ROBUST MEASUREMENT VIA A FUSED LATENT AND GRAPHICAL ITEM RESPONSE THEORY MODEL

Yunxiao Chen, Xiaoou Li, Jingchen Liu, Zhiliang Ying

Abstract

Item response theory (IRT) plays an important role in psychological and educational measurement. Unlike the classical testing theory, IRT models aggregate the item level information, yielding more accurate measurements. Most IRT models assume local independence, an assumption not likely to be satisfied in practice, especially when the number of items is large. Results in the literature and simulation studies in this paper reveal that misspecifying the local independence assumption may result in inaccurate measurements and differential item functioning. To provide more robust measurements, we propose an integrated approach by adding a graphical component to a multidimensional IRT model that can offset the effect of unknown local dependence. The new model contains a confirmatory latent variable component, which measures the targeted latent traits, and a graphical component, which captures the local dependence. An efficient proximal algorithm is proposed for the parameter estimation and structure learning of the local dependence. This approach can substantially improve the measurement, given no prior information on the local dependence structure. The model can be applied to measure both a unidimentional latent trait and multidimensional latent traits.

Key words: item response theory, local dependence, robust measurement, differential item functioning, graphical model, Ising model, pseudo-likelihood, regularized estimator, Eysenck Personality Questionnaire-Revised

1. Introduction

Item response theory (IRT; Rasch, 1960; Lord and Novick, 1968) models play an important role in measurement theory. Unlike classical testing theory, IRT models integrate item level information for measurement and are regarded as being a superior measurement tool to classical test theory (Embretson and Reise, 2000). They have become the preferred method for developing scales, especially when high-stake decisions are involved. In particular, IRT models are used in National Assessment of Education Progress (NAEP), Scholastic Aptitude Test (SAT), and Graduate Record Examination (GRE). Popular IRT models include the single factor models, such as the Rasch model (Rasch, 1960), the two-parameter logistic model, and the three-parameter logistic model (Birnbaum, 1968), and multiple factor models, such as the multidimensional two-parameter logistic (M2PL) model (McKinley and Reckase, 1982; Reckase, 2009).

We use the multidimensional two-parameter logistic model as a building block. Consider an individual responding to J test items and the responses are recorded by a vector $\mathbf{X} = (X_1, ..., X_J)^\top$. To simplify the presentation, we only consider binary items, i.e. $X_j \in \{0, 1\}$, but emphasize that the proposed approach is flexible enough to be generalized to analyzing polytomous items (Chen, 2016). Associated with each response vector is an unobserved continuous latent vector $\boldsymbol{\theta} \in \mathbb{R}^K$, representing the latent characteristics that are measured, where K is the number of latent traits. The model becomes a unidimensional model when $K = 1$. The conditional distribution of each response given the latent vector follows a logistic model

$$f_j(\boldsymbol{\theta}) \triangleq P(X_j = 1|\boldsymbol{\theta}) = \frac{e^{\mathbf{a}_j^\top \boldsymbol{\theta} + b_j}}{1 + e^{\mathbf{a}_j^\top \boldsymbol{\theta} + b_j}},$$

where $f_j(\boldsymbol{\theta})$ is known as the *item response function* and $\mathbf{a}_j = (a_{j1}, ..., a_{jK})^\top$ are known as the *factor loading* parameters. When used in a confirmatory manner, the model imposes constraints on the factor loading parameters, that is, parameter a_{jk} is set to be 0, if item j is not designed to measure the kth latent trait. Such design information is characterized by a $J \times K$ item-trait relationship matrix, which we refer to as the Λ-matrix, $\Lambda = (\lambda_{jk})_{J\times K} = (1_{\{a_{jk}\neq 0\}})_{J\times K}$. The Λ-matrix is usually provided by the item designers and is often assumed to be known. When information about the Λ-matrix is vague, data-driven approaches for learning the Λ-matrix are proposed (Liu *et al.*, 2012, 2013; Chen *et al.*, 2015a; Sun *et al.*, 2016; Chen *et al.*, 2015b; Liu, 2017).

One common assumption of standard IRT models, including the M2PL model, is the so-called *local independence* assumption, i.e. X_1, X_2, ..., X_J are conditionally independent, given the value of $\boldsymbol{\theta}$. That is

$$P(X_1 = x_1, ..., X_J = x_J|\boldsymbol{\theta}) = P(X_1 = x_1|\boldsymbol{\theta})P(X_2 = x_2|\boldsymbol{\theta}) \cdots P(X_J = x_J|\boldsymbol{\theta}), \quad (1.1)$$

for each $\mathbf{x} = (x_1, ..., x_J)^\top \in \{0, 1\}^J$. The local independence assumption implies that, although the items may be highly intercorrelated in the test as a whole, it is only caused by items sharing the common latent traits measured by the test. When the trait levels are controlled, local independence implies that no relationship remains between the items (Embretson and Reise, 2000).

In recent years, computer-based and mobile-app-based instruments are becoming prevalent in educational and psychological studies, where a large number of responses with complex dependence structure are observed. For these tests, a small number of latent traits may not adequately capture the dependence structure among the responses. It is known that there are many possible causes for local dependence, including order effect where responses to early items affect the responses to subsequent items, and shared content effect where additional dependence is caused by a common stimuli from shared content (Hoskens and De Boeck, 1997; Knowles and

Condon, 2000; Schwarz, 1999; Yen, 1993). Generally speaking, the item response process could be complicated, and affected by many external and internal factors. Consequently, a low-dimensional latent factor model may not be adequate to capture all the dependence structure within a test, which may explain the frequently observed phenomenon of model lack of fit in empirical studies (Reise *et al.*, 2011; Yen, 1984, 1993; Ferrara *et al.*, 1999).

In this paper, we propose a *Fused and Latent Graphical IRT* (FLaG-IRT) model to incorporate local dependence as well as to include the test-design information in the Λ-matrix as a priori. The model extends the Fused and Latent Graphical (FLaG) model proposed in Chen *et al.* (2016) by incorporating the loading structure information. The proposed model adds a sparse graphical component upon a multidimensional item response theory (MIRT) model to capture the local dependence. The idea is that for a well designed test, the common dependence among responses has been well explained by the latent traits and the remaining dependence can be characterized by a sparse graphical structure. Moreover, a statistical learning approach is proposed for data-driven learning of the unknown local dependence structure[1].

In psychometrics, there is existing literature on modeling the local dependence structure, including the bi-factor and testlet models (Gibbons and Hedeker, 1992; Gibbons *et al.*, 2007; Reise *et al.*, 2007; Bradlow *et al.*, 1999; Wainer *et al.*, 2000; Wang and Wilson, 2005; Li *et al.*, 2006; Cai *et al.*, 2011), copula based approaches (Braeken *et al.*, 2007; Braeken, 2011), and models with fixed interaction parameters (Hoskens and De Boeck, 1997; Ip, 2002; Ip *et al.*, 2004; Ip, 2010). Most of these approaches require prior information on the local dependence structure, such as knowing the item clusters and assuming the local independence between items clusters, while the proposed approach handles unknown local dependence structure. The proposed FLaG-IRT model is also closely connected to three lines of research in psychometrics:

[1] An R package and example code for the proposed approach can be downloaded from `http://www.scientifichpc.com/flagirt.html`.

(1) psychometric network models and their applications (van der Maas *et al.*, 2006; Cramer *et al.*, 2010, 2012; van Borkulo *et al.*, 2014; Boschloo *et al.*, 2015; Fried *et al.*, 2015; Rhemtulla *et al.*, 2016), (2) log-multiplicative association model (Holland, 1990; Anderson and Vermunt, 2000; Anderson and Yu, 2007; Marsman *et al.*, 2015; Epskamp *et al.*, 2016; Kruis and Maris, 2016), and (3) the use of graphical models for structural violations of local independence (Epskamp *et al.*, 2017; Pan *et al.*, 2017).

The contribution of this paper is of two-folds. First, it provides a rich class of locally dependent IRT models that can capture complex local dependence patterns. Second, a statistically solid and computationally efficient procedure is developed for learning the local dependence structure from data, for which no prior information is needed on the way the items are locally dependent on each other. Consequently, the proposed approach substantially generalizes the traditional methods which may not be flexible enough to capture various types of local dependence patterns and require prior knowledge (e.g. the specification of item clusters in using the bi-factor model).

The rest of the paper is organized as follows. In Section 2, the FLaG-IRT model is introduced and a review of related works is provided. In Section 3, the statistical analysis based on the model, including parameter estimation and model selection, is presented. Results of simulation studies are reported in Section 4. Section 5 contains an application to a real data example.

2. FLaG-IRT Model

2.1. Two Basic Models

We first describe the fused and latent graphical IRT model, which is built upon the multidimensional 2-parameter logistic (M2PL) model and the Ising model (Ising, 1925). To begin with, we describe these two building-block models.

MIRT model. The M2PL model is one of the most popular multidimensional IRT models for binary responses. The item response function of the M2PL model is given by

$$P(X_j = 1|\boldsymbol{\theta}) = \frac{e^{\mathbf{a}_j^\top \boldsymbol{\theta} + b_j}}{1 + e^{\mathbf{a}_j^\top \boldsymbol{\theta} + b_j}}.$$

The item-trait relationship is incorporated by constraints specified by a pre-specified matrix $\Lambda = (\lambda_{jk})_{J \times K}$, $\lambda_{jk} \in \{0, 1\}$, where $\lambda_{jk} = 0$ means that item j is not associated with latent trait k and the corresponding loading a_{jk} is constrained to be 0. The item response function can be further written as

$$P(X_j = x_j|\boldsymbol{\theta}) = \frac{e^{(\mathbf{a}_j^\top \boldsymbol{\theta} + b_j)x_j}}{1 + e^{\mathbf{a}_j^\top \boldsymbol{\theta} + b_j}} \propto \exp\{(\mathbf{a}_j^\top \boldsymbol{\theta} + b_j)x_j\}.$$

The notation "$\propto$" above is used to define probability density or mass functions when the left-hand side and the right-hand side are different by a normalizing constant that depends only on the parameters and is free of the value of the random variable/vector.

Under the M2PL model, the joint distribution of the responses $\mathbf{X} = (X_1, ..., X_J)^\top$ given $\boldsymbol{\theta}$ can be further written as, due to the local independence assumption,

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{J} P(X_j = x_j|\boldsymbol{\theta}) \propto \exp\{\boldsymbol{\theta}^\top A^\top \mathbf{x} + \mathbf{b}^\top \mathbf{x}\}, \tag{2.1}$$

where $A = (a_{jk})_{J \times K}$ is known as the factor loading matrix and $\mathbf{b} = (b_1, ..., b_J)^\top$. In particular, when $K = 1$, the model is known as the two-parameter logistic model (2PL; Birnbaum, 1968).

