov(F, $\frac{1}{\mu_2(1-\mu_2)}$)
Mean	0.00375222926711774	0.00128226971780505	0.00338793663518931
SD	$\doteq 0$	0.00399039021853967	0.0053682934424681
Estimators	Cov(F, $\frac{1}{\mu_3}$)	Cov(δ_{Y_1}, $\frac{1}{\mu_2(1-\mu_2)}$)	Cov(δ_{Y_1}, $\frac{1}{\mu_3}$)
Mean	**-0.566665990748019**	0.000927683128166474	-0.000430536497826355
SD	0.00360798386203851	$\doteq 0$	$\doteq 0$
Estimators	Cov($\frac{1}{\mu_2(1-\mu_2)}$, $\frac{1}{\mu_3}$)		
Mean	**0.0784670653377518**		
SD	$\doteq 0$		

Table 3-2: The Correlation Estimates of the Derived Residuals δ_{Y_1}, δ_{Z_2}, and δ_{Z_3} for the Pseudo-Response Variables Y_1, Z_2, and Z_3 from the Raw Data

$$\begin{aligned} Y_1 &= F + \delta_{Y_1} \\ logit(\mu_2) &= 0.5F \\ log(\mu_3) &= 0.5F \\ \text{Sample size } n = 1000, \quad F &\sim N(0,1), \quad \delta_{Y_1} \sim N(0, 0.25) \end{aligned}$$

Estimators	Corr(F, δ_{Y_1})	Corr(F, δ_{Z_2})	Corr(F, δ_{Z_3})
Mean	0.0028345673646307	0.0000979445140091732	0.00212153666724797
SD	0.0289173297462378	0.0078052289480203	0.00806928280432202
Estimators	Corr(δ_{Y_1}, δ_{Z_2})	Corr(δ_{Y_1}, δ_{Z_3})	Corr(δ_{Z_2}, δ_{Z_3})
Mean	0.000666048183312564	0.00748125926529849	0.00256267478777341
SD	$\doteq 0$	0.00535659923415136	0.00201428345036792
Estimators	Corr(F, $Y_2 - \mu_2$)	Corr(F, $Y_3 - \mu_3$)	Corr(δ_{Y_1}, $Y_2 - \mu_2$)
Mean	0.00014960135394405	0.00207858865524895	0.000762618568084194
SD	0.0064895298526054	$\doteq 0$	$\doteq 0$
Estimators	Corr(δ_{Y_1}, $Y_3 - \mu_3$)	Corr($Y_2 - \mu_2$, $Y_3 - \mu_3$)	Corr(F, $\frac{1}{\mu_2(1-\mu_2)}$)
Mean	0.00706587218221219	0.00244002631560017	0.0082093196641338
SD	$\doteq 0$	0.00803329448853753	0.011709201950677
Estimators	Corr(F, $\frac{1}{\mu_3}$)	Corr(δ_{Y_1}, $\frac{1}{\mu_2(1-\mu_2)}$)	Corr(δ_{Y_1}, $\frac{1}{\mu_3}$)
Mean	**-0.939236332875279**	0.00434795766923775	-0.00140541543868504
SD	0.0018514783958216 9	$\doteq 0$	$\doteq 0$
Estimators	Corr($\frac{1}{\mu_2(1-\mu_2)}$, $\frac{1}{\mu_3}$)		
Mean	**0.32144296061664**		
SD	$\doteq 0$		

Table 3-3: The Covariance Estimates of the Derived Residuals δ_{Y_1}, δ_{Z_2}, and δ_{Z_3} for the Pseudo-Response Variables Y_1, Z_2, and Z_3 after converge of GLM with true F value

$$\begin{aligned} Y_1 &= F + \delta_{Y_1} \\ logit(\mu_2) &= 0.5F \\ log(\mu_3) &= 0.5F \end{aligned}$$