Ising model. We now present the Ising model that is used to characterize the local dependence structure on top of the M2PL model. The Ising model is an undirected graphical model (e.g. Koller and Friedman, 2009). It encodes the conditional independence relationships among X_j's through the topological structure of a graph that can greatly facilitate the interpretation and understanding of the dependence structure. This model is originated in statistical physics (Ising, 1925).

Specification of the Ising model consists of an undirected graph $G = (V, E)$, where V and E are the sets of vertices and edges respectively. The vertex set $V =$

$\{1, 2, ..., J\}$ corresponds to the random variables, $X_1, ..., X_J$. The graph is said to be undirected in the sense that $(i, j) \in E$, if and only if $(j, i) \in E$. The Ising model associated with an undirected graph $G = (V, E)$ is specified as

$$P(\mathbf{X} = \mathbf{x}) \propto \exp\left\{\frac{1}{2}\mathbf{x}^\top S\mathbf{x}\right\}, \tag{2.2}$$

where $S = (s_{ij})_{J\times J}$ is a symmetric matrix such that $s_{ij} \neq 0$ if and only if $(i, j) \in E$.

The conditional independence relationship in the Ising model is encoded by the topological structure of the graph. More precisely, let A, B and C be nonoverlapping subsets of V and $A \cup B \cup C = V$. We further let $\mathbf{X}_A$, $\mathbf{X}_B$, and $\mathbf{X}_C$ be the random vectors associated with the sets A, B, and C, respectively, i.e., $\mathbf{X}_A = (X_i : i \in A)$ and so on. We say A and B are separated by C, if every path from a vertex in A to a vertex in B includes at least one vertex in C, as illustrated by an example in Figure 1. In Figure 1, $A = \{1, 2\}$, $B = \{4, 5\}$, and $C = \{3\}$, and all paths from A to B pass through C. For example, the path $(1 \to 3 \to 4)$ that connects vertices 1 and 4, passes through vertex 3. In particular, $(i, j) \notin E$ implies X_i and X_j are independent given others. When C is an empty set, the separation between A and B implies their independence.

==========================

Insert Figure 1 about here

==========================

The Ising model can be understood based on the conditional distribution of one variable given all the others. Specifically, we denote $\mathbf{X}_{-j} = (X_1, ..., X_{j-1}, X_{j+1}, ..., X_J)$. Then (2.2) implies that

$$P(X_j = 1|\mathbf{X}_{-j} = \mathbf{x}_{-j}) = \frac{\exp\left(\frac{1}{2}s_{jj} + \sum_{i\neq j} s_{ij}x_i\right)}{1 + \exp\left(\frac{1}{2}s_{jj} + \sum_{i\neq j} s_{ij}x_i\right)}, \tag{2.3}$$

which takes a logistic regression form. The model parameters can be interpreted based on (2.3). Specifically, $s_{jj}/2$ is the log-odds of $X_j = 1$ given $\mathbf{X}_{-j} = (0, ..., 0)$ and s_{ij} is the log-odds-ratio of $X_j = 1$ associated with X_i given all the other variables. In particular, based on (2.3), X_i does not affect the conditional distribution (2.3) when $s_{ij} = 0$ (i.e. $(i, j) \notin E$). This relationship is symmetric, in the sense that $s_{ij} = 0$ also implies that X_j does affect the conditional distribution $P(X_i = 1|\mathbf{X}_{-i})$, since S is a symmetric matrix.

2.2. FLaG-IRT Model

The FLaG-IRT model combines the M2PL model (2.1) and the Ising model (2.2) to construct a joint item response function. More precisely, the conditional distribution is assumed to take the form

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, A, S) \propto \exp\left\{\boldsymbol{\theta}^\top A^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top S\mathbf{x}\right\}. \tag{2.4}$$

This conditional model is an Ising model with parameter matrix $S(\boldsymbol{\theta})$, where $s_{ij}(\boldsymbol{\theta}) = s_{ij}$ for $i \neq j$ and $s_{jj}(\boldsymbol{\theta}) = \mathbf{a}_j^\top \boldsymbol{\theta} + s_{jj}$. In addition, the graph of model (2.4) is the same as that encoded by S, that is, $E = \{(i, j) : s_{ij} \neq 0, i \neq j\}$. Moreover, when the graph is degenerate, i.e. $s_{ij} = 0$, for all $i \neq j$,

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, A, S) \propto \exp\left\{\boldsymbol{\theta}^\top A^\top \mathbf{x} + \sum_{j=1}^{J} \frac{1}{2} s_{jj} x_j^2\right\} = \exp\left\{\boldsymbol{\theta}^\top A^\top \mathbf{x} + \sum_{j=1}^{J} \frac{1}{2} s_{jj} x_j\right\},$$

which takes the same form as that of the M2PL model (2.1) if reparameterizing $b_j = s_{jj}/2$. Note $\sum_j s_{jj} x_j^2 = \sum_j s_{jj} x_j$ since $x_j \in \{0, 1\}$.

Similar to (2.3), model (2.4) can be understood through the conditional distribution of X_j given $\boldsymbol{\theta}$ and $\mathbf{X}_{-j}$. More precisely,

$$P(X_j = 1|\boldsymbol{\theta}, \mathbf{X}_{-j} = \mathbf{x}_{-j}) = \frac{\exp(\frac{1}{2}s_{jj} + \sum_{k=1}^{K} a_{jk}\theta_k + \sum_{i\neq j} s_{ij}x_i)}{1 + \exp(\frac{1}{2}s_{jj} + \sum_{k=1}^{K} a_{jk}\theta_k + \sum_{i\neq j} s_{ij}x_i)},$$

taking a logistic form. Consequently, the model parameters can be interpreted similarly based on the log-odds and log-odds-ratios as the ones in (2.3). In particular, a_{jk} is the log-odds-ratio of X_j associated with one unit increase in θ_k. When $s_{ij} = 0$ for all $i \neq j$,

$$P(X_j = 1|\boldsymbol{\theta}, \mathbf{X}_{-j} = \mathbf{x}_{-j}) = \frac{\exp(\frac{1}{2}s_{jj} + \sum_{k=1}^K a_{jk}\theta_k)}{1 + \exp(\frac{1}{2}s_{jj} + \sum_{k=1}^K a_{jk}\theta_k)},$$

implying that X_j and $\mathbf{X}_{-j}$ are conditionally independent given $\boldsymbol{\theta}$ and the item response function takes the same form as in the M2PL model. Moreover, given $\boldsymbol{\theta}$, the distribution of X_is only depends on its neighbors. For example, consider $K = 1$, $J = 3$, $A = (1, 1, 1)^\top$, and

$$S = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}.$$

S-matrix encodes a graph with three nodes: node 1 is connected to both nodes 2 and 3; nodes 2 and 3 are not connected. In this example, the joint distribution of (X_1, X_2, X_3) given θ_1 becomes

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3|\theta_1) \\ =&\frac{\exp(x_1x_2 - x_1x_3 + \theta_1x_1 + \theta_1x_2 + \theta_1x_3)}{\sum_{x_1', x_2', x_3' = 0,1} \exp(x_1'x_2' - x_1'x_3' + \theta_1x_1' + \theta_1x_2' + \theta_1x_3')}. \end{aligned}$$

Simple calculation gives

$$\begin{aligned} P(X_1 = 1|\theta_1, X_2 = x_2, X_3 = x_3) &= \frac{\exp(\theta_1 + x_2 - x_3)}{1 + \exp(\theta_1 + x_2 - x_3)}, \\ P(X_2 = 1|\theta_1, X_1 = x_1, X_3 = x_3) &= \frac{\exp(\theta_1 + x_1)}{1 + \exp(\theta_1 + x_1)}, \\ P(X_3 = 1|\theta_1, X_1 = x_1, X_2 = x_2) &= \frac{\exp(\theta_1 - x_1)}{1 + \exp(\theta_1 - x_1)} \end{aligned}$$

which allow us to interpret the relationship among X_1, X_2, X_3, and θ_1 based on odds-ratios. For example, given X_2 and X_3, the log-odds-ratio of X_1 associated with one unit increase in θ_1 is 1. In addition, given θ_1 and X_2, the log-odds-ratio of X_1 associated with X_3 is -1, implying a negative association between X_1 and X_3 when the other variables are controlled.

To assist understanding, Figure 2 provides graphical representations of the MIRT model and the FLaG-IRT model. The left panel shows a graphical representation of the marginal distribution of responses, where there is an edge between each pair of responses. Under the conditional independence assumption (1.1) of the MIRT model, there exists a latent vector $\boldsymbol{\theta}$. If we include $\boldsymbol{\theta}$ in the graph, then there is no edge among X_js as in the middle panel. The concern is that this conditional independence structure may be oversimplified and there is additional dependence not attributable to the latent traits. The FLaG-IRT model (right panel) is a natural extension of the MIRT model (middle panel), allowing edges among X_js even if $\boldsymbol{\theta}$ is included. The additional edges capture the dependence among X_js not explained by $\boldsymbol{\theta}$. Due to the presence of the latent variables, it is likely that we only need a small number of additional edges to capture the local dependence. Furthermore, the loading structure in Λ is reflected by the edges between θ_ks and the responses X_js in the middle and right panels.

==========================

Insert Figure 2 about here

==========================

We consider the following joint distribution of $(X, \boldsymbol{\theta})$,

$$f(\mathbf{x}, \boldsymbol{\theta} | A, S, \Sigma) = \frac{1}{z_0(A, S, \Sigma)} \exp\Big\{ -\frac{1}{2}\boldsymbol{\theta}^\top \Sigma^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^\top A^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top S \mathbf{x} \Big\}, \tag{2.5}$$

where (A, S, Σ) are the model parameters and $z_0(A, S, \Sigma)$ is the normalizing constant,

$$z_0(A, S, \Sigma) = \sum_{\mathbf{x}\in\{0,1\}^J} \int \exp\Big\{ -\frac{1}{2}\boldsymbol{\theta}^\top\Sigma^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^\top A^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top S\mathbf{x}\Big\} d\boldsymbol{\theta}.$$

Note that under this joint distribution, the joint item response function, i.e., the conditional distribution of $\mathbf{X}$ given $\boldsymbol{\theta}$, is consistent with (2.4). Under this joint distribution, a specific prior distribution of $\boldsymbol{\theta}$ is implicitly assumed, under which the posterior distribution of $\boldsymbol{\theta}$ is Guassian. Moreover, the prior distribution of $\boldsymbol{\theta}$ can be derived from (2.5), that is,

$$\begin{aligned} f(\boldsymbol{\theta}|A, S, \Sigma) &= \sum_{\mathbf{x}\in\{0,1\}^J} f(\mathbf{x}, \boldsymbol{\theta}|A, S, \Sigma) \\ &= \frac{\sum_{\mathbf{x}\in\{0,1\}^J} \exp\Big\{ -\frac{1}{2}\boldsymbol{\theta}^\top\Sigma^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^\top A^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top S\mathbf{x}\Big\}}{z_0(A, S, \Sigma)}, \end{aligned}$$

taking the form of a mixture of Gaussian distributions. This prior distribution of $\boldsymbol{\theta}$ brings technical convenience in the data analysis (see equation (2.8)). More precisely, under this model, $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$ follows Gaussian distribution

$$N(\Sigma A^\top\mathbf{x}, \Sigma), \tag{2.6}$$

for which the posterior variance is Σ and the posterior mean is given by

$$E(\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}) = \Sigma A^\top\mathbf{x}, \tag{2.7}$$

a weighted sum of the responses. Once A and Σ are estimated from the data, it is reasonable to score each individual by $\hat{\Sigma}\hat{A}^\top\mathbf{x}$.

In the specification (2.5), A, Σ, S, and the graph E induced by S (equivalently, the nonzero pattern of matrix S) can be estimated from the data. Similar to the M2PL model, we pre-specify a binary matrix $\Lambda = (\lambda_{jk})_{J\times K}$ for the confirmatory structure and impose constraint that $a_{jk} = 0$ if $\lambda_{jk} = 0$. Since the latent vector $\boldsymbol{\theta}$ is not directly observable, parameter estimation is based on the marginal likelihood,

$$P(\mathbf{X} = \mathbf{x} | A, S, \Sigma) = \int f(\mathbf{x}, \boldsymbol{\theta} | A, S, \Sigma) d\boldsymbol{\theta},$$

where $f(\mathbf{x}, \boldsymbol{\theta} | A, S, \Sigma)$ is given in (2.5). From a straightforward integration over $\boldsymbol{\theta}$, the marginal distribution of $\mathbf{X}$ still follows an Ising model, that is

$$P(\mathbf{X} = \mathbf{x} | A, S, \Sigma) = \int f(\mathbf{x}, \boldsymbol{\theta} | A, S, \Sigma) d\boldsymbol{\theta} \propto \exp\left\{\frac{1}{2}\mathbf{x}^\top (A\Sigma A^\top + S)\mathbf{x}\right\}. \qquad (2.8)$$

It is worth pointing out that this is a second-order generalized log-linear model (Holland, 1990; Laird, 1991). In fact, Holland (1990) considers a special case of (2.8) for which the graph is degenerate (i.e., S is a diagonal matrix). As shown in Corollary 1 of Holland (1990), this second-order generalized log-linear model can be obtained under a joint distribution of $\mathbf{X}$ and $\boldsymbol{\theta}$, under which $\mathbf{X}$ given $\boldsymbol{\theta}$ follows an M2PL model and $\boldsymbol{\theta}$ given $\mathbf{X}$ is multivariate Gaussian.

2.3. Related Works and Discussions

In what follows, we first review related works and make connections to the proposed approach. Then discussions are provided on extending the proposed FLaG-IRT model to more general response types.

FLaG exploratory analysis. The proposed model is similar to the FLaG model considered in Chen *et al.* (2016) except that the loading structure Λ is prespecified for the former. Both papers consider item response analysis in the presence of local dependence. However, the scopes and the goals of the two papers are different, which further lead to different analyses and computational algorithms. Chen *et al.* (2016) focuses on the recovery of the major latent factors underlying an item pool under an exploratory item factor analysis setting, where the number of major latent factors and their loading structures, as well as the local dependence structure, are unknown.

Chen *et al.* (2016) shows that by adjusting for the local dependence using a graphical model component, the number of major latent factors and their loading structure can be consistently recovered. On the other hand, the current paper studies the use of the FLaG model as a measurement model, under a setting similar to confirmatory item factor analysis but with an unknown local dependence structure. As will be shown in the rest of the paper, the proposed approach automatically adjusts for local dependence structure, substantially reducing the measurement bias induced by the local dependence structure.

Bi-factor models. The bi-factor model is one of the most popular models to incorporate dependence. This model is a special case of the M2PL model, assuming that there is a unidimensional general factor θ_g associated with all items and is the target of measurement. Besides the general factor, there exist nuisance factors $\theta_1, ..., \theta_M$ associated with M nonoverlapping item clusters C_1, C_2, ..., C_M, where each item cluster has no less than two items and there may be items not belonging to any of these item clusters. The bi-factor model based on a logistic link (e.g. Cai *et al.*, 2011) is a special M2PL model with

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}) \propto \exp\{\boldsymbol{\theta}^\top A^\top \mathbf{x} + \mathbf{b}^\top \mathbf{x}\}, \tag{2.9}$$

where $\boldsymbol{\theta} = (\theta_g, \theta_1, ..., \theta_M)$, $\mathbf{b} = (b_1, ..., b_J)^\top$ and $A = (\mathbf{a}_g, \mathbf{a}_1, ..., \mathbf{a}_M)$. In particular, the jth element of $\mathbf{a}_k$ is zero if item j is not in the kth item cluster, i.e., $j \notin C_k$.

Such a bi-factor model structure can be captured by the proposed FLaG-IRT model. Specifically, if we use the specific joint distribution of $(\mathbf{X}, \boldsymbol{\theta})$ as in the FLaG-IRT model and further assume Σ to be an identity matrix, i.e.

$$f(\mathbf{x}, \boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top A^\top \mathbf{x} + \mathbf{b}^\top \mathbf{x}\right\},$$

then the marginal distribution of $\mathbf{X}$ becomes

$$P(\mathbf{X} = \mathbf{x}) \propto \exp\left\{\frac{1}{2}\mathbf{x}^\top \mathbf{a}_g \mathbf{a}_g^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top S \mathbf{x}\right\}, \tag{2.10}$$

where $s_{jj} = 2b_j$, and $s_{ij} = s_{ji} = 0$ when items i and j do not belong to the same item cluster and $s_{ij} = s_{ji} = a_{ik}a_{jk}$ when both items belong to the kth cluster, which admits the same form as the marginal FLaG-IRT model in (2.8). In other words, the graphical model component of the FLaG-IRT model can take the place of the specific factors in the bi-factor model. The corresponding graph encoded by the S matrix in (2.10) is sparse, when each item cluster has only a small number of items. For example, if each item cluster has only two items, then the sparsity level of the graph, defined as the ratio between the number of edges in the graph and the total number of item pairs, is $1/(J-1)$, which can be as small as 3% with $J = 30$ items. Figure 3 presents an example of the a bi-factor model, the corresponding FLaG-IRT model, and the local dependence graph. In other words, when the specific prior for $\boldsymbol{\theta}$ is assumed, the bi-factor model becomes a special case of the FLaG-IRT model with one latent trait and a sparse local dependence graph. One of the advantages of the FLaG-IRT model is that there is no need to specify a priori item clusters and they are learned from the data.

==========================

Insert Figure 3 about here

==========================

Psychometric network models. The proposed method is also connected to, but different from, network modeling of psychometric problems (van der Maas *et al.*, 2006; Cramer *et al.*, 2010, 2012; van Borkulo *et al.*, 2014; Boschloo *et al.*, 2015; Fried *et al.*, 2015; Rhemtulla *et al.*, 2016), where no latent variable is considered. In these models, psychometric item responses are conceived of as proxies for variables that directly interact with each other, instead of being dominated by a few latent factors.

In particular, the Ising model is used as a psychometric network model when the item responses are binary. The key difference between the proposed model and the psychometric network models is that the proposed one maintains a latent variable component that can be used for measurement. In addition, upon the existence of latent factors whose effects spread out to the item responses, one typically needs a network model with a dense graph (e.g. the left panel of Figure 2) to fit the data well, resulting in lack of visualizability and interpretability.

Log-multiplicative association model. The proposed FLaG-IRT model, according to the joint distribution of $(\mathbf{X}, \boldsymbol{\theta})$ in (2.5), is also closely related to the log-multiplicative association model. That is, when the graphical component is degenerate, i.e. $s_{ij} = 0$, for all $i \neq j$, the joint model (2.5) of $\mathbf{X}$ and $\boldsymbol{\theta}$ is a log-multiplicative association model, whose use as an IRT model has been discussed in Holland (1990); Anderson and Vermunt (2000); Anderson and Yu (2007); Marsman *et al.* (2015); Epskamp *et al.* (2016); Kruis and Maris (2016). Empirical evidences show that the log-multiplicative association model and traditional IRT models perform similarly (e.g. Anderson and Yu, 2007).

Graphical modeling in structural equation models. Recent works on structural equation modeling, including Epskamp *et al.* (2017) and Pan *et al.* (2017), consider a similar idea of capturing local dependence structure by a sparse graphical model. In these works, the observed variables are continuous and are assumed to follow a multivariate Gaussian model with latent variables. Such a model assumes that given the latent variables, the observed variables, instead of being conditionally independent, follows a sparse Gaussian graphical model (e.g. Koller and Friedman, 2009). Statistical procedures for learning the sparse graphical component are also developed in Epskamp *et al.* (2017) and Pan *et al.* (2017). The developments in the current paper are independent of and parallel to that of Epskamp *et al.* (2017) and Pan *et al.* (2017), under the context of item response analysis where the observed variables are binary.

Extension to more general response types. The proposed FLaG-IRT model can be extended to analyzing responses of mixed types, under an exponential family model framework (Chen, 2016; Lee and Hastie, 2015). Let $\mathbf{X}$ be the response vector, containing discrete variables or a combination of both continuous and discrete variables. Then the joint distribution of $\mathbf{X}$ and $\boldsymbol{\theta}$ can be specified as

$$f(\mathbf{x}, \boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^\top \Sigma^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^\top A^\top \mathbf{s}(\mathbf{x}) + \frac{1}{2}\mathbf{t}(\mathbf{x})^\top S\mathbf{t}(\mathbf{x})\right\}, \tag{2.11}$$

where $\mathbf{s}(\mathbf{x}) = (s_1(x_1), ..., s_J(x_J))^\top$ and $\mathbf{t}(\mathbf{x}) = (t_1(x_1), ..., t_J(x_J))^\top$ are transformations of the original data, where $s_j(x_j)$ and $t_j(x_j)$ can be vectors. For example, if $X_j \in \{0, 1, ..., c_j\}$ is a discrete variable, we can set $s_j(x_j)$ and/or $t_j(x_j)$ to be $(1_{\{x_j=1\}}, ..., 1_{\{x_j=c_j\}})$ and if X_j is continuous, we set $s_j(\cdot)$ and $t_j(\cdot)$ to be identity functions. The dimensions of matrices A and S depend on the choices of $\mathbf{s}(\cdot)$ and $\mathbf{t}(\cdot)$. The S matrix may contain constraints, depending the data types. Specifically, when all items are binary, model (2.11) becomes the same as (2.5). When all item responses are ordinal, model (2.11) can be viewed as a combination of a multidimensional partial credit model (Yao and Schwarz, 2006) and an undirected graphical model for categorical variables. When all the responses are continuous, the model above becomes the same Gaussian model considered in Epskamp *et al.* (2017). The statistical inference and computation procedures described below can be adapted to this generalized FLaG-IRT model.