Sample size $n = 1000$, $F \sim N(0,1)$, $\delta_{Y_1} \sim N(0, 0.25)$

Estimators	Cov(F, δ_{Y_1})	Cov(F, δ_{Z_2})	Cov(F, δ_{Z_3})
Mean	0.00145622575741779	0.000294104772107693	**-0.676858302067822**
SD	0.014568840760596	0.00323734237539355	0.0101322486805901
Estimators	Cov(δ_{Y_1}, δ_{Z_2})	Cov(δ_{Y_1}, δ_{Z_3})	Cov(δ_{Z_2}, δ_{Z_3})
Mean	0.000671591683197729	0.00347084543633434	**0.0186538672128944**
SD	$\doteq 0$	$\doteq 0$	$\doteq 0$
Estimators	Cov(F, $Y_2 - \mu_2$)	Cov(F, $Y_3 - \mu_3$)	Cov(δ_{Y_1}, $Y_2 - \mu_2$)
Mean	8.39354798828077e-006	**-507.972862087114**	0.000180144742518826
SD	$\doteq 0$	0.113433835962197	$\doteq 0$
Estimators	Cov(δ_{Y_1}, $Y_3 - \mu_3$)	Cov($Y_2 - \mu_2$, $Y_3 - \mu_3$)	Cov(F, $\frac{1}{\mu_2(1-\mu_2)}$)
Mean	**-8.58564340366774**	**4.80529291743544**	0.00301277113367695
SD	$\doteq 0$	0.220299729452898	0.00901396999578264
Estimators	Cov(F, $\frac{1}{\mu_3}$)	Cov(δ_{Y_1}, $\frac{1}{\mu_2(1-\mu_2)}$)	Cov(δ_{Y_1}, $\frac{1}{\mu_3}$)
Mean	**-2.60434736151852**	0.00127146745550534	**-0.0222005570433585**
SD	0.0222801312663991	$\doteq 0$	$\doteq 0$
Estimators	Cov($\frac{1}{\mu_2(1-\mu_2)}$, $\frac{1}{\mu_3}$)		
Mean	**2.10707836962572**		
SD	$\doteq 0$		

Table 3-4: The Correlation Estimates of the Derived Residuals δ_{Y_1}, δ_{Z_2}, and δ_{Z_3} for the Pseudo-Response Variables Y_1, Z_2, and Z_3 after converge of GLM with true F value

$$\begin{aligned} Y_1 &= F + \delta_{Y_1} \\ logit(\mu_2) &= 0.5F \\ log(\mu_3) &= 0.5F \\ \text{Sample size } n = 1000, \quad F &\sim N(0,1), \quad \delta_{Y_1} \sim N(0, 0.25) \end{aligned}$$

Estimators	Corr(F, δ_{Y_1})	Corr(F, δ_{Z_2})	Corr(F, δ_{Z_3})
Mean	0.0028345673646307	0.000139992452784524	**-0.26370740504046**
SD	0.0289173297462378	0.0032614878463136	0.0103680400671665
Estimators	Corr(δ_{Y_1}, δ_{Z_2})	Corr(δ_{Y_1}, δ_{Z_3})	Corr(δ_{Z_2}, δ_{Z_3})
Mean	0.000684925270228996	0.00291249486011692	0.00235080309367759
SD	$\doteq 0$	$\doteq 0$	$\doteq 0$
Estimators	Corr(F, $Y_2 - \mu_2$)	Corr(F, $Y_3 - \mu_3$)	Corr(δ_{Y_1}, $Y_2 - \mu_2$)
Mean	0.0000172677623888288	**-0.376283658317126**	0.000717427626712922
SD	$\doteq 0$	$\doteq 0$	$\doteq 0$
Estimators	Corr(δ_{Y_1}, $Y_3 - \mu_3$)	Corr($Y_2 - \mu_2$, $Y_3 - \mu_3$)	Corr(F, $\frac{1}{\mu_2(1-\mu_2)}$)
Mean	-0.00472580890234064	-0.000858457664255342	0.00642724068258737
SD	$\doteq 0$	$\doteq 0$	0.0177467284442302
Estimators	Corr(F, $\frac{1}{\mu_3}$)	Corr(δ_{Y_1}, $\frac{1}{\mu_2(1-\mu_2)}$)	Corr(δ_{Y_1}, $\frac{1}{\mu_3}$)
Mean	**-0.381665511121521**	0.00424188830210559	0.00232851813836744
SD	0.0133310865295649	$\doteq 0$	$\doteq 0$
Estimators	Corr($\frac{1}{\mu_2(1-\mu_2)}$, $\frac{1}{\mu_3}$)		
Mean	**0.53210355784553**		
SD	$\doteq 0$		