3. FLaG-IRT Analysis

3.1. Regularized Pseudo-likelihood Estimation

In this section, we discuss estimation and dimension reduction of the FLaG-IRT model. The most natural approach would be the maximum marginal likelihood function of responses given in (2.8). Unfortunately, the evaluation of (2.8) involves computing the normalizing constant,

$$z(A,S,\Sigma)=\sum_{\mathbf{x}\in\{0,1\}^J}\exp\left\{\frac{1}{2}\mathbf{x}^\top(A\Sigma A^\top+S)\mathbf{x}\right\},$$

which requires a summation over 2^J all possible response patterns and thus is computationally infeasible for even a relatively small J. To bypass this, we propose a pseudo-likelihood as a surrogate (Besag, 1974), which is based on the conditional distribution of X_j given the rest $\mathbf{X}_{-j}=(X_1,...,X_{j-1},X_{j+1},...,X_J)$,

$$P(X_j=1|\mathbf{X}_{-j}=\mathbf{x}_{-j},A,S,\Sigma)=\frac{\exp\{\frac{1}{2}(l_{jj}+s_{jj})+\sum_{i\neq j}(l_{ij}+s_{ij})x_i\}}{1+\exp\{\frac{1}{2}(l_{jj}+s_{jj})+\sum_{i\neq j}(l_{ij}+s_{ij})x_i\}},$$

where $L=(l_{ij})_{J\times J}=A\Sigma A^\top$. Note that the above conditional distribution takes a logistic regression form. Following Besag (1974), we let $\mathcal{L}_j(A,S,\Sigma;\mathbf{x})=P(X_j=x_j|\mathbf{X}_{-j}=\mathbf{x}_{-j},A,S,\Sigma)$ and define the pseudo-likelihood function

$$\mathcal{L}(A,S,\Sigma)=\prod_{i=1}^N\prod_{j=1}^J\mathcal{L}_j(A,S,\Sigma;\mathbf{x}_i),\tag{3.1}$$

where $\mathbf{x}_i$ is the responses from individual i.

The above pseudo-likelihood function is related to, but different from the vertex-wise sparse logistic regression approach for learning a sparse Ising graphical model (e.g. van Borkulo *et al.*, 2014). Under the sparse Ising graphical model, the conditional distribution of each variable X_j given the rest $\mathbf{X}_{-j}$ follows a sparse logistic regression model. Consequently, the neighbors of each vertex j can be learned by regressing X_j on all the other variables $\mathbf{X}_{-j}$ and selecting the variables with nonzero regression coefficients (van Borkulo *et al.*, 2014; Ravikumar *et al.*, 2010; Barber and Drton, 2015). The entire graph is constructed by aggregating vertex-wise information. In the FLaG-IRT model, learning the graphical component requires knowledge about the latent factor component parameterized by A and Σ, which has to be learned from the entire data. Consequently, the learning of the FLaG-IRT model cannot be decomposed into solving vertex-wise regression problems separately. By aggregating the likelihood functions of vertex-wise logistic regressions, the pseudo-likelihood function (3.1) contains information about S, A, and Σ simultaneously and thus can be used for the model selection and parameter estimation.

To incorporate the knowledge of the test items, the factor loading matrix A is constrained such that $a_{jk} = 0$ when $\lambda_{jk} = 0$, noting that the matrix $\Lambda = (\lambda_{jk})$ is pre-specified. Therefore, the unknown parameters in A are $\{a_{jk} : \lambda_{jk} = 1\}$. Since A and Σ appear in the pseudo-likelihood function in the form of $A\Sigma A^\top$, additional constraints are needed to ensure their identifiability. This is because, for example, scaling A by a constant ω can be offset by the corresponding scaling of Σ by ω^{-2}. To identify the scale of latent factors, we impose constraints $\Sigma_{kk} = 1$, $k = 1, ..., K$. To avoid rotational indeterminacy, we assume that with appropriate column swapping, the Λ matrix contains a $K \times K$ identity submatrix. It means that for each latent factor, there is at least one item that only measures that factor.

When the graph for local dependence is known, we estimate A, S, and Σ using a maximum pseudo-likelihood function

$$\begin{aligned}
(\hat{A}, \hat{S}, \hat{\Sigma}) = &\operatorname*{arg\,min}_{A,S,\Sigma} \left\{ -\frac{1}{N} \log \mathcal{L}(A, S, \Sigma) \right\} \\
&s.t.\ a_{jk} = 0 \text{ if } \lambda_{jk} = 0, j = 1, ..., J, k = 1, ..., K, \\
&\qquad S = S^\top, s_{ij} = 0 \text{ if } (i, j) \notin E, \\
&\qquad \text{and } \Sigma \text{ is positive semidefinite, } \Sigma_{kk} = 1, k = 1, ..., K,
\end{aligned} \tag{3.2}$$

where E is the set of edges of the known graph.

When the graph for local dependence is unknown, which is typically the case in practice, we impose an assumption that the graph is sparse, that is, the number of edges in $E = \{(i, j) : s_{ij} \neq 0\}$ is relatively small. The rationale is that most of the dependence among responses has been captured by the common latent traits, leaving the local dependence structure sparse. This assumption is incorporated in the analysis through selecting a sparse graphical model component based on the data. We'd like to point out that even for a sparse local dependence structure (i.e. a local dependence graph with a relatively small number of edges), if ignored in the measurement, can result in measurement bias, as illustrated by simulated examples. In addition, the sparse local dependence graph, once learned from the data, facilitates the understanding of the measurement and may be used to improve the test design. For example, patterns (e.g. item clusters) identified from the graph may help the test designers to review the items and improve the wording.

We propose to use the regularized pseudo-likelihood for simultaneous estimation and model selection

$$\begin{aligned}(\hat{A}^{\gamma}, \hat{S}^{\gamma}, \hat{\Sigma}^{\gamma}) = \underset{A,S,\Sigma}{\arg\min} &\left\{ -\frac{1}{N} \log \mathcal{L}(A, S, \Sigma) + \gamma \sum_{i \neq j} |s_{ij}| \right\} \\ s.t.\ a_{jk} = 0 \text{ if } &\lambda_{jk} = 0, j = 1, ..., J, k = 1, ..., K, \\ &S = S^{\top},\ \text{and } \Sigma \text{ is positive semidefinite}, \sigma_{kk} = 1, k = 1, ..., K,\end{aligned} \tag{3.3}$$

where γ is a tuning parameter that controls the sparsity level of the estimated graph $\hat{E}^{\gamma} = \{(i, j) : \hat{s}^{\gamma}_{ij} \neq 0, i \neq j\}$. At one extreme, when γ is sufficiently large, the estimated graph becomes degenerate, i.e., no edge, and the responses are conditionally independent given the latent variables that are measured. The graph becomes more and more dense as γ decreases.

The optimization problem (3.3) is nonconvex and nonsmooth, and thus is computationally nontrivial. An efficient and stable algorithm is developed, which alternates between minimizing A, S, and Σ. In particular, an proximal gradient based method (Parikh and Boyd, 2014) is used in updating S, which avoids the issues due to the nonsmoothness of the function that may occur in standard gradient based optimization approaches. Details of the algorithm are provided in the appendix in the online supplementary material.

3.2. Choice of Tuning Parameters

In the estimation, we construct a solution path of $(\hat{A}^\gamma, \hat{S}^\gamma, \hat{\Sigma}^\gamma)$ for a sequence of γ values. We then choose γ based on an extended Bayes information criterion (EBIC; Chen and Chen, 2008; Foygel and Drton, 2010; Barber and Drton, 2015), which takes the form

$$\mathrm{EBIC}_\rho(\mathcal{M}) = -2\log L(\hat{\beta}(\mathcal{M})) + |\mathcal{M}|(\log N + 4\rho\log(J)),$$

where $\mathcal{M}$ is the model under consideration, $L(\hat{\beta}(\mathcal{M}))$ is the maximal likelihood for model $\mathcal{M}$, $|\mathcal{M}|$ is the number of free parameters, and $\rho \in [0,1]$ is a parameter that indexes the criterion and has a Bayesian interpretation (Chen and Chen, 2008). When $\rho = 0$, the criterion becomes the classical Bayes information criterion (Schwarz, 1978). Positive ρ leads to stronger penalization when the model space is large (i.e. when J is large). In this study, we replace the likelihood function with the pseudo-likelihood function. Specifically, let

$$\mathcal{M}^\gamma = \big\{(A, S, \Sigma) : a_{jk} = 0 \text{ if } \lambda_{jk} = 0, S = S^\top, s_{ij=0} \text{ if } \hat{s}_{ij} = 0,$$
$$\text{and } \Sigma \text{ is positive semidefinite}, \sigma_{kk} = 1, k = 1, ..., K\}$$

be the model selected by tuning parameter γ, containing all models having the same support as $\hat{S}^\gamma$. We select the tuning parameter γ, such that the corresponding model minimizes the pseudo-likelihood-based EBIC

$$\mathrm{EBIC}_\rho(\mathcal{M}^\gamma) = -2 \max_{(A,S,\Sigma)\in M^\gamma} \{\log \mathcal{L}(A, S, \Sigma)\} + |\mathcal{M}^\gamma|(\log N + 4\rho\log(J)), \tag{3.4}$$

where the number of parameters in $\mathcal{M}^\gamma$ is

$$|\mathcal{M}^\gamma| = \sum_{j,k} \lambda_{jk} + J + \sum_{i<j} 1_{\{\hat{s}^\gamma_{ij}\neq 0\}} + \frac{(K-1)K}{2}.$$

Here, $\sum_{j,k} \lambda_{jk}$ counts the number of free parameters in the loading matrix A, J and $\sum_{i<j} 1_{\{\hat{s}^\gamma_{ij}\neq 0\}}$ are the numbers of diagonal and off-diagonal parameters in $\hat{S}^\gamma$, and

$K(K-1)/2$ is the number of parameters in Σ.