Table 3-5: The Covariance Estimates of the Derived Residuals δ_{Y_1}, δ_{Z_2}, and δ_{Z_3} for the Pseudo-Response Variables Y_1, Z_2, and Z_3 after converge of LM with true F value

$$\begin{aligned} Y_1 &= F + \delta_{Y_1} \\ logit(\mu_2) &= 0.5F \\ log(\mu_3) &= 0.5F \\ \text{Sample size } n = 1000, \quad F &\sim N(0,1), \quad \delta_{Y_1} \sim N(0, 0.25) \end{aligned}$$

Estimators	Cov(F, δ_{Y_1})	Cov(F, δ_{Z_2})	Cov(F, δ_{Z_3})
Mean	0.00145622575741779	-6.37919258069205e-016	-1.30797019305941e-014
SD	0.014568840760596	$\doteq 0$	$\doteq 0$
Estimators	Cov(δ_{Y_1}, δ_{Z_2})	Cov(δ_{Y_1}, δ_{Z_3})	Cov(δ_{Z_2}, δ_{Z_3})
Mean	0.000653984643916516	0.00399136365924357	**0.00582289303679342**
SD	$\doteq 0$	0.00251405828046894	0.00165182045698597
Estimators	Cov(F, $Y_2 - \mu_2$)	Cov(F, $Y_3 - \mu_3$)	Cov(δ_{Y_1}, $Y_2 - \mu_2$)
Mean	-0.000164903734209712	0.000477548846080722	0.000176219796114244
SD	$\doteq 0$	$\doteq 0$	$\doteq 0$
Estimators	Cov(δ_{Y_1}, $Y_3 - \mu_3$)	Cov($Y_2 - \mu_2$, $Y_3 - \mu_3$)	Cov(F, $\frac{1}{\mu_2(1-\mu_2)}$)
Mean	0.00380409356892336	0.00127948449740848	0.00258503749764027
SD	$\doteq 0$	0.003993404308313	0.00916176299987637
Estimators	Cov(F, $\frac{1}{\mu_3}$)	Cov(δ_{Y_1}, $\frac{1}{\mu_2(1-\mu_2)}$)	Cov(δ_{Y_1}, $\frac{1}{\mu_3}$)
Mean	**-0.577355347966404**	0.00124578427075383	-0.000543822652515056
SD	$\doteq 0$	$\doteq 0$	$\doteq 0$
Estimators	Cov($\frac{1}{\mu_2(1-\mu_2)}$, $\frac{1}{\mu_3}$)		
Mean	**0.0842934951191868**		
SD	$\doteq 0$		

Table 3-6: The Correlation Estimates of the Derived Residuals δ_{Y_1}, δ_{Z_2}, and δ_{Z_3} for the Pseudo-Response Variables Y_1, Z_2, and Z_3 after converge of LM with true F value

$$\begin{aligned} Y_1 &= F + \delta_{Y_1} \\ logit(\mu_2) &\doteq 0.5F \\ log(\mu_3) &= 0.5F \\ \text{Sample size } n = 1000, \quad F &\sim N(0,1), \quad \delta_{Y_1} \sim N(0, 0.25) \end{aligned}$$