The tuning parameter is finally selected by

$$\hat{\gamma}_\rho = \arg\min_{\gamma} \mathrm{EBIC}_\rho(\mathcal{M}^\gamma). \tag{3.5}$$

In addition, the corresponding maximal pseudo-likelihood estimates of A, S, and Σ are used as the final estimate of A, S, and Σ:

$$(\hat{A}, \hat{S}, \hat{\Sigma})_\rho = \arg\max_{(A,S,\Sigma)\in\mathcal{M}^{\hat{\gamma}_\rho}} \{\mathcal{L}(A, S, \Sigma)\}. \tag{3.6}$$

In the rest of the paper, following Barber and Drton (2015), $\rho = 0, 0.25$, and 0.5 are used.

3.3. Summary

We summarize the procedure of FLaG-IRT analysis, when the graph for local dependence is unknown.

1. Select a sequence of γ values, denoted by Γ.

2. Obtain a sequence of models indexed by $\gamma \in \Gamma$, based on the regularized estimates $(\hat{A}^\gamma, \hat{S}^\gamma, \hat{\Sigma}^\gamma)$ from (3.3).

3. Among the sequence of models above, select the best fitted model in terms of EBIC_ρ value, using (3.5).

4. Report $(\hat{A}, \hat{S}, \hat{\Sigma})_\rho$ from the selected model given by (3.6), as well as the local dependence graph given by $\hat{E}_\rho = \{(i, j) : (\hat{s}_{ij})_\rho \neq 0\}$.

The default values ρ are chosen as 0, 0.25, and 0.5, reflecting different prior beliefs on the size of the model space. In practice, the sequence of γ values in step 1 is chosen in two stages. First, coarse grid points (e.g. $\gamma = 10^{-3}, 10^{-2.5}, 10^{-2}, ...$) are used to anchor a reasonable range, for which the sparsity level of the estimated graph is of interest (e.g. from below 5% to above 40%). Then finer grids are placed in this range for more refined analysis. We also remark that the regularized estimator is mainly used to produce a short list of candidate models, which are further compared and selected by the EBIC. Unregularized parameter estimate is reported for the selected model, which has the advantage of a smaller bias comparing to the regularized one (e.g. Belloni and Chernozhukov, 2013).

4. Simulation Studies

In this section, we report two simulation studies. First, we provide a study exploring the consequence of ignoring local dependence and the effectiveness of the proposed FLaG-IRT model. Second, we evaluate the performance of the FLaG-IRT analysis, when data are generated from a FLaG-IRT model. An additional simulation study is reported in the supplementary material that assesses the performance of FLaG-IRT analysis under model misspecification.

4.1. Study 1

Data generation. We generate data from the bi-factor model (2.9), with $N = 1000$, $J = 15$, and only one item cluster $C_1 = \{1, 2, 3, 4, 5\}$. Note that the general factor θ_g and the nuisance factor θ_1 are assumed to be independent and follow the standard normal distribution. The setting mimics a test that aims at measuring the general factor θ_g, and thus every item is designed to be associated with this dimension. In addition, θ_1 is a nuisance dimension that is only associated with five items and is not included in the design. For ease of exposition, we set $a_{jg} = 1.5$, $j = 1, 2, ..., J$ and $a_{j1} = c$, $j = 1, ..., 5$. The value of c is positive and will be varied to account for different local dependence levels. In addition, b_js are sampled from uniform distribution over interval $[-2, 2]$. For each value of c, 100 independent data sets are generated.

Comparison. In this study, we compare three models, including (1) the unidimensional 2PL model, (2) the bi-factor model with known nuisance factor, and (3) the proposed FLaG-IRT model with known local dependence graph. Specifically, the graph of the FLaG-IRT model is set to be $E = \{(i,j) : i,j \leq 5\}$ and the specific values of s_{ij} remain to be estimated. Note that this FLaG-IRT model is a misspecified model that approximates the generating one.

The measurement of the general factor is compared for the three models above. For a given model, a two-stage procedure is adopted. In the first stage, the model parameters are estimated, and then in the second stage, each person i is measured by the expected a posteriori (EAP) score $\hat{\theta}_i$ computed under the estimated model. Note that for the bi-factor model, $\hat{\theta}_i$ refers to the EAP score of the general factor. We investigate the *measurement accuracy* based on sample correlation between $\hat{\theta}_i$ and the true general factor score θ_{ig}s. In addition, *measurement bias* is investigated based on the sample correlation between $\hat{\theta}_i$ and the nuisance factor score θ_{i1}. For better comparison, we consider three correlation measures, including Kendall's tau rank correlation, Spearman's rho rank correlation, and Pearson's correlation. We point out that as Kendall's tau and Spearman's rho are both rank-based measures that do no rely on specific distribution assumptions, they may be more objective measures for the comparison than Pearson's correlation.

Results. Results are shown in Figure 4, where the left and right panels reflect the measurement accuracy (correlations between $\hat{\theta}_i$s and θ_{ig}s) and the measurement bias (correlations between $\hat{\theta}_i$s and θ_{i1}s), respectively. In each panel, the x-axis records the value of c, where the level of local dependence increases as c increases. Each point is an average over 100 independent data sets. From Figure 4, under all local dependence levels, the proposed FLaG-IRT model with a known graph performs similarly as the bi-factor model, in terms of both measurement accuracy and bias. Moreover, the 2PL model that ignores the local dependence structure performs poorly. Specifically, when local dependence becomes severe, the Kendall's tau, Spearman's rho, and Pearson's correlations between $\hat{\theta}_i$s and θ_{ig}s based on the 2PL model can drop to 0.3, 0.4, and 0.4, respectively, while they remain to be 0.7, 0.9, and 0.9, respectively, for both the bi-factor and FLaG-IRT models. In addition, when local dependence becomes more severe, the three correlation measures between $\hat{\theta}_i$s and θ_{i1}s based on the 2PL model increase and can be as high as 0.6, 0.8, and 0.8, respectively, while the ones based on the bi-factor and FLaG-IRT models are all below 0.1. In other words, the latent trait being measured under the 2PL model deviates from what is designed to measure. This could lead to the issue of test fairness that could especially be of concern in educational testing. That is, for two examinees with the same θ_g value, the one with a higher nuisance trait level tends to be scored higher. This phenomenon is known as differential item functioning (Holland and Wainer, 2012).

==========================

Insert Figure 4 about here

==========================

4.2. Study 2

In this study, we evaluate the performance of the FLaG-IRT analysis in Section 3, under the settings that data are generated from a FLaG-IRT model. In this FLaG-IRT analysis, the local dependence structure is completely unspecified and learned from data.

Data generation. We consider the following model settings.

S1. We consider $J = 45$, $K = 3$ and the local dependence graph $E = \{(i, j) : |i - j| \leq 1\}$. For the loading structure, items 1-15, 16-30, and 31-45 measure the three latent traits, respectively. If particular, we set $a_{jk} = 0.4$ for $q_{jk} \neq 0$, $s_{jj} = -4$, $j = 1, ..., J$, $s_{ij} = 0.5$ for $(i, j) \in E$, and $\sigma_{kk} = 1$, $k = 1, ..., K$ and $\sigma_{kl} = 0.1$, $k \neq l$.

S2. We consider $J = 100$, $K = 5$ and the local dependence graph $E = \{(i, j) :$

$|i-j| \leq 1\}$. For the loading structure, items 1-20, 21-40, 41-60, 61-80, and 81-100 measure the five latent traits, respectively. If particular, we set $a_{jk} = 0.35$ for $q_{jk} \neq 0$, $s_{jj} = -4.5$, $j = 1, ..., J$, $s_{ij} = 1$ for $(i,j) \in E$, and $\sigma_{kk} = 1$, $k = 1, ..., K$ and $\sigma_{kl} = 0.1$, $k \neq l$.

For each setting, sample sizes $N = 500, 1000,$ and 3000 are considered. For each setting and each sample size, 100 independent data sets are generated.

Evaluation criteria. For each data set, model selection results are obtained from the FLaG-IRT analysis under the extended Bayesian criterion with $\rho = 0, 0.25, 0.5$. The selected models are evaluated based on the following criteria.

1. The Kendall's tau correlation between the EAP score $\hat{\theta}_{ik}$s and the corresponding true factor score, θ_{ik}s, $k = 1, ..., K$. An average of the Kendall's tau correlations over K latent traits is reported.

2. The true positive rate of graph estimation, defined as

$$TPR = \frac{\sum_{i<j} 1_{\{(i,j)\in\hat{E},(i,j)\in E\}}}{\sum_{i<j} 1_{\{(i,j)\in E\}}}.$$

3. The false positive rate of graph estimation, defined as

$$FPR = \frac{\sum_{i<j} 1_{\{(i,j)\in\hat{E},(i,j)\notin E\}}}{\sum_{i<j} 1_{\{(i,j)\notin E\}}}.$$

4. The accuracy in parameter estimation is also evaluated for the selected model based on the mean square error (MSE).

Results. Results are presented in Tables 1 and 2. In Table 1, the column "Oracle" gives the values of Kendall's tau, TPR, and FPR when the true model and its parameters are known. Given the true model, the oracle values of TPR and FPR are 1 and 0, respectively. The oracle value of Kendall's tau is the correlation between the EAP scores under the true model and the true scores. According to Table 1, under both settings, all sample sizes, and all values of ρ in the EBIC, the models selected by the FLaG-IRT analysis has high measurement accuracy. The Kendall's tau correlation between the EAP scores under the selected model and the true factor scores is very close to the oracle one. In addition, it is observed that a larger value of ρ in the EBIC yields both lower TPR and lower FPR. This is because a larger value of ρ penalizes more on the model complexity, resulting in a more sparse graph. Furthermore, as sample size increases, the TPR and FPR tend to increase and decrease, respectively. When the sample size is as large as 3000, under both settings, the TPR and FPR are close to 1 and 0, respectively, implying that the true model is accurately selected. Table 2 shows the results on parameter estimation. In particular, we show the MSE for the estimation of a_{11}, s_{11}, and σ_{12}, calculated based on the 100 independent replications. According to the data generation model, these results are representative of those of nonzero a_{jk}s, s_{jj}s, and σ_{kl}s, respectively, which are freely estimated and are not under model selection. According to Table 2, we see that the MSEs become smaller when the sample size increases. In addition, the models selected by the EBIC ($\rho = 0.25, 0.5$) tend to have smaller MSEs than the ones selected by the BIC ($\rho = 0$) and thus have more accurate estimates. Finally, we point out that even under the setting S2 where $J = 100$, $K = 5$, and under the sample size $N = 3000$, the proposed algorithm solves the optimization problem (3.3) for the regularized estimator efficiently. For a given tuning parameter, (16) can be solved within three minutes on an Intel(R) machine (Core(TM) i5-5300U CPU @ 2.30GHz), with code written in R. The algorithm can be further speeded up by writing the code in a more efficient language such as C++ and by parallel computing.

==========================

Insert Table 1 about here

==========================

==========================

Insert Table 2 about here

==========================

5. Real Data Analysis

We illustrate the use of FLaG-IRT analysis through an application to the Extroversion short scale of the Eysenck's Personality Questionnaire-Revised (EPQ-R; Eysenck *et al.*, 1985; Eysenck and Barrett, 2013). The data set contains the responses to 12 items from 842 females in the United States. All these items are designed to measure a single personality trait *Extroversion*, characterized by personality patterns such as sociability, talkativeness, and assertiveness. The items are shown in Table 3, and the data are preprocessed so that the responses to the reversely worded items are flipped.

==========================

Insert Table 3 about here

==========================

We start with fitting the unidimensional 2PL model whose unidimensional latent trait follows a standard Gaussian distribution and then check the model fit. The estimated 2PL parameters are shown in Table 4. Under the fitted model, the expected two-by-two tables for item pairs can be evaluated by

$$E_{x_i x_j} = N \times \widehat{P}(X_i = x_i, X_j = x_j) = N \int \frac{\exp{(\hat{a}_i\theta + \hat{b}_i)x_i}}{1+\exp{(\hat{a}_i\theta + \hat{b}_i)}} \frac{\exp{(\hat{a}_j\theta + \hat{b}_j)x_j}}{1+\exp{(\hat{a}_j\theta + \hat{b}_j)}} \phi(\theta) d\theta,$$

where $\phi(\theta)$ is the density function of a standard normal distribution. We first check the fit of item pairs by comparing the expected two-by-two tables with the observed ones, using the X^2 local dependence index (Chen and Thissen, 1997) as a descriptive statistic. For each item pair i and j, the X^2 statistic is defined as

$$X_{ij}^2 = \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} \frac{(O_{x_i x_j} - E_{x_i x_j})^2}{E_{x_i x_j}},$$

where $O_{x_i x_j}$ is the observed number of (x_i, x_j) pairs. A large value of X^2_{ij} indicates a lack of fit. In addition, based on simulation studies, Chen and Thissen (1997) suggest that the marginal distribution of each X^2_{ij} is roughly a chi-square distribution with one degree of freedom when data are generated from the 2PL model. We visualize $(X^2_{ij})_{J\times J}$ using a heat map in the left panel of Figure 5. For a better visualization, we plot a monotone transformation of X^2_{ij},

$$T_{ij} = X^2_{ij}/(Q^{Chi}_{1,95\%} + X^2_{ij}),$$

where $Q^{Chi}_{1,95\%}$ is the 95% quantile of the chi-square distribution with one degree of freedom. Thus, $T_{ij} > 1/2$ suggests that item pair (i, j) is not fitted well. In the heat map, the value of T_{ij} is presented according to the color key above the heat map. The top four item pairs with highest levels of T_{ij} are shown in Table 5, where items within a pair tend to share common content/stimuli. To further assess the overall fit of the 2PL model and to compare it with that of the selected FLaG-IRT model, we consider a parametric bootstrap test, using the total sum of the X^2 statistics as the test statistic $SX_{2PL} = \sum_{i<j} X^2_{ij}$. That is, we generate 500 bootstrap data sets, each of which has 842 samples drawn from the estimated 2PL model. For each bootstrap data set, we fit the 2PL model again and compute the corresponding total sum of X^2s, denoted by $SX^{(b)}_{2PL}$. The empirical distribution of $SX^{(b)}_{2PL}$ is used as the reference distribution. The histogram of $SX^{(1)}_{2PL}, ..., SX^{(500)}_{2PL}$ is shown in the left panel of Figure 6. The observed value of SX_{2PL} based on the fitted model is 192, much larger than the ones from bootstrap data. Consequently, the p-value of this bootstrap test is 0, indicating the lack-of-fit of the 2PL model.

==========================

Insert Figure 5 about here

==========================

==========================

Insert Figure 6 about here

===========================

We apply the FLaG-IRT analysis. Using the BIC for model selection, the local dependence graph of the selected model has 12 edges, as shown in Figure 7, where the positive and negative edges are in black and red, respectively. In particular, the most locally dependent item pairs also correspond to the most positive edges in Figure 7. Similar to the analysis above, we compute the local independence indices for all the items pairs and visualize them in the right panel of Figure 5, where no X_{ij}^2 is found to exceed $Q_{1,95\%}^{Chi}$. Moreover, 500 bootstrap data sets are generated from the selected FLaG-IRT model and the bootstrap distribution of SX_{FLaG} is shown in the right panel of Figure 6. As we can see, the observed value of SX_{FLaG} for the selected model is within the range of the bootstrap distribution with a p-value 9%, which does not show strong evidence of model lack-of-fit.

===========================

Insert Figure 7 about here

===========================

Based on the above analysis, we see that even a well designed 12-item EPQ-R short form displays significant level of local dependence, which, if not adjusted, may result in measurement bias. The proposed FLaG-IRT model automatically adjusts for the local dependence based on the data, while maintaining the unidimensional latent trait as the key source of dependence among responses. As a result, the FLaG-IRT model learned from data fits well, at both the item pair level and the test level.

References

Anderson, C. J. and Vermunt, J. K. (2000) Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, **30**, 81–121.

Anderson, C. J. and Yu, H.-T. (2007) Log-multiplicative association models as item response models. *Psychometrika*, **72**, 5–23.

Barber, R. F. and Drton, M. (2015) High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics*, **9**, 567–607.

Belloni, A. and Chernozhukov, V. (2013) Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–547.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 192–236.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (eds. F. M. Lord and M. R. Novick), 395–479. Reading, MA: Addison-Wesley.

Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D. and Schoevers, R. A. (2015) The network structure of symptoms of the diagnostic and statistical manual of mental disorders. *PLoS One*, **10**, e0137621.

Bradlow, E. T., Wainer, H. and Wang, X. (1999) A Bayesian random effects model for testlets. *Psychometrika*, **64**, 153–168.

Braeken, J. (2011) A boundary mixture approach to violations of conditional independence. *Psychometrika*, **76**, 57–76.

Braeken, J., Tuerlinckx, F. and De Boeck, P. (2007) Copula functions for residual dependency. *Psychometrika*, **72**, 393–411.

Cai, L., Yang, J. S. and Hansen, M. (2011) Generalized full-information item bifactor analysis. *Psychological Methods*, **16**, 221–248.

Chen, J. and Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

Chen, W.-H. and Thissen, D. (1997) Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, **22**, 265–289.

Chen, Y. (2016) Latent variable modeling and statistical learning. Available at `http://academiccommons.columbia.edu/catalog/ac:198122`. PhD thesis, Columbia University.

Chen, Y., Li, X., Liu, J. and Ying, Z. (2016) A fused latent and graphical model for multivariate binary data. Available at `https://arxiv.org/pdf/1606.08925v1.pdf`. ArXiv preprint.

Chen, Y., Liu, J., Xu, G. and Ying, Z. (2015a) Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, **110**, 850–866.

Chen, Y., Liu, J. and Ying, Z. (2015b) Online item calibration for Q-matrix in CD-CAT. *Applied Psychological Measurement*, **39**, 5–15.

Cramer, A. O., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S. and Borsboom, D. (2012) Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, **26**, 414–431.

Cramer, A. O., Waldorp, L. J., van der Maas, H. L. and Borsboom, D. (2010) Complex realities require complex theories: Refining and extending the network approach to mental disorders. *Behavioral and Brain Sciences*, **33**, 178–193.

Embretson, S. E. and Reise, S. P. (2000) *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Epskamp, S., Maris, G. K., Waldorp, L. J. and Borsboom, D. (2016) Network psychometrics. *arXiv preprint arXiv:1609.02818.*

Epskamp, S., Rhemtulla, M. and Borsboom, D. (2017) Generalized network pschometrics: Combining network and latent variable models. *Psychometrika*, **82**, 904–927.

Eysenck, S. and Barrett, P. (2013) Re-introduction to cross-cultural studies of the EPQ. *Personality and Individual Differences*, **54**, 485–489.

Eysenck, S. B., Eysenck, H. J. and Barrett, P. (1985) A revised version of the Psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Ferrara, S., Huynh, H. and Michaels, H. (1999) Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement*, **36**, 119–140.

Foygel, R. and Drton, M. (2010) Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, 604–612.

Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O., Epskamp, S., Tuerlinckx, F., Carr, D. and Stroebe, M. (2015) From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology*, **124**, 256–265.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J. and Stover, A. (2007) Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, **31**, 4–19.

Gibbons, R. D. and Hedeker, D. R. (1992) Full-information item bi-factor analysis. *Psychometrika*, **57**, 423–436.

Holland, P. W. (1990) The Dutch identity: A new tool for the study of item response models. *Psychometrika*, **55**, 5–18.

Holland, P. W. and Wainer, H. (2012) *Differential item functioning.* New York, NY: Routledge.

Hoskens, M. and De Boeck, P. (1997) A parametric model for local dependence among test items. *Psychological Methods*, **2**, 261–277.

Ip, E. H. (2002) Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, **67**, 367–386.

Ip, E. H. (2010) Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, **63**, 395–416.

Ip, E. H., Wang, Y. J., De Boeck, P. and Meulders, M. (2004) Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika*, **69**, 191–216.

Ising, E. (1925) Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, **31**, 253–258.

Knowles, E. S. and Condon, C. A. (2000) Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment*, **12**, 245–252.

Koller, D. and Friedman, N. (2009) *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT press.

Kruis, J. and Maris, G. (2016) Three representations of the Ising model. *Scientific Reports*, **6**.

Laird, N. M. (1991) Topics in likelihood-based methods for longitudinal data analysis. *Statistica Sinica*, **1**, 33–50.

Lee, J. D. and Hastie, T. J. (2015) Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, **24**, 230–253.

Li, Y., Bolt, D. M. and Fu, J. (2006) A comparison of alternative models for testlets. *Applied Psychological Measurement*, **30**, 3–21.

Liu, J. (2017) On the consistency of Q-matrix estimation: A commentary. *Psychometrika*, **82**, 523–527.

Liu, J., Xu, G. and Ying, Z. (2012) Data-driven learning of Q-matrix. *Applied Psychological Measurement*, **36**, 548–564.

Liu, J., Xu, G. and Ying, Z. (2013) Theory of the self-learning Q-matrix. *Bernoulli*, **19**, 1790–1817.

Lord, F. M. and Novick, M. R. (1968) *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marsman, M., Maris, G., Bechger, T. and Glas, C. (2015) Bayesian inference for low-rank Ising networks. *Scientific Reports*, **5**.

McKinley, R. L. and Reckase, M. D. (1982) The use of the general Rasch model with multidimensional item response data. Iowa City, IA: American College Testing.