Estimators	Corr(F, δ_{Y_1})	Corr(F, δ_{Z_2})	Corr(F, δ_{Z_3})
Mean	0.0028345673646307	-3.0514209163557e-016	-1.21146968498384e-014
SD	0.0289173297462378	$\doteq 0$	1.74065577553716e-009
Estimators	Corr(δ_{Y_1}, δ_{Z_2})	Corr(δ_{Y_1}, δ_{Z_3})	Corr(δ_{Z_2}, δ_{Z_3})
Mean	0.000669452200100126	0.00747509838488736	0.00257984114156831
SD	$\doteq 0$	0.00554438256447983	0.00147153606964539
Estimators	Corr(F, $Y_2 - \mu_2$)	Corr(F, $Y_3 - \mu_3$)	Corr(δ_{Y_1}, $Y_2 - \mu_2$)
Mean	-0.000329480994746257	0.000404443598243263	0.000701573271903001
SD	$\doteq 0$	$\doteq 0$	$\doteq 0$
Estimators	Corr(δ_{Y_1}, $Y_3 - \mu_3$)	Corr($Y_2 - \mu_2$, $Y_3 - \mu_3$)	Corr(F, $\frac{1}{\mu_2(1-\mu_2)}$)
Mean	0.00717976099675851	0.00243512067156767	0.00463904414910382
SD	$\doteq 0$	0.00804802634749676	0.0183376025573628
Estimators	Corr(F, $\frac{1}{\mu_3}$)	Corr(δ_{Y_1}, $\frac{1}{\mu_2(1-\mu_2)}$)	Corr(δ_{Y_1}, $\frac{1}{\mu_3}$)
Mean	**-0.93806825660643**	0.0042501570935359	-0.00135764594156089
SD	0.00414662270356817	$\doteq 0$	$\doteq 0$
Estimators	Corr($\frac{1}{\mu_2(1-\mu_2)}$, $\frac{1}{\mu_3}$)		
Mean	**0.320363054856871**		
SD	$\doteq 0$		

Table 4: The Simulation Results for a One-Factor Three-Indicator GFA Model

$Var(\delta_{Y_1}) = 0.25$. Sample size $(n) = 1000$. Number of repetitions $(m) = 500$.

Estimators		λ_2	λ_3
True Value		0.25	0.25
EM-CFA	Mean	0.25092649693187	0.25077280393814
	SD	(0.27188917804651)	(0.25779815451420)
IV1	Mean	0.23917803559980	0.24175813023387
	SD	(0.25016339007073)	(0.26395594629856)
IV2	Mean	0.18820484324542	0.18296473232450
	SD	(0.18962223003058)	(0.18430516904135)
Naive	Mean	0.20385240382896	0.19896037286348
	SD	(0.05616941128867)	(0.02829132854139)
True Value		0.5	0.5
EM-CFA	Mean	0.506262456399002	0.52789336850906
	SD	(0.169496397921747)	(0.164492467426508)
IV1	Mean	0.479948009028105	0.49864687709039
	SD	(0.143898300631222)	(0.140166490686705)
IV2	Mean	0.491471572991417	0.528938137102815
	SD	(0.14695309863652)	(0.14027973639975)
Naive	Mean	0.394203707107843	0.39879624899906
	SD	(0.060496465011727)	(0.026884175697182)
True Value		0.75	0.75
EM-CFA	Mean	0.770805776916108	0.803566453545486
	SD	(0.148548512532787)	(0.112550796094448)
IV1	Mean	0.755992332797742	0.755170694969821
	SD	(0.125516560084339)	(0.101351033826447)
IV2	Mean	0.741523132695775	0.75203472185432
	SD	(0.122236971276649)	(0.10145359485326)
Naive	Mean	0.584053624312387	0.598897441281704
	SD	(0.061932729997401)	(0.028905630245148)
True Value		1.0	1.0
EM-CFA	Mean	1.06302978540374	1.0371487334856
	SD	(0.185307670785744)	(0.090552921306518 4)
IV1	Mean	1.24528805940761	1.00881819975644
	SD	(0.235194514154233)	(0.0949600132097685)
IV2	Mean	0.977987102582031	1.01534135675285
	SD	(0.139790973079505)	(0.08842560485614)
Naive	Mean	0.765614819012559	0.799832361822084
	SD	(0.071004841008492)	(0.0335064195148907)