Pan, J., Ip, E. H. and Dubé, L. (2017) An alternative to post hoc model modification in confirmatory factor analysis: The bayesian lasso. *Psychological methods*, **22**, 687–704.

Parikh, N. and Boyd, S. P. (2014) Proximal algorithms. *Foundations and Trends in Optimization*, **1**, 127–239.

Rasch, G. (1960) Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010) High-dimensional ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, **38**, 1287–1319.

Reckase, M. (2009) *Multidimensional item response theory.* New York, NY: Springer.

Reise, S. P., Horan, W. P. and Blanchard, J. J. (2011) The challenges of fitting an item response theory model to the social anhedonia scale. *Journal of Personality Assessment*, **93**, 213–224.

Reise, S. P., Morizot, J. and Hays, R. D. (2007) The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, **16**, 19–31.

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S. and Borsboom, D. (2016) Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence*, **161**, 230–237.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Schwarz, N. (1999) Self-reports: How the questions shape the answers. *American Psychologist*, **54**, 93–105.

Sun, J., Chen, Y., Liu, J., Ying, Z. and Xin, T. (2016) Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika*, **81**, 921–939.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A. and Waldorp, L. J. (2014) A new method for constructing networks from binary data. *Scientific Reports*, **4**.

van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M. and Raijmakers, M. E. (2006) A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, **113**, 842–861.

Wainer, H., Bradlow, E. T. and Du, Z. (2000) Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized adaptive testing: Theory and practice* (eds. W. J. van der Linden and G. A. Glas), 245–269. New York, NY: Springer.

Wang, W.-C. and Wilson, M. (2005) The Rasch testlet model. *Applied Psychological Measurement*, **9**, 126–149.

Yao, L. and Schwarz, R. D. (2006) A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, **30**, 469–492.

Yen, W. M. (1984) Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, **8**, 125–145.

Yen, W. M. (1993) Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, **30**, 187–213.

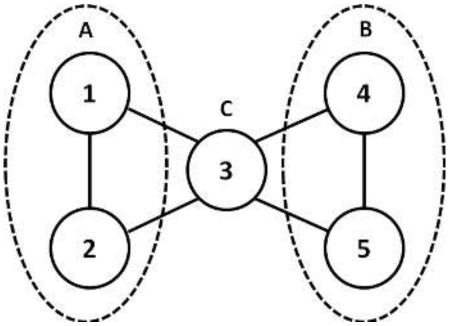

Figure 1.
The set C separates A from B. All paths from A to B pass through C.

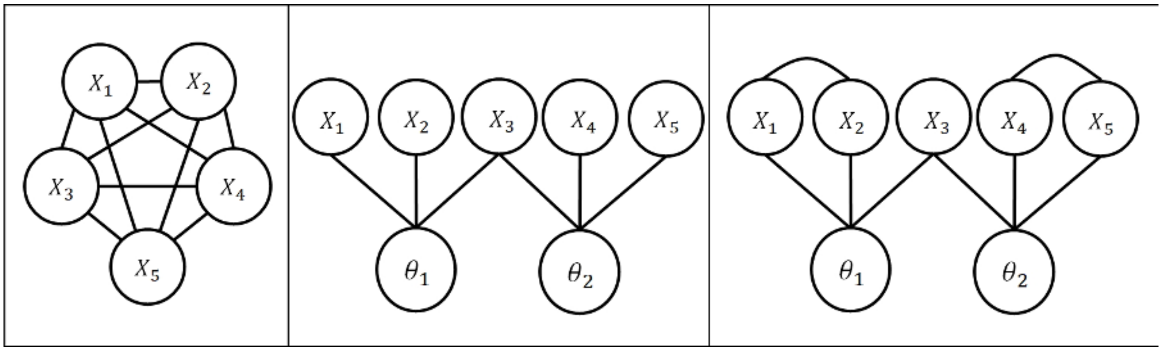

Figure 2.
Graphical illustration of the MIRT model and the FLaG-IRT model.

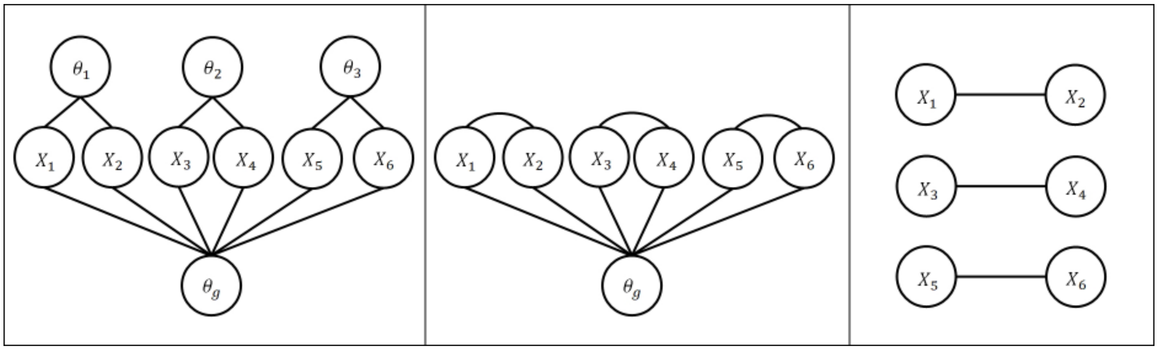

Figure 3.
Graphical representation of a bi-factor model, the corresponding FLaG-IRT model, and the local dependence graph.

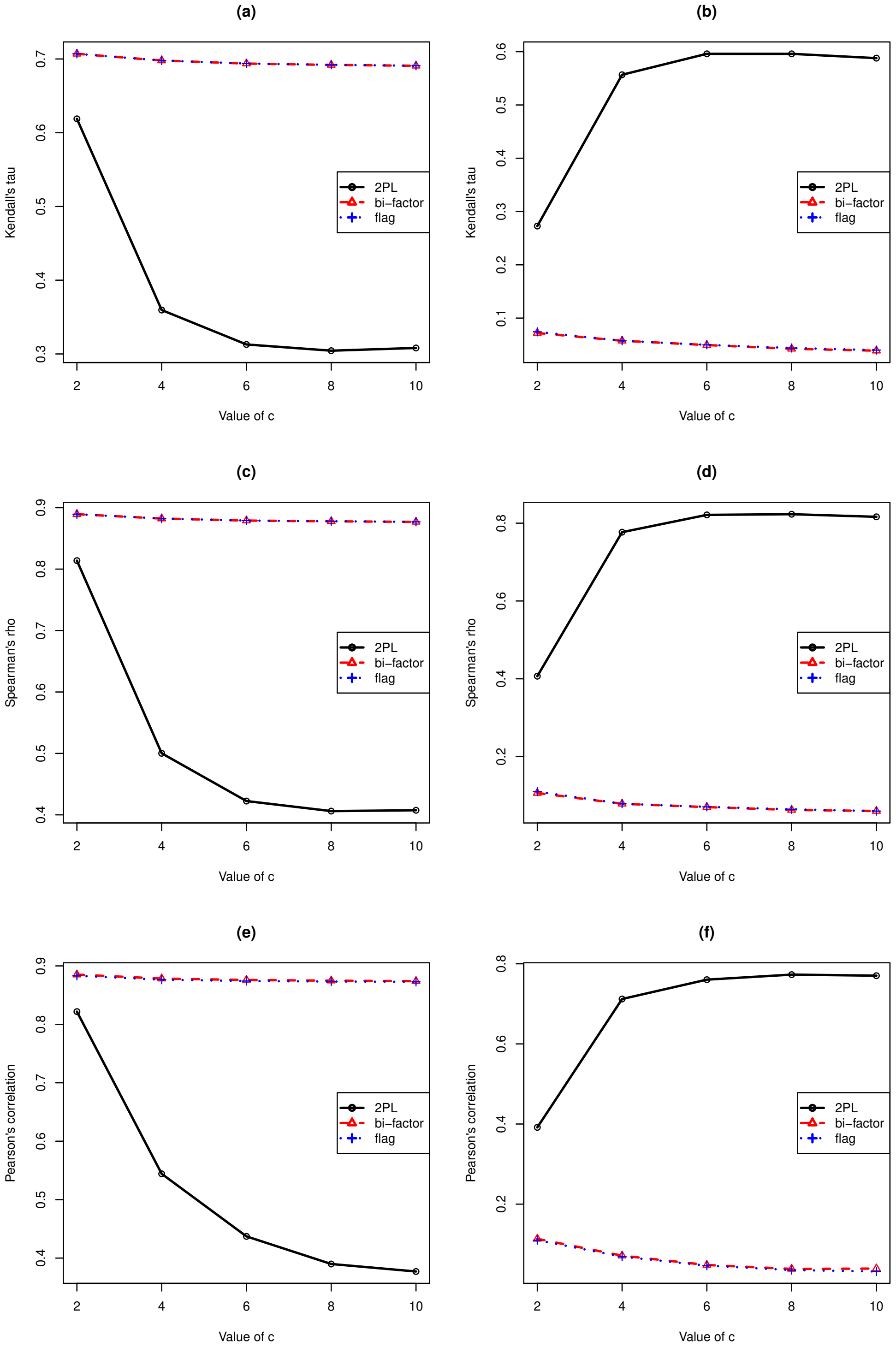

Figure 4.

Study 1: (a) Kendall' tau correlation between $\hat{\theta}_i$s and θ_{ig}s. (b) Kendall' tau correlation between $\hat{\theta}_i$s and θ_{i1}s. (c) Spearman' rho correlation between $\hat{\theta}_i$s and θ_{ig}s. (d) Spearman' rho correlation between $\hat{\theta}_i$s and θ_{i1}s. (e) Pearson' correlation between $\hat{\theta}_i$s and θ_{ig}s. (f) Pearson' correlation between $\hat{\theta}_i$s and θ_{i1}s.

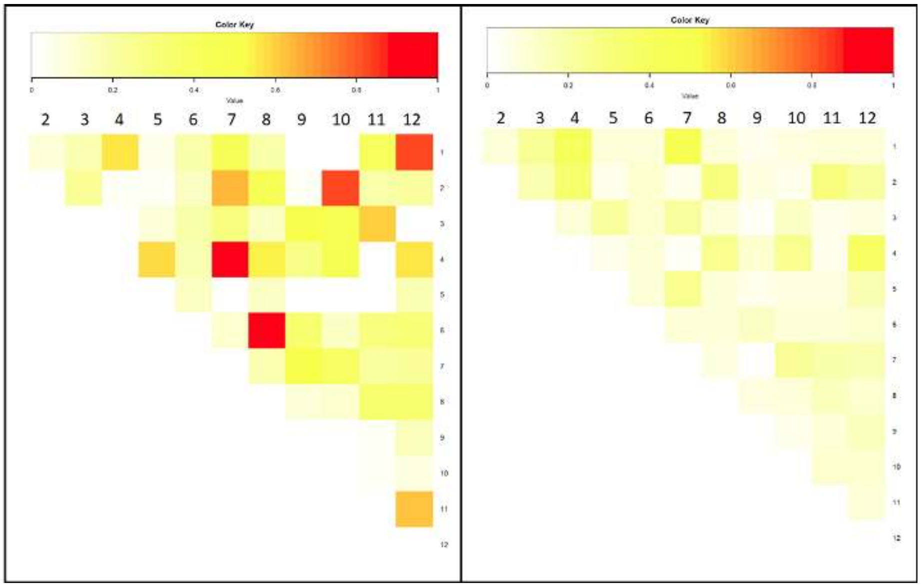

Figure 5.
Real data: The heat maps for visualizing the fit of all item pairs under the 2PL model (left) and selected FLaG-IRT model (right).

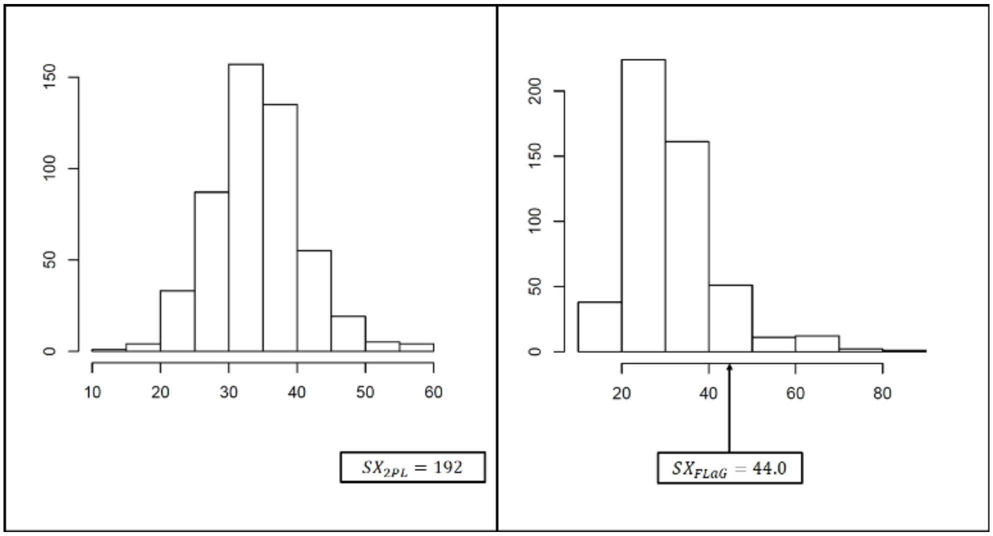

Figure 6.
Real data: The results of a parametric bootstrap test for the 2PL model (left) and the selected FLaG-IRT model (right)

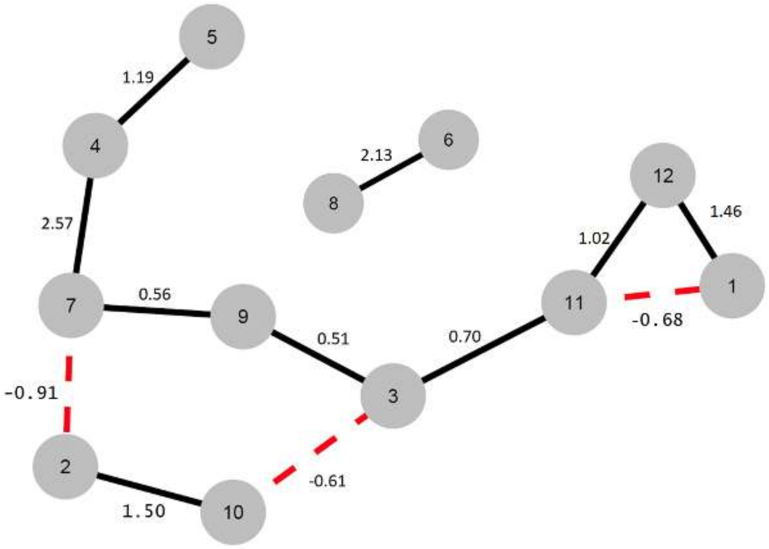

Figure 7.
Real data: The local dependence graph of the selected FLaG-IRT model.

S1		$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	Oracle
	Kendall's tau	0.61 (0.001)	0.61 (0.001)	0.61 (0.001)	0.63
$N = 500$	TPR	0.61 (0.007)	0.44 (0.007)	0.30 (0.008)	1
	FPR	0.08 (0.002)	0.02 (0.001)	0.01 (0.000)	0
	Kendall's tau	0.62 (0.001)	0.62 (0.001)	0.62 (0.001)	0.63
$N = 1000$	TPR	0.86 (0.006)	0.73 (0.007)	0.62 (0.007)	1
	FPR	0.06 (0.002)	0.02 (0.001)	0.01 (0.000)	0
	Kendall's tau	0.62 (0.000)	0.62 (0.001)	0.62 (0.001)	0.63
$N = 3000$	TPR	1.00 (0.000)	0.99 (0.001)	0.98 (0.002)	1
	FPR	0.04 (0.001)	0.01 (0.000)	0.00 (0.000)	0
S2		$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	Oracle
	Kendall's tau	0.67 (0.001)	0.67 (0.001)	0.67 (0.001)	0.68
$N = 500$	TPR	0.63 (0.005)	0.39 (0.004)	0.24 (0.004)	1
	FPR	0.08 (0.000)	0.02 (0.000)	0.00 (0.000)	0
	Kendall's tau	0.67 (0.001)	0.68 (0.001)	0.68 (0.001)	0.68
$N = 1000$	TPR	0.87 (0.003)	0.70 (0.005)	0.58 (0.006)	1
	FPR	0.06 (0.000)	0.01 (0.000)	0.01 (0.000)	0
	Kendall's tau	0.68 (0.000)	0.68 (0.000)	0.68 (0.000)	0.68
$N = 3000$	TPR	1.00 (0.000)	0.99 (0.001)	0.98 (0.001)	1
	FPR	0.03 (0.000)	0.01 (0.000)	0.00 (0.000)	0

Table 1.

Study 2: Performance of FLaG-IRT analysis when data are generated from a FLaG-IRT model. The average of each evaluation measure and its standard error over 100 independent replications are reported.

S1

$\rho = 0$	$N = 500$	$N = 1000$	$N = 3000$
$a_{11} = 0.4$	1.4×10^{-2}	8.4×10^{-3}	1.7×10^{-3}
$s_{11} = -4$	3.5×10^{-1}	2.2×10^{-1}	4.8×10^{-2}
$\sigma_{12} = 0.1$	2.6×10^{-3}	1.6×10^{-3}	3.9×10^{-4}
$\rho = 0.25$	$N = 500$	$N = 1000$	$N = 3000$
$a_{11} = 0.4$	1.2×10^{-2}	5.6×10^{-3}	1.5×10^{-3}
$s_{11} = -4$	2.9×10^{-1}	1.8×10^{-1}	4.5×10^{-2}
$\sigma_{12} = 0.1$	1.8×10^{-3}	8.9×10^{-4}	1.8×10^{-4}
$\rho = 0.5$	$N = 500$	$N = 1000$	$N = 3000$
$a_{11} = 0.4$	9.2×10^{-3}	5.5×10^{-3}	1.3×10^{-3}
$s_{11} = -4$	2.6×10^{-1}	1.6×10^{-1}	4.6×10^{-2}
$\sigma_{12} = 0.1$	1.6×10^{-3}	7.6×10^{-4}	1.3×10^{-4}

S2

$\rho = 0$	$N = 500$	$N = 1000$	$N = 3000$
$a_{11} = 0.35$	1.2×10^{-2}	6.5×10^{-3}	1.6×10^{-3}
$s_{11} = -4.5$	4.8×10^{-1}	2.3×10^{-1}	4.6×10^{-2}
$\sigma_{12} = 0.1$	2.5×10^{-3}	9.8×10^{-4}	3.2×10^{-4}
$\rho = 0.25$	$N = 500$	$N = 1000$	$N = 3000$
$a_{11} = 0.35$	7.2×10^{-3}	4.9×10^{-3}	9.6×10^{-4}
$s_{11} = -4.5$	3.4×10^{-1}	1.8×10^{-1}	4.5×10^{-2}
$\sigma_{12} = 0.1$	1.7×10^{-3}	6.8×10^{-4}	2.0×10^{-4}
$\rho = 0.5$	$N = 500$	$N = 1000$	$N = 3000$
$a_{11} = 0.35$	4.8×10^{-3}	3.8×10^{-3}	7.0×10^{-4}
$s_{11} = -4.5$	3.1×10^{-1}	1.7×10^{-1}	4.3×10^{-2}
$\sigma_{12} = 0.1$	1.3×10^{-3}	6.0×10^{-4}	1.4×10^{-4}

Table 2.

Study 2: Performance of FLaG-IRT analysis when data are generated from a FLaG-IRT model. The MSEs for the estimation of a_{11}, s_{11}, and σ_{12} calculated based on 100 independent replications are reported.

1	Are you a talkative person?
2	Are you rather lively?
3	Can you usually let yourself go and enjoy yourself at a lively party?
4	Do you enjoy meeting new people?
5	Do you usually take the initiative in making new friends?
6	Can you easily get some life into a rather dull party?
7	Do you like mixing with people?
8	Can you get a patty going?
9	Do you like plenty of bustle and excitement around you?
10	Do other people think of you as being very lively?
11(R)	Do you tend to keep in the background on social occasions?
12(R)	Are you mostly quiet when you are with other people?

Table 3.
Real data: The revised Eysenck Personality Questionnaire short form of Extroversion scale.

	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{a}_j$	1.90	2.13	1.82	1.67	1.53	2.48	2.27	2.25	0.85	2.49	1.74	2.05
$\hat{b}_j$	1.16	2.35	1.71	3.13	0.66	-0.51	2.80	0.53	0.91	1.81	0.60	1.13

Table 4.
Real data: The estimated 2PL model for the EPQ-R data.

	T_{ij}	X^2_{ij}	
1	0.89	32	6. Can you easily get some life into a rather dull party?
			8. Can you get a patty going?
2	0.88	28	4. Do you enjoy meeting new people?
			7. Do you like mixing with people?
3	0.83	18	2. Are you rather lively?
			10. Do other people think of you as being very lively?
4	0.83	18	1. Are you a talkative person?
			12. Are you mostly quiet when you are with other people?

Table 5.
Real data: Item pairs with largest values of local dependence indices.