Estimators		λ_2	λ_3
True Value		1.25	1.25
EM-CFA	Mean	1.40220449509103	1.16619704621377
	SD	(0.22217889066854)	(0.14573630432616)
IV1	Mean	2.26072152841269	1.24919145135057
	SD	(0.59010017172323)	(0.09365187995617)
IV2	Mean	1.23995950680057	1.27412203789911
	SD	(0.18319664780851)	(0.13824740378273)
Naive	Mean	0.93681694772132	0.98987065901224
	SD	(0.07050462839213)	(0.04518235185069)
True Value		0.5	1.0
EM-CFA	Mean	0.51501650180935	1.01830285913052
	SD	(0.12144561744652)	(0.1272447814200)
IV1	Mean	0.57788827699316	1.00381246264769
	SD	(0.16056702818523)	(0.16801078221817)
IV2	Mean	0.49950397735152	1.06309367337369
	SD	(0.11158406617320)	(0.26484241789319)
Naive	Mean	0.39673543565992	0.79998715259196
	SD	(0.06377698891446)	(0.03481135326788)
True Value		1.0	0.5
EM-CFA	Mean	1.02781523803823	0.525066319497778
	SD	(0.232544586692135)	(0.0894179464868579)
IV1	Mean	0.881558911810525	0.502297498691286
	SD	(0.146926308125686)	(0.0783581593224056)
IV2	Mean	0.98426117666135	0.505227777936546
	SD	(0.189398517849977)	(0.0929382297309208)
Naive	Mean	0.772928740970611	0.4002852853845
	SD	(0.071622819141420)	(0.0289400950712675)

Table 5: The Simulation Results for a One-Factor Three-Indicator GFA Model

$Var(\delta_{Y_1}) = 0.75$. Sample size $(n) = 1000$. Number of repetitions $(m) = 500$.

Estimators		λ_2	λ_3
True Value		0.5	0.5
EM-CFA	Mean	0.50980269530201	0.54086567539394
	SD	(0.16297579284051)	(0.16893066097260)
IV1	Mean	0.48909950906730	0.50672422925337
	SD	(0.14336438742163)	(0.14818503770210)
IV2	Mean	0.48567788071855	0.50633361540587
	SD	(0.13555834589625)	(0.16620030854094)
Naive	Mean	0.28302514728299	0.28543378712282
	SD	(0.05333027571444)	(0.02520645751251)
True Value		0.75	0.75
EM-CFA	Mean	0.761863519261033	0.821334437208613
	SD	(0.149544362544804)	(0.118584875356677)
IV1	Mean	0.761533690363771	0.758161354341419
	SD	(0.129500897862386)	(0.104205697202766)
IV2	Mean	0.712425748525573	0.768824280159445
	SD	(0.116430454571507)	(0.159133311767691)
Naive	Mean	0.410762197762189	0.429172468525754
	SD	(0.0502338619818798)	(0.024651456028187)
True Value		1.0	1.0
EM-CFA	Mean	1.00648708302048	1.06342719346418
	SD	(0.17235372188075)	(0.12208564398683)
IV1	Mean	1.25428239402012	1.01478619076639
	SD	(0.24589805486487)	(0.11023633023196)
IV2	Mean	0.92693799536161	1.04670254551449
	SD	(0.13418419966657)	(0.18470383374334)
Naive	Mean	0.52305716392208	0.56889119940112
	SD	(0.05613333518339)	(0.03508548872038)